Umjetna inteligencija

Uvod u strojno učenje

Marko Čupić

IZDAVAČ

JAVNO DOSTUPNO NA WEB STRANICI JAVA.ZEMRIS.FER.HR/NASTAVA/UI

Ovo je popratni materijal za kolegij Umjetna inteligencija na Fakultetu elektrotehnike i računarstva Sveučilišta u Zagrebu. Tekst je usklađen s prezentacijama koje se koriste na tom kolegiju i okvirno ih prati. Uz tekst je pripremljena dodatna biblioteka napisana u programskom jeziku Java koja ilustrira algoritme koji se obrađuju u tekstu. Čitatelj se upućuje da istu skine i prati te pokreće primjere o koji se obrađuju u tekstu.

Prvo izdanje, travanj 2020.

Sadržaj

1. Uvod

Kroz tekst koji slijedi (a i neke kasnije cjeline), upoznat ćemo se s pojmom *strojno učenje* te nekim od pristupa i algoritama koje ono obuhvaća. Postoji mnoštvo načina na koje možemo definirati što bi bilo strojno učenje. Tako primjerice Kevin P. Murphy u svojoj knjizi *Machine Learning - A Probabilistic Perspective* strojno učenje definira kao "skup metoda koje u podatcima mogu automatski otkrivati obrasce, i potom te otkrivene obrasce iskorištavati pri budućem predviđanju podataka, ili obavljati druge zadatke odlučivanja u prisustvu nesigurnosti".

Ono što je zajedničko mnogim definicijama jest spoznaja da danas živimo u svijetu u kojem smo okruženi obiljem podataka, te da nam je interesantno razvijati programske sustave koji su sposobni iskorištavati te podatke, učiti iz njih i na temelju toga nuditi korisna ponašanja.

Podatci s kojima raspolažemo mogu biti *numerički* ili *kategorički*. Numerički podatci su podatci poput visine čovjeka, težine čovjeka, vremena koje je čovjeku potrebno da pretrči 100 metara i slično. Nad numeričkim podatcima možemo raditi računske operacije i imamo intepretaciju dobivenog rezultata. Kategorički podatci su skup informacija koji je podijeljen u određene grupe. Kategorički podatci dijele se u nominalne i ordinalne. Nominalni podatci su imenovani podatci; primjerice, zamislimo situaciju gdje studente anketiramo te ih kao jedno pitanja pitamo kako se osjećaju, i nudimo opcije "veselo", "tužno", "nervozno". Sličan primjer: pitamo za spol i nudimo "ženski", "muški". Nominalne podatke karakterizira činjenica da nemaju numeričke vrijednosti (ne možemo raditi aritmetičke operacije; što bi bio rezultat oduzimanja "nervozno" od "veselo"?) kao niti poretka (ne možemo reći je li "nervozno" veće/manje/jednako "tužno"). Ordinalni podatci su podatci s kojima i dalje ne možemo raditi aritmetiku, ali imaju definiran poredak. Primjerice, pitamo li studenta u anketi kako je zadovoljan kolegijem Umjetna inteligencija, te ponudimo opcije "jako zadovoljan", "umjereno zadovoljan", "ni zadovoljan, ni nezadovoljan", "umjereno nezadovoljan", "jako nezadovoljan" - dobili smo ordinalni podatak. Te podatke možemo uspoređivati, no s njima ne možemo raditi aritmetiku (koje bi bilo značenje "umjereno zadovoljan" minus "jako nezadovoljan"?).

Strojno učenje dijelimo na tri glavna područja:

- nadzirano učenje,
- nenadzirano učenje te

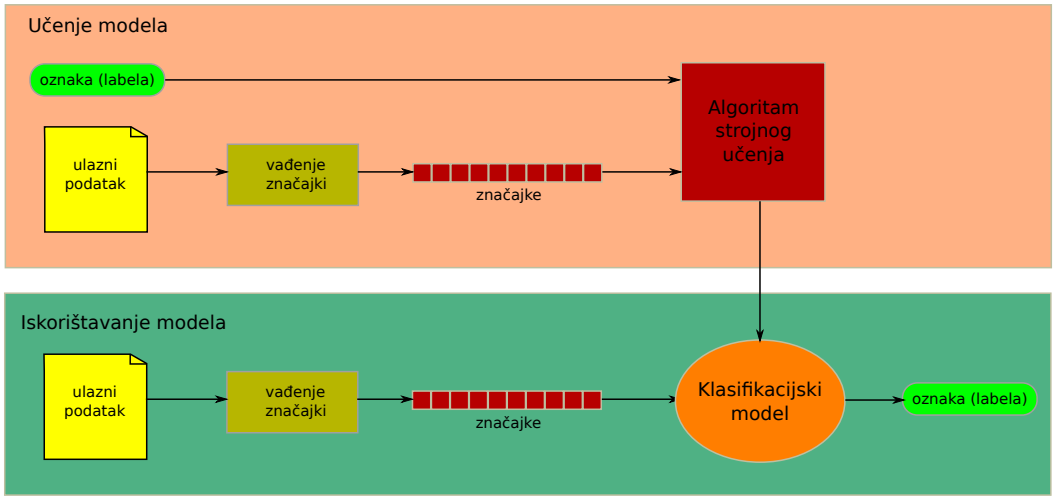
- podržano učenje.

1.1 Nadzirano učenje

Nadzirano učenje (engl. *supervised learning*) ima za cilj naučiti odnosno omogućiti obavljanje preslikavanja ulaza $\mathbf{x}$ na izlaz y, u skladu sa skupom uzoraka za učenje (engl. *training set*) koji je oblika $\mathscr{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Ovdje $\mathbf{x}_i$ predstavlja informacije o uzorku (to može biti vektor brojeva, slika, zvučni zapis, itd.) a y_i predstavlja podatak koji treba pridružiti tom uzorku. $\mathbf{x}_i$ ćemo zvati vektorom značajki uzorka, gdje svaka značajka govori o nekom aspektu uzorka. Primjerice, radimo li klasifikator koji će na temelju lista odrediti o kojem se drvu radi, značajke bi mogle biti: visina lista, širina lista, oblik lista, dominantna boja lista itd.

Ako je y_i kategorička varijabla, tada govorimo o klasifikacijskom problemu: uzorak je potrebno smjestiti u jedan od razreda; to bi bio slučaj s našim klasifikatorom, gdje bi y_i poprimao vrijednosti iz skupa {bor, hrast, jablan, breza, ...}. Ako je y_i numerička varijabla, tada govorimo o funkcijskoj aproksimaciji.

Kod modela strojnog učenja razlikujemo dvije faze: *faza učenja modela*, te *faza iskorištavanja modela*. Aktivnosti u svakoj od njih ilustrira slika u nastavku.



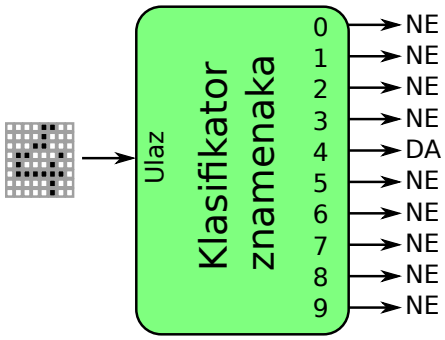
U fazi učenja, nad svakim se ulaznim podatkom obavlja vađenje značajki (engl. *feature extraction*). Na primjeru klasifikacije drveća an temelju lista, za svaki list koji imamo odredili bi prethodno spomenute značajke. Taj se vektor značajki zajedno s uparenom oznakom razreda šalje u algoritam strojnog učenja. Algoritam strojnog učenja na temelju ovih uzoraka uči optimalne parametre odabranog modela strojnog učenja - u ovom slučaju, odabranog modela koji obavlja zadaću klasifikacije (primjerice, uči težine umjetne neuronske mreže koju koristimo kao klasifikator).

U fazi iskorištavanja, korisnik postavlja ulazni podatak koji se ponovno obrađuje na jednak način: računaju se vrijednosti svih značajki i formira vektor značajki. Taj se vektor potom postavlja na ulaz naučenog modela strojnog učenja, koji na izlazu generira predviđanje (u našem slučaju, oznaku razreda u koji uzorak pripada). Jednom kad smo došli do faze iskorištavanja, gornji dio prethodne slike postaje višak i može se odbaciti.

Spomenimo da danas u određenim situacijama prethodno opisani "tijek" može biti izmijenjen tako da se izbaci središnji dio koji se bavi vađenjem ručno specificiranih značajki. Naime, ako je ulaz slika, jedan od tipičnih modela strojnog učenja koji se koristi za klasifikaciju su duboke neuronske mreže, a posebice konvolucijske neuronske mreže. Takvi modeli na ulaz mogu dobiti izvorni podatak ("čistu sliku") i tijekom učenja sami mogu u ranijim slojevima mreže otkriti koje su značajke bitne i kako se računaju, te u kasnijim slojevima mreže razviti klasifikacijsko ponašanje.

Za takve modele kažemo da rade posao od-kraja-do-kraja (engl. *end-to-end*): na jednom kraju dobiju "sirovi" ulazni podatak, a na drugom generiraju oznaku pripadnog razreda. Više o ovakvim modelima i njihovoj primjeni na obradu slika moći ćete naučiti na kolegiju Duboko učenje koji se nudi na diplomskom studiju.

Primjeri klasifikacijskih zadataka su razvrstavanje slika prema životinji koja je na slici, prepoznavanje rukom pisane znamenke sa slike, detekcija neželjene elektroničke pošte (engl. *spam*) i slično. Slika 1.1 ilustrira klasifikacijski problem gdje je ulaz skenirana slika znamenke, veličine 8×8 slikovnih elemenata; svaki slikovni element je crni ili bijeli. Klasifikatorski sustav treba na izlazu odrediti koju znamenku predstavlja ta slika, odnosno dani ulazni uzorak smjestiti u jedan od 10 razreda.



Slika 1.1: Klasifikacija znamenaka

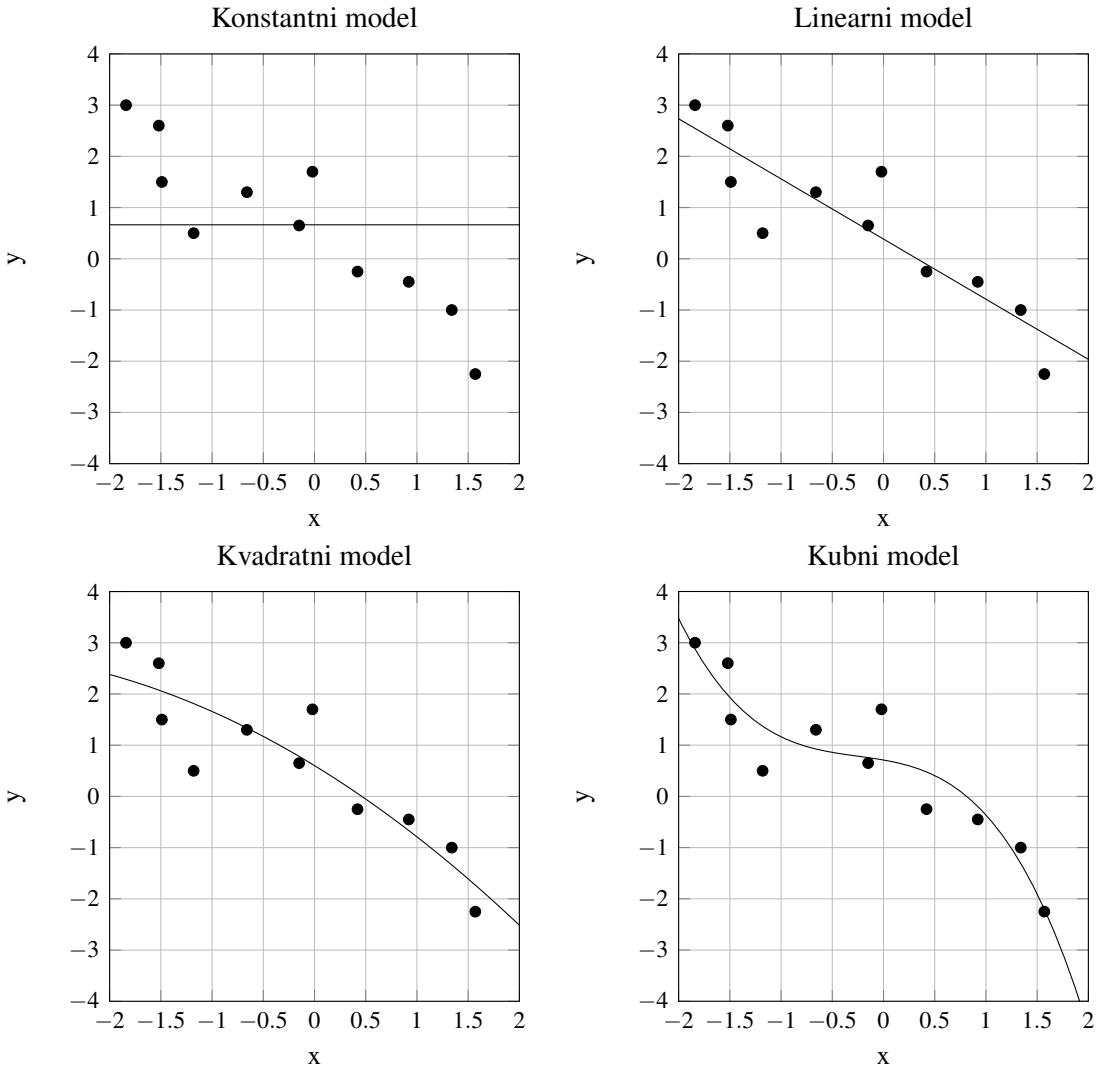
Primjeri regresije su predviđanje ukupnog broja bodova koje će student ostvariti na kolegiju na temelju bodova komponenata prve polovice semestra, predviđanje potrošnje plina na temelju povijesnih podataka, predviđanje potrošnje električne energije kućanstva na temelju vremenskih parametara (temperatura, tlak) i drugih podataka, itd.

Kao primjer regresije razmotrit ćemo skup podataka prikazan tablicom u nastavku.

x	y
-1.84	3.00
-1.52	2.60
-1.49	1.50
-1.18	0.50
-0.66	1.30
-0.15	0.65
-0.02	1.70
0.42	-0.25
0.92	-0.45
1.34	-1.00
1.57	-2.25

Potrebno je pronaći prikladan model $y = f(x)$ kako bismo iz modela mogli dobiti informaciju koliki će biti y za x-eve koji nisu u skupu uzoraka za učenje. Modela koji se mogu koristiti za regresiju ima mnoštvo i o nekima od njih (poput umjetnih neuronskih mreža) pričat ćemo kroz kasnije teme. Ovdje smo u nastavku prikazali četiri parametarska modela: konstantni ($y = a$), linearni ($y = ax + b$), kvadratni ($y = ax^2 + bx + c$) i kubni ($y = ax^3 + bx^2 + cx + d$). Svaki od ovih modela ima fiksan broj parametara koje je potrebno prilagoditi tako da model najbolje odgovara podatcima iz skupa za učenje.

Pri tome, pojam "najbolje odgovara" općenito može značiti različite stvari. Primjerice, može podrazumijevati da imamo definiranu funkciju pogreške (još je nazivamo i funkcijom gubitka) i da tijekom učenja želimo pronaći one parametre modela uz koje je ta funkcija pogreške nad podatcima koje imamo minimalna. Ako je to slučaj, tada govorimo o klasičnom optimizacijskom problemu. Kod regresije, uobičajena funkcija pogreške je polovično kvadratno odstupanje vrijednosti koju daje model i vrijednosti koja je zapisana uz uzorak, pa sumirano po svim uzorcima.



Uz tako definiranu funkciju pogreške, prethodno spomenuti modeli uz koju je pogreška mala bili bi:

$$y = 0.6636 \quad \text{(konstantni model)}$$

$$y = 0.3850 - 1.1743 \cdot x \quad \text{(linearni model)}$$

$$y = 0.6019 - 1.2239 \cdot x - 0.1671 \cdot x^2 \quad \text{(kvadratni model)}$$

$$y = 0.7103 - 0.3496 \cdot x - 0.3102 \cdot x^2 - 0.4129 \cdot x^3 \quad \text{(kubni model)}$$

Što je model složeniji (ima više parametara), to će se moći bolje prilagoditi podatcima, što može biti nepoželjno.

Pogledajte sada još jednom prethodne četiri slike i četiri modela koja smo naučili. Ako smo svjesni da su podatci koje imamo samo jedanaest mjerenja ulaza i izlaza nekog procesa (a teoretski

bismo mogli napraviti još tisuće i tisuće mjerenja), odnosno predstavljaju samo mali slučajni uzorak iz inače potencijalno beskonačno velikog skupa podataka, te ako smo svjesni da je prilikom mjerenja nužno djelovao i određen šum zbog kojeg nam podatci nisu savršeni, s kojim od prethodna četiri modela bismo bili zadovoljni?

Primijetite da je prvi model vrlo krut - ima samo jedan parametar i vrlo smo ga slabo mogli prilagoditi skupu uzoraka koji imamo na raspolaganju. S druge strane, četvrti model je već dovoljno ekspresivan da je uspio modelirati i određeno "vijuganje" u skupu za učenje. Da smo uzeli još bogatiji model (kažemo model višeg kapaciteta), isti smo mogli još bolje prilagoditi podatcima, a uz dovoljno kapacitivan model mogli smo dobiti funkciju koja doslovno vijuga od mjerenja do mjerenja te izmjerenih 11 uzoraka modelira savršeno (naša definirana greška tada bi pala na nulu).

Pitanje koje si ovdje trebamo postaviti – i to je pravi problem strojnog učenja – koji od ovih modela je "najbolji" model? Da bismo odgovorili na to pitanje, uzet ćemo u obzir da ako u podatcima postoji šum ili ako postoje stršeće vrijednosti, presložen model moći će jako dobro modelirati i te pojave na uštrb dobre generalizacije. Prisjetite se: naš skup uzoraka za učenje nije kompletan opis modeliranog procesa: to je samo mali slučajan uzorak. Ono što želimo jest odabrati i naučiti onaj model koji bi savršeno odgovarao našem procesu, odnosno koji bi na beskonačno velikom skupu podataka koji bi opisivali naš proces imao minimalnu pogrešku. Problem je što nam taj skup nije dostupan (a i da jest, ako je beskonačno velik, nikada ne bismo završili postupak učenja pa opet nemamo koristi). Stoga bismo htjeli učiti na skupu uzoraka koji imamo, ali tako da minimiziramo grešku na "stvarnom" skupu. Htjeli bismo, dakle, da naš naučeni model i na podatcima koje nije vidio tijekom učenja radi dobro. Za model koji ima to svojstvo kažemo da dobro **generalizira**. Za model koji smo naučili do te mjere da ima jako malu pogrešku na skupu za učenje, ali radi veliku pogrešku na podatcima koje nije vidio, kažemo da je **prenaučen** (ili pretreniran). Takav model nam je u praksi beskoristan.

Postoji više načina kako se može utjecati na generalizacijske sposobnosti modela. Mi ćemo ovdje spomenuti pristup koji se naziva *unakrsna provjera*. Kako nemamo "pravi" beskonačno veliki skupo podataka, a htjeli bismo naučiti model koji dobro radi i na neviđenim primjerima, skup uzoraka s kojim raspolažemo, razdijelit ćemo u dva podskupa. Veći dio uzoraka smjestit ćemo u *skup za učenje* (engl. *training set*). Preostali manji dio uzoraka činit će skup za provjeru (engl. *validation set*). Koliko ćemo točno uzoraka uzeti u skup za učenje, a koliko u skup za provjeru, nije presudno. Neki autori (primjerice, Kevin P. Murphy u knjizi *Machine Learning - A Probabilistic Perspective*) uzimaju 80% raspoloživih uzoraka u skup za učenje, a preostalih 20% u skup za provjeru. Drugi autori navode drugačije podatke (primjerice, 70% u skup za učenje, 30% u skup za provjeru). Jednom kad smo uzorke razdvojili u dva skupa, provodimo učenje modela, ali samo na skupu za učenje (drugim riječima, kod naših modela iz prethodnog primjera, na temelju skupa za učenje korigirali bismo parametre modela). Povremeno, kako se model ponaša provjeravamo predočavanjem uzoraka iz skupa za provjeru; ta predočavanja koristimo isključivo da bismo odredili koliko model griješi nad uzorcima tog skupa, ali pri tim predočavanjima ne korigiramo parametre (drugim riječima, model nad tim uzorcima ne uči). Tijekom učenja, dogodit će se nešto interesantno: u ranim fazama učenja, kako prilagođavamo parametre modela, greška na skupu na učenje, kao i greška na skupu za provjeru, polagano će padati. Model će se prilagođavati generalnim trendovima u podatcima (učit će generalizirati). Međutim, u jednom trenutku, kako nastavljamo s učenjem nad uzorcima iz skupa za učenje, model koji je dovoljno ekspresivan počet će se prilagođavati specifičnom šumu i stršećim vrijednostima iz tog skupa, a nad neviđenim podatcima počet će sve više griješiti. To ćemo identificirati time što ćemo uočiti da od tog trenutka na dalje pogreška koju računamo na skupu za provjeru počinje rasti! Trenutak u kojem to identificiramo je trenutak u kojem treba prekinuti učenje: nastavkom učenja model se počinje prenaučavati, a to ne želimo.

Ako naš model ima i parametre koji utječu na njegovu složenost, tada ćemo trebati na neki način odrediti i te parametre. Konkretno, četiri modela koja smo prikazali u prethodnom primjeru

zapravo bismo mogli tretirati kao jedan model: polinomijalni, uz parametar r koji predstavlja red polinoma. Tako je naš konstantni model zapravo polinomijalni model reda 0, linearni model je zapravo polinomijalni model reda 1, i tako dalje. Koji bismo od modela u konačnici htjeli odabrati kao "najbolji" model? Da bismo odgovorili na to pitanje, skup podataka s kojim raspolažemo morat ćemo još razdijeliti:

- jedan dio uzoraka smjestit ćemo u *skup za učenje* (engl. *training set*; primjerice 40%),
- jedan dio uzoraka smjestit ćemo u *skup za provjeru* (engl. *validation set*; primjerice 30%),
- jedan dio uzoraka smjestit ćemo u *skup za ispitivanje* (engl. *test set*; primjerice 30%).

Nad skupom za učenje učit ćemo polinomijalni model reda 0, a s učenjem ćemo stati kad greška na skupu za provjeru počne rasti. Potom ćemo nad skupom za učenje učit ćemo polinomijalni model reda 1 i s učenjem stati kad greška na skupu za provjeru počne rasti. Potom ćemo nad skupom za učenje učit ćemo polinomijalni model reda 2 i s učenjem stati kad greška na skupu za provjeru počne rasti. Potom ćemo nad skupom za učenje učit ćemo polinomijalni model reda 3 i s učenjem stati kad greška na skupu za provjeru počne rasti. Jednom kad smo naučili modele uz svaki od parametara koji određuju kompleksnost modela, svaki od naučenih modela ispitat ćemo nad skupom za konačno ispitivanje. Model koji na tom skupu ima minimalnu pogrešku uzet ćemo kao najbolji model.

Primijetite da kod ovog pristupa i uzorci skupa za provjeru i uzorci skupa za konačno ispitivanje glume uzorke koje model prethodno nije vidio (u kontekstu učenja modela). Svaki od ta dva skupa, međutim, ima različitu zadaću.

Ako je skup uzoraka s kojim raspolažemo malen, tada podjela na opisani može biti problematična. Stoga su razvijene modifikacije opisanog postupka o kojima ćete imati prilike više naučiti na diplomskom studiju.

1.2 Nenadzirano učenje

Nenadzirano učenje (engl. *unspervised learning*) čine pristupi kod kojih je skup podataka oblika $\mathscr{D} = \{(\mathbf{x}_i)\}_{i=1}^{N}$. Drugim riječima, ništa osim samih uzoraka nije dano: nemamo numeričke ili kategoričke vrijednosti y_i koja je povezana s uzorkom. U postupke nenadziranog učenja spadaju postupci grupiranja (engl. *clustering*), postupci otkrivanja stršećih ili novih vrijednosti (engl. *outlier detection*, *novelty detection*), postupci smanjenja dimenzionalnosti (engl. *dimensionality reduction*), postupci otkrivanja veza između uzoraka (engl. *discovering graph structure*) te drugi.

Zadaća grupiranja jest na temelju sličnosti između uzoraka iste razdijeliti u određen broj razreda. Primjerice, slike životinja želimo grupirati u grupe, pri čemu se u jednoj grupi nalaze slike jedne životinjske vrste. Ovisno što znamo o podatcima, zadaća grupiranja može biti razdijeliti uzorke u unaprijed definirane razrede, ili pak može biti odrediti i potreban broj razreda i koji uzorak pripada u koji od razreda. Da bismo uzorke mogli grupirati, potrebno je biti u stanju računati njihovu udaljenost: trebamo neku metriku.

Nove ili stršeće vrijednosti su vrijednosti atributa koje imamo u skupu uzoraka, no koje su po nekom kriteriju dovoljno drugačije da ih želimo detektirati. Primjerice, zamislite neki proces kod kojega je izlaz y funkcija ulaza x u skladu s modelom $y = 2x$. Radite niz mjerenja, i dobijete skup uzoraka {(0,0),(1,2),(2,4),(3,60),(4,8),(5,10),(6,12)}. Podatak (3,60) očito je stršeći uzorak i ovdje predstavlja grešku (prilikom zapisivanja, osoba koja je bilježila podatke greškom je dopisala 0). S druge strane, podatci koji su dovoljno drugačiji ponekad mogu predstavljati i nešto novo odnosno nešto specifično za proces što u ostatku podataka nismo vidjeli. U oba slučaja, željeli bismo biti u stanju detektirati takve podatke.

Zada postupaka smanjenja dimenzionalnosti jest smanjiti broj atributa kojima opisujemo uzorke. Evo jednostavnog primjera: imamo skup uzoraka koji su točke u trodimenzijskom prostoru, i svi leže u istoj ravnini. Zamislimo sada da smo u toj ravnini nacrtali novi koordinatni sustav: to bi bio dvodimenzijski koordinatni sustav, i svaki bismo uzorak mogli zapisati uporabom dva broja,

umjesto tri. Ovo je trivijalan primjer, ali bitan je zbog vrlo značajnog problema: ako naši uzorci imaju mnoštvo atributa (a danas nije rijetkost da imamo uzorke kod kojih se broj atributa mjeri u tisućama), tada je algoritmima strojnog učenja tipično vrlo teško kvalitetno raditi s takvim uzorcima (primjerice, ako razmatramo klasifikator uzoraka, postići dobru generalizaciju). Postupci smanjenja dimenzionalnosti mogu iz mnoštva atributa kojima su opisani uzorci identificirati (ili ponekad osmisliti nove) manji broj onih koji su relevantni za značenje uzoraka.

1.3 Podržano učenje

Podržano učenje predstavlja dio strojnog učenja koji se bavi optimizacijom ponašanja. Za razliku od nadziranog i nenadziranog učenja, kod podržanog učenja razmatramo interakciju agenta i okoline u kojoj se agent nalazi. Agent na temelju informacija iz okoline obavlja akcije, i kao odgovor za svaku akciju od okoline dobiva nagradu ili kaznu. Zadaća podržanog učenja jest razviti "upravljački sustav" agenta, odnosno otkriti optimalnu strategiju njegovog ponašanja, tako da agent maksimizira nagrade koje dobiva "na duge staze".

Više o podržanom učenju čeka nas u zasebnoj, za to predviđenoj, temi.

2. Naivni Bayesov klasifikator

Naivni Bayesov klasifikator model je strojnog učenja koji omogućava klasifikaciju uzoraka. Kako se radi o modelu koji je utemeljen na teoriji vjerojatnosti, najprije ćemo se podsjetiti osnova teorije vjerojatnosti, a potom pogledati primjenu iste na razvoj naivnog Bayesovog klasifikatora.

2.1 Podsjetnik na teoriju vjerojatnosti

Kada razmišljamo o vjerojatnostima, i što iste zapravo znače, često se razmišljamo o eksperimentu čiji je ishod slučajan. Primjeri takvih eksperimenata su bacanje novčića (zanima nas hoće li rezultat biti glava ili pismo), bacanje šesterostrane kocke (zanima nas hoćemo li dobiti broj 1, 2, 3, 4, 5 ili 6), izvlačenje "na slijepo" kuglice iz kutije u koju je ubačeno 36 kuglica pri čemu na svakoj piše jedan broj od 0 do 35 i nema dvije kuglice s istim brojem (zanima nas koji ćemo broj izvući) i slično.

Slučajni događaj u tom je kontekstu događaj koji se može ili ne mora dogoditi (primjerice, rezultat bacanja novčića je glava, kocka je pala na broj manji ili jednak 3, izvučena je kuglica s brojem 17, izvučena je kuglica s parnim brojem). *Elementarni događaji* su svi mogući ishodi eksperimenta: u kontekstu bacanja novčića, to bi bio dvočlani skup {glava, pismo}, u kontekstu bacanja kocke šesteročlani skup brojeva {1, 2, 3, 4, 5, 6}, a u kontekstu izvlačenja kuglica trideset i šesteročlani skup {0, 1, 2, …, 34, 35}. Skup svih mogućih ishoda eksperimenta (odnosno skup svih elementarnih događaja) nazivamo *prostor elementarnih događaja* (u engleskoj terminologiji najčešće engl. *Sample space*) i taj ćemo skup označiti slovom S. Stoga imamo:

$$S_{\text{novčić}} = \{\text{glava}, \text{pismo}\},$$
$$S_{\text{kocka}} = \{1,2,3,4,5,6\},$$
$$S_{\text{kuglice}} = \{0,1,2,\ldots,34,35\}.$$

U slučaju bacanja dva novčića, elementarni događaj opisali bismo uređenim parom koji bi se sastojao od rezultata bacanja prvog novčića te rezultata bacanja drugog novčića pa bi skup elementarnih događaja imao 4 elementa:

$$S_{\text{dva novčića}} = \{(\text{glava}, \text{glava}), (\text{glava}, \text{pismo}), (\text{pismo}, \text{glava}), (\text{pismo}, \text{pismo})\},$$

a bacamo li kocku dva puta, elementarni bi događaj opisali uređenim parom koji bi se sastojao od broja koji smo dobili prvim bacanjem te broja koji smo dobili drugim bacanjem pa bi skup elementarnih događaja imao 36 elemenata:

$$S_{\text{dvije kocke}} = \{(1,1),(1,2),\ldots,(2,1),(2,2),\ldots(6,5),(6,6)\}.$$

Svaki podskup skupa elementarnih događaja nazivamo *događajem* (engl. *event*). Tako skup $S_{\text{novčić}}$ ima četiri podskupa: $\emptyset$ (prazan skup), $\{\text{glava}\}$, $\{\text{pismo}\}$ te , $\{\text{glava}, \text{pismo}\}$. Prvi skup bi opisivao događaj "novčić nije bačen", drugi "dobili smo glavu", treći "dobili smo pismo", a četvrti "novčić je bačen" (čime je jasno da smo dobili glavu ili pismo, no ne zanima nas što).

Kod bacanja kocke mogli bismo definirati mnoštvo događaja, uključivo 6 događaja koji odgovaraju "dobili smo broj i", događaj "dobili smo paran broj" (pridruženi skup bio bi $\{2,4,6\}$), "dobili smo neparan broj" (pridruženi skup bio bi $\{1,3,5\}$), "dobili smo prost broj" (pridruženi skup bio bi $\{2,3,5\}$), "dobili smo broj manji od 4" (pridruženi skup bio bi $\{1,2,3\}$), i tako dalje.

Za neki događaj E kažemo da se dogodio, ako pridruženi skup sadrži elementarni događaj koji je rezultat provedenog slučajnog eksperimenta. Primjerice, bacimo li kocku i dobijemo broj 2, od prethodno navedenih događaja dogodili su se "dobili smo broj 2", "dobili smo paran broj", "dobili smo prost broj", "dobili smo broj manji od 4", a nisu se dogodili događaji "dobili smo broj 1", "dobili smo broj 3", "dobili smo broj 4", "dobili smo broj 5", "dobili smo broj 6", "dobili smo neparan broj".

Imamo li dva događaja E_1 i E_2 nad skupom elementarnih događaja S, možemo definirati:

- događaj $E_1 \cup E_2$ koji se događa svaki puta kada je rezultat eksperimenta elementarni događaj koji je bilo u skup događaja E_1, bilo u skupu događaja E_2 (drugim riječima, događa se kad se dogodi ili E_1 ili E_2 (ili oba),
- događaj $E_1 \cap E_2$ koji se događa svaki puta kada je rezultat eksperimenta elementarni događaj koji je i u skupu događaja E_1 i u skupu događaja E_2 (drugim riječima, događa se samo kad se dogodi i E_1 i E_2 istovremeno.

Dva događaja E_1 i E_2 su međusobno isključiva ako je događaj $E_1 \cap E_2$ opisan praznim skupom (drugim riječima, ne dijele niti jedan elementarni događaj). Primjer takvih događaja su "dobili smo paran broj" i "dobili smo neparan broj".

Komplementaran događaj događaju E označit ćemo s E^c i definirati kao događaj čiji pridruženi skup čine svi elementarni događaji iz S koji nisu u skupu događaja E. Primijetite u našem primjeru da je komplementarni događaj događaja "dobili smo paran broj" događaj "dobili smo neparan broj", komplementarni događaj događaja "dobili smo glavu" bilo bi događaj "dobili smo pismo", itd.

Sada možemo definirati i vjerojatnost nekog događaja.

Theorem 2.1 — Vjerojatnost događaja. Neka je S skup elementarnih događaja, E neki događaj definiran nad S. *Vjerojatnost događaja* E, oznaka $P(E)$, je broj koji zadovoljava sljedeća svojstva:

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$

Za svaki niz međusobno isključivih događaja $E_1, E_2, \ldots$, mora vrijediti:

$$P(\bigcup_i E_i) = \sum_i E_i$$

Pogledajmo primjer. Pretpostavimo da je kocka koju bacamo takva da je vjerojatnost dobivanja

svakog od brojeva jednaka (označimo je s p); možemo pisati:

$$P(1) = p$$
$$P(2) = p$$
$$P(3) = p$$
$$P(4) = p$$
$$P(5) = p$$
$$P(6) = p.$$

Prema (2) iz prethodnog teorema mora vrijediti $P(\{1,2,3,4,5,6\}) = 1$ što možemo zapisati i kao $P(\{1\} \cup \{2\} \cup \{3\} \cup \{4\} \cup \{5\} \cup \{6\}) = 1$, a kako su ovi događaji međusobno isključivi, po istom teoremu slijedi:

$$\begin{aligned} P(\{1\}) + P(\{2\}) + P(\{3\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) &= 1 \\ p + p + p + p + p + p &= 1 \\ 6 \cdot p &= 1 \\ p &= \frac{1}{6} \end{aligned}$$

odnosno:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

što je u skladu s očekivanim.

Primjetimo također da su događaji E i E^c uvijek međusobno isključivi pa je $P(E \cup E^c) = P(E) + P(E^c)$. S druge strane, $E \cup E^c = S$, pa je $P(E \cup E^c) = P(S) = 1$, čime imamo $P(E) + P(E^c) = 1$ odnosno:

$$P(E^c) = 1 - P(E).$$

Napravimo još jedan primjer.

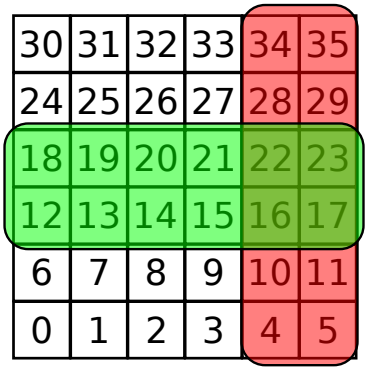
■ **Primjer 2.1** Ponovno razmatramo eksperiment izvlačenja jedne kuglice iz kutije. U kutiju je ubačeno 36 kuglica: kuglica s brojem 0, kuglica s brojem 1, itd. Sve su kuglice istih dimenzija i na opip djeluju isto. Prije izvlačenja, kutija je protresena i izvlačenje se radi na slijepo. Definirajmo dva događaja:

$$A = \{12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23\}$$

$$B = \{4, 5, 10, 11, 16, 17, 22, 23, 28, 29, 34, 35\}$$

Trebamo odrediti $P(A)$, $P(B)$, $P(A \cap B)$ i $P(A \cup B)$.

Rješenje. Skup elementarnih događaja koji ima 36 elemenata vizualiziran je na slici u nastavku. Na istoj slici označeni su i skupovi A (zeleno) i B (crveno).



Kako ima 36 elementarnih događaja i svi su jednako vjerojatni, vjerojatnost svakog od njih je $1/36$. Skupovi A i B imaju po 12 elementarnih događaja pa je:

$$\begin{aligned}
P(A) &= P(\{12,13,14,15,16,17,18,19,20,21,22,23\}) \\
&= P(\{12\}\cup\{13\}\cup\{14\}\cup\{15\}\cup\{16\}\cup\{17\}\cup\{18\}\cup\{19\}\cup\{20\}\cup\{21\} \\
&\quad \cup\{22\}\cup\{23\}) \\
&= P(\{12\})+P(\{13\})+P(\{14\})+P(\{15\})+P(\{16\})+P(\{17\})+P(\{18\}) \\
&\quad +P(\{19\})+P(\{20\})+P(\{21\})+P(\{22\})+P(\{23\}) \\
&= \frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36} \\
&= \frac{12}{36}=\frac{1}{3}
\end{aligned}$$

Na analogan način dobivamo:

$$P(B)=\frac{1}{3}.$$

Vjerojatnost događa $P(A\cap B)$ je:

$$\begin{aligned}
P(A\cap B) &= P(\{16,17,22,23\}) \\
&= P(\{16\})+P(\{17\})+P(\{22\})+P(\{23\}) \\
&= \frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36} \\
&= \frac{4}{36}=\frac{1}{9}
\end{aligned}$$

Vjerojatnost unije dvaju događaja prema teoremu s početka poglavlja jednaka je sumi vjerojatnosti tih događaja, ako su isti međusobno isključivi. Uvidom u sliku vidimo da to ovdje nije slučaj. Uvidom u sliku vidimo da unija događaja A i B sadrži 20 elementarnih događaja, dok svaki od događaja A i B sadrži po 12 elementarnih događaja. Pri izračunu $P(A)$ uzet ćemo u obzir sve ishode izvlačenja koji su brojevi od 12 do 23. Ako tome nadodamo $P(B)$, vidimo da smo dva puta pribrojili vjerojatnosti elementarnih događa 16, 17, 22 i 23, a oni su upravo presjek skupova A i B. Stoga je vjerojatnost unije događaja A i B jednaka sumi vjerojatnosti događaja A i događaja B umanjenoj za vjerojatnost presjeka tih događaja:

$$\begin{aligned}
P(A\cup B) &= P(A)+P(B)-P(A\cap B) \\
&= \frac{1}{3}+\frac{1}{3}-\frac{1}{9} \\
&= \frac{5}{9}
\end{aligned}$$

■

Ako su $E_1,\ldots,E_n$ međusobno isključivi događaji čija je unija jednaka S (drugim riječima čine particiju skupa S), tada je:

$$P(A)=\sum_{i=1}^{n}P(A,E_i)$$

Pogledajmo to na prethodnom primjeru. Skup elementarnih događaja koji čine događaj A možemo podijeliti na uniju skupa elementarnih događaja koji čine A i istovremeno su izvan B

(12,13,14,15,18,19,20,21) te skupa elementarnih događaja koji čine A i istovremeno su na B (16,17,22,23). Vrijedi:

$$\begin{aligned}
P(A) &= \frac{|A|}{|S|} \\
&= \frac{|\{12,13,14,15,16,17,18,19,20,21,22,23\}|}{|0,1,\ldots,35|} \\
&= \frac{|\{12,13,14,15,18,19,20,21\} \cup \{16,17,22,23\}|}{|0,1,\ldots,35|} \\
&= \frac{|\{12,13,14,15,18,19,20,21\}| + |\{16,17,22,23\}|}{|0,1,\ldots,35|} \\
&= \frac{|\{12,13,14,15,18,19,20,21\}|}{|0,1,\ldots,35|} + \frac{|\{16,17,22,23\}|}{|0,1,\ldots,35|} \\
&= \frac{|A \cap B^c|}{|S|} + \frac{|\{A \cap B\}|}{|S|} \\
&= P(A,B^c) + P(A,B)
\end{aligned}$$

pri čemu su B i B^c disjunktni skupovi čija je unija jednaka skupu S.

2.1.1 Nezavisni događaji

Bitan pojam koji ćemo trebati pri objašnjenju naivnog Bayesovog klasifikatora jest nezavisnost događaja.

Theorem 2.2 — Nezavisnost događaja. Za dva događaja A i B kažemo da su nezavisni, ako i samo ako vrijedi:

$$P(A,B) = P(A) \cdot P(B)$$

odnosno da je vjerojatnost da nastupe oba događaja jednaka umnošku vjerojatnosti nastupanja svakog od događaja.

Pogledajmo to na primjeru.

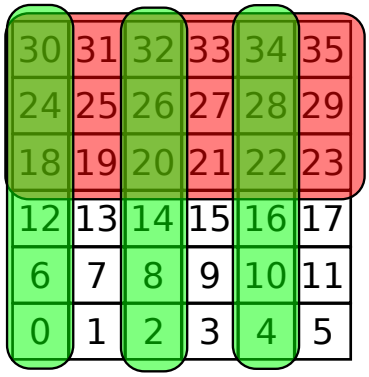
■ **Primjer 2.2** Ponovno razmatramo prethodno opisani eksperiment izvlačenja jedne kuglice iz kutije. Definirajmo dva događaja:

$$A = \{0,2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34\}$$

$$B = \{18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35\}$$

Događaj A predstavlja izvlačenje parnog broja. Događaj B predstavlja izvlačenje broja većeg ili jednakog 18. Odredite $P(A)$, $P(B)$, $P(A \cap B)$.

Rješenje. Skup elementarnih događaja koji ima 36 elemenata vizualiziran je na slici u nastavku. Na istoj slici označeni su i skupovi A (zeleno) i B (crveno).



Skup A sadrži 18 elementarnih događaja, pa uz činjenicu da je vjerojatnost svakog elementarnog događaja jednaka (u eksperimentu koji razmatramo), slijedi:

$$P(A) = \frac{18}{36} = \frac{1}{2}.$$

Skup B sadrži također 18 elementarnih događaja pa je:

$$P(B) = \frac{18}{36} = \frac{1}{2}.$$

Uvidom na sliku vidimo da je skup $A \cap B = \{18, 20, 22, 24, 26, 28, 30, 32, 36\}$, pa je:

$$P(A \cap B) = \frac{9}{36} = \frac{1}{4}.$$

Vidimo i da vrijedi:

$$P(A \cap B) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A) \cdot P(B)$$

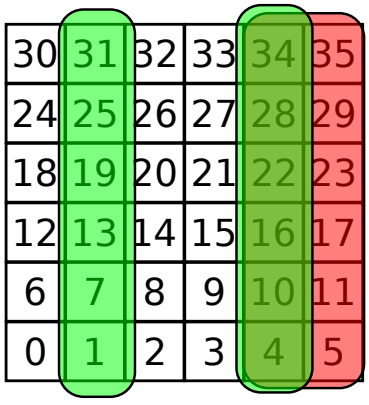
pa su događaji A i B nezavisni događaji.

Evo još par primjera međusobno nezavisnih događaja:

- "izvučen je paran broj" i "izvučen je broj 30 ili veći",
- "izvučen je neparan broj" i "izvučen je broj 23 ili manji",
- "izvučen je paran broj" i "izvučen je broj između 12 i 23" (uz 12 i 23 uključene).

Odredite za svaki od ovih primjera vjerojatnost nastupanja pojedinih događaja te oba događaja. ■

Kako bi izgledao primjer događaja koji nisu nezavisni? Koristeći i dalje eksperiment s izvlačenjem kuglica, pogledajmo sljedeća dva događaja ilustrirana slikom u nastavku.



Događaj A bi odgovarao izvlačenju broja čiji je ostatak djeljenja s 3 jednak 1 (zeleno na slici). Događaj B (crveno na slici) bi odgovarao broju koji je u desna dva stupca (da ne izmišljamo neki pametniji opis). Elementi skupa A pokrivaju 2 od 6 stupaca, pa je vjerojatnost P(A) jednaka:

$$P(A) = \frac{2}{6} = \frac{1}{3}$$

Elementi skupa B također pokrivaju 2 od 6 stupaca, pa je vjejatnost P(B) jednaka:

$$P(B) = \frac{2}{6} = \frac{1}{3}$$

Istovremeno nastupanje događaja A i B opisano je presjekom njihovih skupova i taj skup na slici su pokriva samo predzadnji od 6 stupaca; stoga je:

$$P(A,B) = \frac{1}{6}.$$

Sada vidimo da ne vrijedi da je umnožak $P(A) \cdot P(B) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$ jednak vjerojatnosti nastupanja oba događaja koja je $\frac{1}{6}$. Činjenica da A i B nisu nezavisni trebala bi biti i intuitivno jasna: događaj A u čitavom prostoru elementarnih događaja pokriva trećinu elementarnih događaja; međutim, ako razmotrimo samo elementarne događaje koje pokriva B, u njemu čak polovica elementarnih događaja ispunjava događaj A. Stoga je očito da, ako znamo da je nastupio događaj B, mijenja se vjerojatnost da je nastupio i događaj A, pa ova dva događaja nisu međusobno nezavisna.

Ako razmatramo više od dva događaja, tada treba biti oprezan s terminologijom jer je moguće da, primjerice, u skupu od tri događaja A, B i C, svi u parovima budu nezavisni A i B, A i C, B i C, ali da zajedno ne budu, tj. moguće je da vrijedi $P(A,B) = P(A) \cdot P(B)$, $P(A,C) = P(A) \cdot P(C)$, $P(B,C) = P(B) \cdot P(C)$, ali da ne vrijedi $P(A,B,C) = P(A) \cdot P(B) \cdot P(C)$. Stoga ako kažemo da su događaji nezavisni, podrazumijevati ćemo u svim kombinacijama (dvojkama, trojkama, ...), a ako kažemo da su međusobno nezavisni, podrazumijevat ćemo poseban slučaj gdje vrijedi samo nezavisnost u parovima, ali ne i šire.

2.1.2 Uvjetna vjerojatnost

Uvjetna vjerojatnost ograničava prostor elementarnih događaja na neki podskup. Vratimo se ponovno na prethodnu sliku gdje smo imali događaje A (ostatak dijeljenja jednak 1) i B (dva najdesnija stupca). Vjerojatnost nastupanja događaja A jednaka je omjeru kardinaliteta pridruženog skupa elementarnih događaja te kardinaliteta čitavog prostora elementarnih događaja; u našem slučaju to je bilo:

$$P(A) = \frac{|A|}{|S|} = \frac{12}{36} = \frac{1}{3}.$$

Uvjetna vjerojatnost ograničava elementarne događaje koje razmatramo. Vjerojatnost da nastupi događaj A ako je nastupio događaj B (dakle, uvjetovano nastupanjem događaja B) označit ćemo s P(A|B) i ona će (u našem primjeru gdje su svi elementarni događaji jednako vjerojatni) biti jednaka broju elementarnih događaja koji ispunjavaju događaj A razmotrenih samo na podskupu elementarnih događaja koji ispunjavaju B, podijeljen kardinalitetom skupa B. Primijetite da je to zapravo omjer kardinaliteta presjeka A i B, i kardinaliteta skupa B:

$$P(A|B) = \frac{|A \cap B|}{|B|}$$

U našem konkretnom primjeru, kardinalitet skupa B je 12 (dva najdesnija stupca). Od tih 12 elemenata, 6 (predzadnji stupac) ih ispunjava i događaj A. Stoga je:

$$P(A|B) = \frac{6}{12} = \frac{1}{2}$$

Na istom primjeru vidimo i da je primjerice:

$$P(A|B^c) = \frac{6}{24} = \frac{1}{4}$$

Općenito, P(A|B) i P(B|A) ne moraju biti jednake. Zamislite samo da je A pokrivao 1., 2. i 5. stupac. tada bi bilo $P(A|B) = \frac{1}{2}$ ali $P(B|A) = \frac{1}{3}$.

Vjerojatnost istovremenog nastupanja događaja A i B možemo zapisati i preko uvjetnih vjerojatnosti:

$$P(A,B) = \frac{|A \cap B|}{|S|} = \frac{|A \cap B|}{|S|} \cdot \frac{|B|}{|B|} = \frac{|A \cap B|}{|B|} \cdot \frac{|B|}{|S|} = P(A|B) \cdot P(B)$$

odnosno analogno:

$$P(A,B) = \frac{|A \cap B|}{|S|} = \frac{|A \cap B|}{|S|} \cdot \frac{|A|}{|A|} = \frac{|A \cap B|}{|A|} \cdot \frac{|A|}{|S|} = P(B|A) \cdot P(A)$$

čime smo došli do vrlo važnog izraza:

$$P(A,B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{2.1}$$

Ovaj izraz, samo malo drugačije zapisan, zapravo je poznati Bayesov teorem.

Theorem 2.3 — Bayesov teorem. Neka su A i B dva događaja. Vrijedi:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}.$$

Sada možemo pokazati još jednu važnu relaciju koja vrijedi. Već smo pokazali da je, primjerice:

$$P(A|B) = \frac{P(A,B)}{P(B)}.$$

Ako su događaji A i B međusobno nezavisni, tada je $P(A,B) = P(A) \cdot P(B)$, pa ako to uvrstimo, dobivamo:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A).$$

Drugim riječima (i ovo je alternativna definicija nezavisnosti događaja), ako su A i B nezavisni događaji, onda činjenica da znamo da je nastupio jedan od njih ne mijenja vjerojatnost da je nastupio drugi: $P(A|B) = P(A)$ i $P(B|A) = P(B)$.

Konačno, za dva događaja A i B kažemo da su uvjetno nezavisni s obzirom na C, ako vrijedi:

$$P(A,B|C) = P(A|C) \cdot P(B|C).$$

2.1.3 Zakon ulančavanja

Posljednje na što ćemo se pozvati kod Byesovog klasifikatora jest izraz koji nam govori kako možemo odrediti vjerojatnost istovremenog nastupanja više od jednog događaja. Prisjetimo se, u slučaju dva događaja mogli smo pisati:

$$P(A,B) = P(A|B) \cdot P(B).$$

Ako razmatramo tri događaja, vrijedi:

$$\begin{aligned} P(A,B,C) &= P(A|B,C)\cdot P(B,C) \\ &= P(A|B,C)\cdot P(B|C)\cdot P(C). \end{aligned}$$

U slučaju četiri događaja imamo:

$$\begin{aligned} P(A,B,C,D) &= P(A|B,C,D)\cdot P(B,C,D) \\ &= P(A|B,C,D)\cdot P(B|C,D)\cdot P(C,D) \\ &= P(A|B,C,D)\cdot P(B|C,D)\cdot P(C|D)\cdot P(D). \end{aligned}$$

Na analogan način možemo raspisati izraz za proizvoljan broj događaja.

2.2 Naivni Bayesov klasifikator

Na sportskom ste natjecanju i navijač ste domaćeg tima. Vrijeme je hladno i poprilici 15% navijača nosi šal. Među svim navijačima koji su došli popratiti natjecanje, 10% su navijači protivničkog tima. Međutim, poznato je da je jedan od njihovih tradicionalnih rekvizita upravo šal: čak 80% navijača protivničkog tima ga nosi. Ako vidite navijača koji nosi šal, možete li zaključiti da se radi o navijaču protivničkog tima?

Odgovor je, dakako, ne. Naime, pitanje je ekvivalentno onoj poznatoj priči: ako pada kiša, ceste su mokre. Ako vidimo da su ceste mokre, možemo li zaključiti da pada kiša? Pravilo zaključivanja koje to omogućava zove se abdukcija, i znamo da to nije ispravno pravilo zaključivanja. Možemo li, međutim, pojedina opažanja i zaključke na odgovarajući način kvantificirati s određenim mjerama, htjet ćemo dopustiti i uporabu abdukcije kao pravila zaključivanja. Mjera o kojoj je ovdje riječ bit će vjerojatnost.

Vratimo se na početni primjer. Eksperiment koji radimo je gledanje navijača. Pitamo se je li navijač kojeg gledamo navijač protivničkog tima ili nije. Definirat ćemo dva događaja koje ćemo ujedno zvati i hipotezama: s H_{JE} ćemo uznačiti hipotezu "navijač je navijač protivničkog tima", a s H_{NIJE} ćemo uznačiti hipotezu "navijač nije navijač protivničkog tima". Iz početnog opisa znamo da 10% navijača navija za protivnički tim:

$$P(H_{\text{JE}}) = 0.1.$$

Znamo također da 15% svih navijača nosi šal:

$$P(\text{nosi šal}) = 0.15$$

te da 80% navijača protivničkog tima nosi šal - čime je zapravo iskazana uvjetna vjerojatnost:

$$P(\text{nosi šal}|H_{\text{JE}}) = 0.8$$

Zadatak nas traži da odredimo vjerojatnost da je navijač kojeg smo sreli i koji nosi šal navijač protivničkog tima. Sjetimo li se da je vjerojatnost istovremenog nastupanja oba događaja jednaka:

$$P(\text{nosi šal}, H_{\text{JE}}) = P(\text{nosi šal}|H_{\text{JE}})\cdot P(H_{\text{JE}}) = P(H_{\text{JE}}|\text{nosi šal})\cdot P(\text{nosi šal})$$

iz čega možemo odrediti traženu uvjetnu vjerojatnost $P(H_{\text{JE}}|\text{nosi šal})$:

$$P(H_{\text{JE}}|\text{nosi šal}) = \frac{P(\text{nosi šal}|H_{\text{JE}})\cdot P(H_{\text{JE}})}{P(\text{nosi šal})} = \frac{0.8\cdot 0.1}{0.15} = 0.53$$

Vjerojatnost da gledamo navijača protivničkog tima je dakle preko 50%.

Činjenicu da navijač nosi šal općenito ćemo zvati uočenim dokazima, i označavati velikim slovom E.

Pogledajmo još jedan primjer: liječnik obrađuje niz pacijenata i sve bilježi u kartonima. Nakon mnoštva obrađenih pacijenata, liječnik je identificirao niz bolesti: gripa, prehlada, upala pluća, meningitis, itd. Na temelju svih obrađenih pacijenata i pretpostavke da niti jedan pacijent nije istovremeno imao više od jedne bolesti, liječnik može procijeniti kolika je vjerojatnost da pacijent ima svaku od bolesti tako da podijeli broj pacijenata koji su imali promatranu bolest s ukupnim brojem obrađenih pacijenata.

Za svakog pacijenta, liječnik iz zapisa može odrediti i koliko je pacijenata imalo temperaturu, koliko ih je imalo začepljen nos, koliko ih je imalo upaljeno grlo i slično, te ponovno dijeljenjem s ukupnim brojem pacijenata može procijeniti kolika je vjerojatnost da pacijent ima temperaturu, kolika je vjerojatnost da ima začepljen nos, kolika je vjerojatnost da ima upaljeno grlo i slično.

Međutim, iz podataka kojima raspolaže, liječnik može odrediti i niz uvjetnih vjerojatnosti: od svih pacijenata koji su imali gripu, koliko ih je imalo temperaturu (što omogućava procjenu P(temperatura|gripa)), koliko ih je imalo začepljen nos (što omogućava procjenu P(začepljen nos|gripa)), koliko ih je imalo upaljeno grlo (što omogućava procjenu P(upaljeno grlo|gripa)), te analogne procjene može izvući i za sve ostale bolesti.

Opažanja poput *Ima temperaturu*, *ima začepljen nos* i *Ima upaljeno grlo* zvat ćemo dokazima. Primijetite da u stvarnom životu, bolest uzrokuje opažene simptome; obrat ne vrijedi. Prehlada kao jedan od simptoma može imati upaljeno grlo. Upaljeno grlo međutim nije uzrok prehlade. Izjave "pacijent boluje od prehlade", "pacijent boluje od gripe" i slično, zvat ćemo hipotezama.

Ako imam jedan dokaz E i skup od m međusobno isključivih hipoteza $\{H_j\}_{j=1}^{m}$ (koje pokrivaju čitav skup elementarnih događaja), tada prema pojašnjenju iz prethodnog poglavlja možemo pisati:

$$P(E) = \sum_{j=1}^{m} P(E, H_j) = \sum_{j=1}^{m} P(H_j) \cdot P(E|H_j)$$

prva jednakost vrijedi jer pretpostavljamo da su hipoteze međusobno isključive te da pokrivaju čitav prostor događaja. Druga jednakost raspisuje vjerojatnost para preko uvjetne vjerojatnosti. Ono što nas zanima jest, ako smo opazili dokaz E, kolika je vjerojatnost da pacijent ima određenu bolest (odnosno da je nastupio događaj H_i). Istu prema Bayesovom teoremu možemo izračunati kao:

$$P(H_i|E) = \frac{P(E|H_i) \cdot P(H_i)}{P(E)} = \frac{P(E|H_i) \cdot P(H_i)}{\sum_{j=1}^{m} P(H_j) \cdot P(E|H_j)}$$

Da bismo ovo mogli izračunati, trebamo samo vjerojatnosti svake od hipoteza (drugim riječima, bolesti), te uvjetne vjerojatnosti dokaza uz svaku od hipoteza (drugim riječima, opaženog simptoma po svakoj od bolesti). To su sve podatci koje liječnik iz zapisa pacijenata može procijeniti.

U praksi uobičajeno razmatramo situaciju u kojoj imamo više dokaza: imamo pacijenta koji ima temperaturu (dokaz E_1) i upaljeno grlo (dokaz E_2). Zanima nas kolika je vjerojatnost da ima određenu bolest:

$$P(H_i|E_1, E_2) = \frac{P(E_1, E_2|H_i) \cdot P(H_i)}{P(E_1, E_2)} = \frac{P(E_1, E_2|H_i) \cdot P(H_i)}{\sum_{j=1}^{m} P(H_j) \cdot P(E_1, E_2|H_j)}.$$

Da bismo ovo izračunali, sada nam treba puno više podataka: primjerice, za svaku od bolesti, trebamo vjerojatnost da su istovremeno uočena oba simptoma. kako općenito možemo imati proizvoljan broj simptoma, broj podataka koji trebamo da bismo odredili točne vjerojatnosti postaje nedostupan. Stoga ćemo uvesti "naivnu" pretpostavku da su svi dokazi međusobno uvjetno neovisni

s obzirom na hipoteze; drugim riječima, da vrijedi $P(E_1, E_2|H_i) = P(E_1|H_i) \cdot P(E_2|H_i)$. Uz tu pretpostavku, prethodni izraz se pretvara u:

$$P(H_i|E_1,E_2) = \frac{P(E_1|H_i) \cdot P(E_2|H_i) \cdot P(H_i)}{P(E_1,E_2)} = \frac{P(E_1|H_i) \cdot P(E_2|H_i) \cdot P(H_i)}{\sum_{j=1}^{m} P(H_j) \cdot P(E_1|H_j) \cdot P(E_2|H_j)}.$$

odnosno u najopćenitijem slučaju za n dokaza i m hipoteza:

$$P(H_i|E_1,\ldots,E_n) = \frac{P(H_i)\prod_{k=1}^{n} P(E_k|H_i)}{P(E_1,\ldots,E_n)} = \frac{P(H_i)\prod_{k=1}^{n} P(E_k|H_i)}{\sum_{j=1}^{m} P(H_j)\prod_{k=1}^{n} P(E_k|H_j)}. \tag{2.2}$$

Upoznajmo se još i s terminologijom koju ćemo koristiti u nastavku.

- $P(H_i)$ ćemo zvati apriorna vjerojatnost hipoteze H_i; to je naprosto vjerojatnost da hipoteza vrijedi prije no što dobijemo bilo kakve dokaze.
- $P(H_i|E)$ je aposteriorna vjerojatnost hipoteze H_i nakon što smo dobili dokaz E; drugim riječima, ako imamo dokaz E, kolika je vjerojatnost na vrijedi hipoteza H_i.
- $P(E|H_i)$ nazivamo izglednost dokaza E uz danu hipotezu H_i.

Da bismo došli do klasifikatora, preostalo je još samo postaviti posljednje pitanje: s obzirom na dokaze koje smo opazili, od koje bolesti boluje promatrani pacijent? Ovo je klasifikacijski problem u kojem pacijenta treba smjestiti u jedan od razreda.

Prvi način kako to možemo učiniti jest da prema izrazu (2.2) za sve hipoteze odredimo aposteriorne vjerojatnosti, te za pacijenta odaberemo onu hipotezu koja ima maksimalnu aposteriornu vjerojatnost; tu hipotezu nazivamo H_{MAP} i računamo prema izrazu:

$$H_{\text{MAP}} = \arg\max_{H_i} P(H_i|E_1,\ldots,E_n) = \arg\max_{H_i} \frac{P(H_i)\prod_{k=1}^{n} P(E_k|H_i)}{\sum_{j=1}^{m} P(H_j)\prod_{k=1}^{n} P(E_k|H_j)}.$$

Uočimo li da je nazivnik razlomka zapravo vjerojatnost $P(E_1,\ldots,E_n)$, i da je stoga isti za sve hipoteze čije vjerojatnosti računamo, H_{MAP} možemo odrediti i malo učinkovitije ne računajući nepotrebno nazivnik:

$$H_{\text{MAP}} = \arg\max_{H_i} P(H_i|E_1,\ldots,E_n) = \arg\max_{H_i} P(H_i)\prod_{k=1}^{n} P(E_k|H_i). \tag{2.3}$$

U posebnom slučaju da su sve hipoteze jednako vjerojatne, tj. $P(H_1) = P(H_2) = \cdots = P(H_m)$, iz izraza (2.3) možemo izbaciti i množenje s apriornim vjerojatnostima. U tom slučaju, dobivenu hipotezu zovemo hipotezom maksimalne izglednosti i računamo prema izrazu:

$$H_{\text{ML}} = \arg\max_{H_i} P(H_i|E_1,\ldots,E_n) = \arg\max_{H_i} \prod_{k=1}^{n} P(E_k|H_i). \tag{2.4}$$

2.2.1 Primjer klasifikatora na skupu *Dan za sport*

Jedan od čestih primjera na kojem se analizira izgradnja naivnog Bayesovog klasifikatora jest skup uzoraka *Dan za sport*. Skup uzoraka *Dan za sport* prikazan je tablicom u nastavku.

Redni broj	Vrijeme	Temperatura	Vlažnost	Vjetar	Igra
1.	sunčano	vruće	velika	slab	NE
2.	sunčano	vruće	velika	jak	NE
3.	oblačno	vruće	velika	slab	DA
4.	kišno	ugodno	velika	slab	DA
5.	kišno	hladno	normalna	slab	DA
6.	kišno	hladno	normalna	jak	NE
7.	oblačno	hladno	normalna	jak	DA
8.	sunčano	ugodno	velika	slab	NE
9.	sunčano	hladno	normalna	slab	DA
10.	kišno	ugodno	normalna	slab	DA
11.	sunčano	ugodno	normalna	jak	DA
12.	oblačno	ugodno	velika	jak	DA
13.	oblačno	vruće	normalna	slab	DA
14.	kišno	ugodno	velika	jak	NE

Uzorci su definirani kao uređene četvorke (vrijeme,temperatura,vlažnost,vjetar) koje klasificiramo u dva razreda: DA i NE. Pretpostavimo sada da imamo uzorak (kišno,vruće,velika,jak); bi li to bio dobar dan za baviti se sportom ili ne? Uvidom u prethodnu tablicu vidimo da taj uzorak nije sadržan u tablici, pa ne možemo pročitati u koji bismo razred uzorak trebali smjestiti. Stoga ćemo primijeniti izraz (2.3) kako bismo odredili koja je od hipoteza najizglednija: hipoteza H_{DA} ili hipoteza H_{NE}. Što su naši opaženi dokazi? Atribut *vrijeme* imao je vrijednost "kišno", atribut *temperatura* imao je vrijednost "vruće", atribut *vlažnost* imao je vrijednost "velika" i atribut *vjetar* imao je vrijednost "jak". Prilagodimo li izraz (2.3) ovom konkretnom slučaju, imamo:

$$H_{\text{MAP}} = \arg\max_{H_i} P(H_i|\text{vrijeme=kišno}, \text{temperatura=vruće}, \text{vlažnost=velika}, \text{vjetar=jak})$$

Stoga trebamo odrediti apriorne vjerojatnosti obje hipoteze. To radimo uvidom u skup podataka kojim raspolažemo. Skup ima 14 uzoraka pri čemu ih je 9 u razredu DA, a 5 u razredu NE. Ovo ćemo iskoristiti da bismo procijenili apriorne vjerojatnosti hipoteza:

$$P(H_{DA}) = \frac{9}{14},$$

$$P(H_{NE}) = \frac{5}{14}.$$

Trebat ćemo i 8 aposteriornih vjerojatnosti koje također možemo odrediti brojanjem po uzorcima:

$$P(\text{vrijeme=kišno}|H_{DA}) = \frac{3}{9},$$

$$P(\text{temperatura=vruće}|H_{DA}) = \frac{2}{9},$$

$$P(\text{vlažnost=velika}|H_{DA}) = \frac{3}{9},$$

$$P(\text{vjetar=jak}|H_{DA}) = \frac{3}{9},$$

$$P(\text{vrijeme=kišno}|H_{NE}) = \frac{2}{5},$$

$$P(\text{temperatura=vruće}|H_{NE}) = \frac{2}{5},$$

$$P(\text{vlažnost=velika}|H_{NE}) = \frac{4}{5},$$

$$P(\text{vjetar=jak}|H_{NE}) = \frac{3}{5},$$

Primjerice, aposteriornu vjerojatnost $P(\text{vrijeme=kišno}|H_{DA})$ odredili smo kao omjer broja uzorka koji su u razredu DA i imaju vrijeme=kišno i broja uzoraka koji su u razredu DA.

Za razred DA aposteriorna vjerojatnost proporcionalna je:

$$P(H_{DA})\cdot P(\text{vrijeme=kišno}|H_{DA})\cdot P(\text{temperatura=vruće}|H_{DA})\cdot P(\text{vlažnost=velika}|H_{DA})\cdot P(\text{vjetar=jak}|H_{DA})$$

što je:

$$\frac{9}{14}\cdot\frac{3}{9}\cdot\frac{2}{9}\cdot\frac{3}{9}\cdot\frac{3}{9} = \frac{1}{189} \approx 0.00529$$

Za razred NE aposteriorna vjerojatnost proporcionalna je:

$$P(H_{NE})\cdot P(\text{vrijeme=kišno}|H_{NE})\cdot P(\text{temperatura=vruće}|H_{NE})\cdot P(\text{vlažnost=velika}|H_{NEA})\cdot P(\text{vjetar=jak}|H_{NE})$$

što je:

$$\frac{5}{14}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{4}{5}\cdot\frac{3}{5} = \frac{24}{875} \approx 0.02743$$

Kako je $\max(0.00353, 0.02743) = 0.02743$, MAP-hipoteza je H_{NE}, čime uzorak razvrstavamo u razred NE.

Programska implementacija u memoriji ne bi pamtila čitav skup uzoraka jer isti može biti ogroman. Umjesto toga, u fazi *učenja* klasifikatora na temelju uzoraka izračunala bi i u prikladnim podatkovnim strukturama (primjerice, mapama) zapamtila apriorne vjerojatnosti svakog od razreda, te aposteriorne vjerojatnosti vrijednosti svakog od atributa po razredima. U fazi uporabe (iskorištavanja), na temelju predanog uzorka iz mapa bi samo dohvatila potrebne podatke, izmnožila ih i odabrala MAP-hipotezu.

Isprobajte

Postupak klasifikacije Naivnim Bayesovim klasifikatorom možete isprobati i sami. U naredbenom retku zadajte naredbu:

```
java -cp book-ml.jar ml.bayes.Classifier sport.txt
```

Kroz argument zadajete definiciju datoteke sa skupom uzoraka za učenje. `sport.txt` je ugrađena datoteka. Pokrenut će se program i ispisati nekoliko informacija, kao i napomena da za izlazak zadate "exit". Program će čekati u interaktivnoj ljusci da upisujete uzorke koje želite klasificirati. Primjerice, ako upišete:

```
(kišno,vruće,velika,jak)
```

dobit ćete ispis i klasifikaciju za primjer koji smo analizirali i u tekstu.

Želite li klasifikator isprobati na vlastitom primjeru, možete i sami pripremiti svoju datoteku: u prvi redak upišite nazive atributa odvojene tabom (posljednji će se smatrati ciljnim), a u preostale retke upišite specifikacije uzoraka (navodite samo vrijednosti atributa i također ih razdvajate tabom). Potom prilikom pokretanja programa zadajte stazu do Vaše datoteke.

3. Stabla odluke

Stabla odluke su formalizam koji omogućava rješavanje klasifikacijskih te aproksimacijskih zadataka temeljeći se na slijedu ispitivanja vrijednosti atributa uzorka. U okviru ovog poglavlja razmotrit ćemo uporabu stabla odluke za klasifikacijske svrhe te njihovu izgradnju uporabom algoritma ID3.

Pogledajmo najprije skup uzoraka koji će nam biti na raspolaganju. Radi se o uzorcima koji opisuju oštećenje rotacijskog elementa određenog stroja te na temelju karakteristika oštećenja definiraju je li popravak elementa hitan ili ne. Skup uzoraka koji nam je na raspolaganju prikazan je u tablici 3.1.

Tablica 3.1: Skup uzoraka: *Hitnost popravka rotacijskog elementa*.

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
2.	srednje	bliži	svijetlo	DA
3.	malo	dalji	svijetlo	NE
4.	veliko	bliži	svijetlo	DA
5.	veliko	dalji	tamno	DA
6.	veliko	dalji	svijetlo	DA
7.	srednje	dalji	svijetlo	NE
8.	veliko	bliži	tamno	DA
9.	srednje	dalji	tamno	NE
10.	srednje	bliži	tamno	DA
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

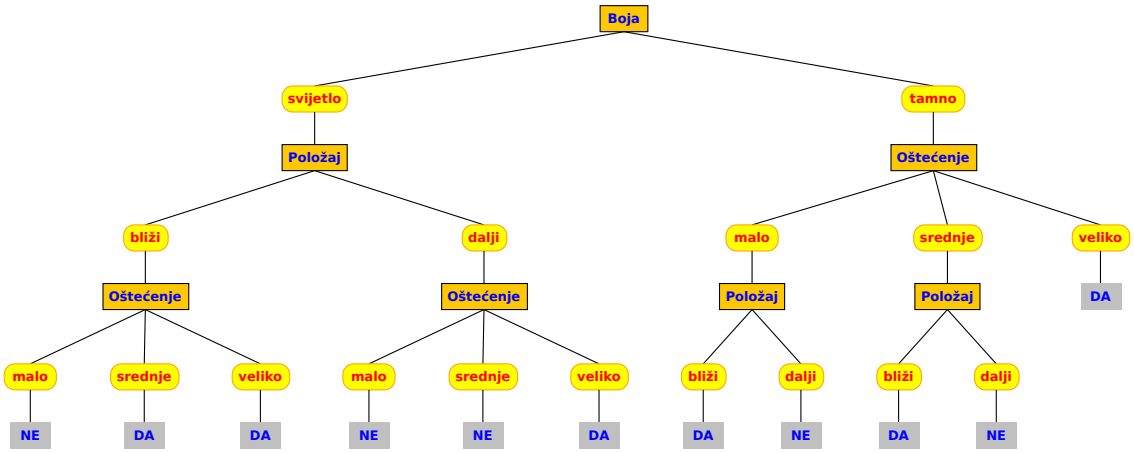
Svaki uzorak opisan je s tri atributa: *Oštećenje*, *Položaj* i *Boja*. Ove atribute zvat ćemo još i značajkama (engl. *features*) uzorka. Uzorke razvrstavamo u dva razreda: one koje je potrebno hitno popraviti te one kod kojih popravak nije hitan. Posljednji stupac tablice 3.1 sadrži ovu informaciju:

ako je vrijednost DA, rotacijski element potrebno je hitno popraviti. Ako je vrijednost NE, popravak nije hitan. Evo i detaljnijeg opisa svakog od atributa.

- *Oštećenje* - specificira kolika je veličina uočenog oštećenja; vrijednosti koje atribut poprima su *malo*, *srednje* te *veliko*.
- *Položaj* - specificira gdje se nalazi oštećenje s obzirom na centar rotacije elementa; vrijednosti koje atribut poprima su *bliži* i *dalji*.
- *Boja* - specificira nijansu boje uočenog oštećenja; vrijednosti koje atribut poprima su *svijetlo* i *tamno*.

Atribut *Hitno* koji je prikazan u posljednjem stupcu tablice još nazivamo i ciljnim atributom. Njegove su vrijednosti u ovom primjeru *DA* i *NE*, a razmatrat ćemo klasifikatorske sustave koji će promatranom uzorku na temelju vrijednosti atributa *Oštećenje*, *Položaj* i *Boja* pokušati dodijeliti ispravnu vrijednost ciljnog atributa.

Pogledajmo sada jedno moguće stablo odluke. Slika 3.1 prikazuje stablo odluke koje sve uzorke iz skupa uzoraka iz tablice 3.1 klasificira korektno.



Slika 3.1: Primjer stabla odluke za skup *Hitnost popravka rotacijskog elementa*.

Korijen stabla je čvor u kojem se ispituje vrijednost atributa *Boja*. Čvor ima dvije grane jer atribut boja može poprimiti dvije vrijednosti. Lijevu granu potrebno je slijediti ako uzorak koji ispitujemo ima *svijetlo* kao vrijednost atributa *Boja*, a desnu ako uzorak koji ispitujemo ima *tamno* kao vrijednost atributa *Boja*. U oba slučaja dolazimo do podstabla koje obrađujemo na analogan način. Iznimka su listovi stabla u kojima se ne ispituje vrijednost nekog od atributa već donosi konačna odluka o razredu kojem uzorak pripada. Ti su čvorovi na slici prikazani kao sivi pravokutnici u koje je upisana vrijednost ciljnog atributa.

Kako bismo uporabom prikazanog stabla odluke zaključili koja je vrijednost ciljnog atributa za uzorak (veliko,bliži,tamno)? Korijen stabla sadrži čvor koji ispituje atribut *Boja*; stoga gledamo koja je boja u našem uzorku koji klasificiramo (*Boja*=tamno), pa u stablu slijedimo desnu granu. Dolazimo do čvora koji ispituje vrijednost atributa *Oštećenje*; stoga gledamo koju vrijednost ima taj atribut u našem uzorku koji klasificiramo (*Oštećenje*=veliko), pa u stablu ponovno slijedimo desnu granu. Time smo došli do lista koji uzorku pridjeljuje vrijednost DA kao vrijednost ciljnog atributa, odnosno uzorak razvrstava u razred uzoraka koje je potrebno hitno popraviti.

Primijetite da je utvrđeni razred u skladu s onime što je za uzorak (veliko,bliži,tamno) definirano u tablici 3.1. Primijetite također da za donošenje ispravne odluke nismo morali razmotriti vrijednosti svih atributa uzorka koji klasificiramo.

Prilikom izgradnje stabala odluke htjet ćemo da izgrađeno stablo ima određene karakteristike. Neiznenađujuće, za početak ćemo htjeti da isto u većini slučajeva ispravno radi svoj posao na skupu

uzoraka na temelju kojeg je izgrađeno, odnosno da uzorcima pridjeljuje vrijednosti ciljnog atributa kako je specificirano u uzorcima za učenje. Ako se pitate zašto ne u svim slučajevima, odgovor leži u činjenici da želimo stabla koja dobro generaliziraju, pa ako za neke uzorke utvrdimo da su *stršeće vrijednosti* (engl. *outliers*), neće nas smetati što će konstruirano stablo njima pridijeljivati drugačiju vrijednost ciljnog atributa od one koja je zapisana u skupu uzoraka za učenje.

Od svih stabala koja su usklađena s prethodnim zahtjevom, preferirat ćemo ona koja su manja, odnosno koja imaju manji broj čvorova.

Klasifikatorska stabla možemo graditi na različite načine. U ovom poglavlju, upoznat ćemo se algoritmom ID3 (engl. *Iterative Dichotomiser 3*) kojeg je osmislio Ross Quinlan, a kako se ovaj postupak temelji na pojmovima entropije i informacijske dobiti, najprije ćemo se upoznati s tim pojmovima.

Razmotrimo četiri skupa uzoraka za učenje, čije su karakteristike specificirane u sljedećoj tablici.

Ime skupa	Ukupan broj uzoraka	Broj uzoraka u razredu DA	Broj uzoraka u razredu NE
$\mathscr{S}_1$	10	10	0
$\mathscr{S}_2$	10	3	7
$\mathscr{S}_3$	10	5	5
$\mathscr{S}_4$	10	0	10

Entropija je mjera neuređenosti skupa. Intuitivno, mogli bismo razmišljati ovako: kada bismo iz skupa nasumice izvlačili uzorke i pokušavali pogoditi (bez uvida u atribute izvučenog uzorka) kojem će razredu pripadati atribut, što je mjera neuređenosti skupa veća, više bismo griješili pri pogađanju.

Pogledajmo skup $\mathscr{S}_1$: on sadrži 10 elemenata, i svi pripadaju razredu DA. Uz pretpostavku da tu informaciju znamo, svaki slučajno izvučeni uzorak savršeno bismo klasificirali uvijek pridjeljujući uzorke u razred DA. Ovaj skup ima minimalnu neuređenost. Analogno vrijedi i za skup $\mathscr{S}_4$ kod kojeg su svi uzorci smješteni u razred NE; i ovaj skup ima minimalnu neuređenost.

Kod razreda $\mathscr{S}_2$ i $\mathscr{S}_3$ nećemo više moći uvijek korektno pogađati ispravan razred izvučenog uzorka. Pri tome bismo s razredom $\mathscr{S}_2$ mogli proći bolje no s razredom $\mathscr{S}_3$. U razredu $\mathscr{S}_2$ 70% uzoraka pripada razredu NE; stoga, ako svakom izvučenom uzorku iz razreda $\mathscr{S}_2$ uvijek pridijelimo oznaku NE, u prosjeku ćemo griješiti u 30% slučajeva. Za razliku od toga, u razredu $\mathscr{S}_3$ po 50% uzoraka pripada razredima DA odnosno NE; stoga ako ćemo uvijek uzorke svrstavati u razred DA (ili NE), griješit ćemo čak u 50% slučajeva.

Uzevši ova zapažanja u obzir, mogli bismo reći da je mjera neuređenosti skupa $\mathscr{S}_3$ veća od mjere neuređenosti skupa $\mathscr{S}_2$ koja je pak veća od mjera neuređenosti skupova $\mathscr{S}_1$ i $\mathscr{S}_4$, koje su minimalne moguće.

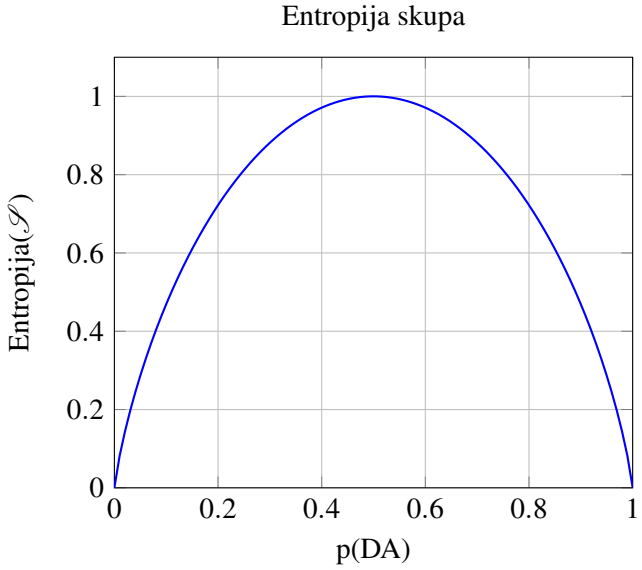
Entropiju skupa $\mathscr{S}$ definiramo s obzirom na vjerojatnosti da slučajno izvučeni uzorak pripada pojedinom razredu. Mjera je definirana kao negativna suma vjerojatnosti da uzorak pripada određenom razredu pomnoženoj dualnim logaritmom iste. U našem konkretnom slučaju imamo dva razreda: DA i NE. Stoga će entropija biti određena izrazom:

$$\text{Entropija}(\mathscr{S}) = -p(\text{DA}) \cdot \log_2 (p(\text{DA})) - p(\text{NE}) \cdot \log_2 (p(\text{NE}))$$

Ako je vjerojatnost nekog razreda 0, odnosno trebamo izračunati $0 \cdot log_2 0$, uzimat ćemo da je to jednako 0. Za skupove $\mathscr{S}_1$ do $\mathscr{S}_4$ entropije izračunate prema prethodnom izrazu prikazane su u nastavku.

Ime skupa	$p(DA)$	$p(NE)$	Entropija($\mathscr{S}_i$)
$\mathscr{S}_1$	$\frac{10}{10} = 1$	$\frac{0}{10} = 0$	0
$\mathscr{S}_2$	$\frac{3}{10} = 0.3$	$\frac{7}{10} = 0.7$	0.8813
$\mathscr{S}_3$	$\frac{5}{10} = 0.5$	$\frac{5}{10} = 0.5$	1
$\mathscr{S}_4$	$\frac{0}{10} = 0$	$\frac{10}{10} = 1$	0

U skladu s našim očekivanjem, vidimo da je entropija minimalna za skupove $\mathscr{S}_1$ i $\mathscr{S}_4$ i taj minimum je 0, entropija je veća za skup $\mathscr{S}_2$, a najveća za skup $\mathscr{S}_3$. Kako se iznos entropije mijenja ovisno o vjerojatnosti da uzorak pripada razredu DA prikazano je na slici 3.2, gdje vidimo da je maksimalni iznos koji entropija poprima jednak 1 (ovo vrijedi za slučaj binarne klasifikacije).



Slika 3.2: Entropija skupa u kojem uzorci pripadaju jednom od dva razreda: DA i NE, s obzirom na vjerojatnost da uzorak pripada razredu DA. Uočimo da je u ovom slučaju $p(NE) = 1 - p(DA)$.

Općenito, ako uzorci mogu pripadati jednom od K razreda, entropija skupa definira se izrazom u nastavku koji je analogan prethodnoj definiciji samo poopćen na više razreda:

$$\text{Entropija}(\mathscr{S}) = -\sum_{c \in \mathscr{C}} p(\text{c}) \cdot \log_2 (p(\text{c})) \tag{3.1}$$

gdje smo s $\mathscr{C}$ označili skup svih razreda, pa prikazana suma ide po razredima. $p(c)$ je vjerojatnost da uzorak pripada razredu c.

Ideja algoritma ID3 jest u svakom koraku razmotriti kolika je korist ako se razmatrani skup podataka podijeli po svakom od atributa. Zatim pohlepno napravi podjelu skupa uzoraka po tom atributu, stvori čvor s oznakom odabranog atributa te rekurzivno gradi podstabla nad napravljenim podskupovima. Mjera kvalitete podjele koju koristi ID3 je *informacijska dobit*: informacijska dobit jednaka je entropiji početnog skupa umanjenoj za entropije napravljenih podskupova skalirane omjerom veličine podskupa i početnog skupa. Intuitivno, ona nam govori koliko smo uspjeli smanjiti neuređenost podjelom skupa u manje podskupove. Možda zvuči komplicirano, no idemo pogledati na konkretnom primjeru skupa prikazanog u tablici 3.1.

U početnom skupu imamo 12 uzoraka. 7 od njih pripada razredu DA pa je $p(DA) = 7/12 = 0.5833$, a 5 ih pripada razredu NE pa je $p(NE) = 5/12 = 0.4167$. Entropija ovog skupa je dakle $-0.5833 * \log_2(0.5833) - 0.4167 * log_2(0.4167) = 0.9799$.

Uzorke ovog skupa možemo podijeliti prema vrijednostima triju atributa. Razmotrimo najprije podjelu po atributu *Oštećenje*. Ovaj atribut može poprimiti 3 različite vrijednosti, što znači da ćemo napraviti 3 podskupa.

Pogledajmo podskup koji čine svi uzorci koji imaju *Oštećenje*=malo. Taj podskup čine uzorci 1, 3, 11 i 12:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
3.	malo	dalji	svijetlo	NE
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

Skup ima 4 uzorka. 1 pripada razredu DA pa je $p(DA) = 1/4 = 0.25$, a 3 pripadaju razredu NE pa je $p(NE) = 3/4 = 0.75$. Stoga ovaj skup ima entropiju:

$$Entropija(\text{Oštećenje=malo}) = -0.25 * log_2(0.25) - 0.75 * log_2(0.75) = 0.8113.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (4) i početnog skupa (12), odnosno vrijednost 0.8113*4/12=0.2704.

Pogledajmo podskup koji čine svi uzorci koji imaju *Oštećenje*=srednje. Taj podskup čine uzorci 2, 7, 9 i 10:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
7.	srednje	dalji	svijetlo	NE
9.	srednje	dalji	tamno	NE
10.	srednje	bliži	tamno	DA

Skup ima 4 uzorka. 2 pripadaju razredu DA pa je $p(DA) = 2/4 = 0.5$, a 2 razredu NE pa je $p(NE) = 2/4 = 0.5$; primijetite da je skup maksimalno neuređen. Ovaj skup ima entropiju:

$$Entropija(\text{Oštećenje=srednje}) = -0.5 * log_2(0.5) - 0.5 * log_2(0.5) = 1.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (4) i početnog skupa (12), odnosno vrijednost $1 * 4/12 = 0.3333$.

Konačno, pogledajmo podskup koji čine svi uzorci koji imaju *Oštećenje*=veliko. Taj podskup čine uzorci 4, 5, 6 i 8:

Redni broj	Oštećenje	Položaj	Boja	Hitno
4.	veliko	bliži	svijetlo	DA
5.	veliko	dalji	tamno	DA
6.	veliko	dalji	svijetlo	DA
8.	veliko	bliži	tamno	DA

Skup ima 4 uzorka i sva četiri pripadaju razredu DA. Stoga je $p(DA) = 4/4 = 1$, a $p(NE) = 0/4 = 0$; primijetite da je skup maksimalno uređen; ima minimalnu neuređenost. Skup ima entropiju:

$$Entropija(\text{Oštećenje=veliko}) = -1 * log_2(1) - 0 * log_2(0) = 0.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (4) i početnog skupa (12), odnosno vrijednost 0*4/12=0.

Sada smo spremni izračunati informacijsku dobit podjele početnog skupa u tri podskupa prema vrijednostima atributa *Oštećenje*. Prisjetimo se, entropija početnog skupa veličine 12 bila je 0.9799. Vrijednosti koje smo izračunali za svaki od podskupova ponovljene su u tablici u nastavku.

Vrijednost atributa	Veličina podskupa	Entropija podskupa
malo	4	0.8113
srednje	4	1
veliko	4	0

Informacijska dobit ove podjele stoga je:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Oštećenje}) = \textit{Entropija}(\text{Početni skup}) \\ - \frac{4}{12} \cdot \textit{Entropija}(\text{Oštećenje=malo}) \\ - \frac{4}{12} \cdot \textit{Entropija}(\text{Oštećenje=srednje}) \\ - \frac{4}{12} \cdot \textit{Entropija}(\text{Oštećenje=veliko}) \end{aligned}$$

čime nakon uvrštavanja dobivamo:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Oštećenje}) &= 0.9799 - \frac{4}{12} \cdot 0.8113 - \frac{4}{12} \cdot 1 - \frac{4}{12} \cdot 0 \\ &= 0.9799 - 0.2704 - 0.3333 - 0 \\ &= 0.3761. \end{aligned}$$

Da bismo odlučili hoćemo li početni skup uzoraka doista razdijeliti prema vrijednostima atributa *Oštećenje*, trebamo pogledati i kolike su informacijske dobiti ako skup razdijelimo prema atributu *Položaj* odnosno prema atributu Boja. Proanalizirajmo najprije podjelu prema atributu *Položaj*.

Pogledajmo podskup koji čine svi uzorci koji imaju *Položaj*=bliži. Taj podskup čine uzorci 2, 4, 8, 10, 11 i 12:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
4.	veliko	bliži	svijetlo	DA
8.	veliko	bliži	tamno	DA
10.	srednje	bliži	tamno	DA
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

Skup ima 6 uzorka od kojih pet pripada razredu DA, a jedan razredu NE. Stoga je $p(DA) = 5/6 = 0.8333$, a $p(NE) = 1/6 = 0.1667$. Skup ima entropiju:

$$\textit{Entropija}(\text{Položaj=bliži}) = -0.8333 * log_2(0.8333) - 0.1667 * log_2(0.1667) = 0.65.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (6) i početnog skupa (12), odnosno vrijednost 0.65*6/12=0.325.

Pogledajmo podskup koji čine svi uzorci koji imaju *Položaj*=dalji. Taj podskup čine uzorci 1, 3, 5, 6, 7 i 9:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
3.	malo	dalji	svijetlo	NE
5.	veliko	dalji	tamno	DA
6.	veliko	dalji	svijetlo	DA
7.	srednje	dalji	svijetlo	NE
9.	srednje	dalji	tamno	NE

Skup ima 6 uzorka od kojih dva pripada razredu DA, a četiri razredu NE. Stoga je $p(DA) = 2/6 = 0.3333$, a $p(NE) = 4/6 = 0.6667$. Skup ima entropiju:

$$Entropija(\text{Položaj=dalji}) = -0.3333 * log_2(0.3333) - 0.6667 * log_2(0.6667) = 0.9183.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (6) i početnog skupa (12), odnosno vrijednost 0.9183*6/12=0.4591.

Informacijska dobit ove podjele stoga je:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Položaj}) = \textit{Entropija}(&\text{Početni skup}) \\ &- \frac{6}{12} \cdot \textit{Entropija}(\text{Položaj=bliži}) \\ &- \frac{6}{12} \cdot \textit{Entropija}(\text{Položaj=dalji}) \end{aligned}$$

čime nakon uvrštavanja dobivamo:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Položaj}) &= 0.9799 - \frac{6}{12} \cdot 0.65 - \frac{6}{12} \cdot 0.9183 \\ &= 0.9799 - 0.325 - 0.4591 \\ &= 0.1957. \end{aligned}$$

Konačno, razmotrimo i podjeli prema atributu *Boja*. Pogledajmo podskup koji čine svi uzorci koji imaju *Boja*=svijetlo. Taj podskup čine uzorci 2, 3, 4, 6, 7 i 11:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
3.	malo	dalji	svijetlo	NE
4.	veliko	bliži	svijetlo	DA
6.	veliko	dalji	svijetlo	DA
7.	srednje	dalji	svijetlo	NE
11.	malo	bliži	svijetlo	NE

Skup ima 6 uzorka od kojih tri pripadaju razredu DA i tri razredu NE. Stoga je $p(DA) = 3/6 = 0.5$, a $p(NE) = 3/6 = 0.5$. Skup ima entropiju:

$$Entropija(\text{Boja=svijetlo}) = -0.5 * log_2(0.5) - 0.5 * log_2(0.5) = 1.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (6) i početnog skupa (12), odnosno vrijednost 1*6/12=0.5.

Pogledajmo podskup koji čine svi uzorci koji imaju *Boja*=tamno. Taj podskup čine uzorci 1, 5, 8, 9, 10, 12:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
5.	veliko	dalji	tamno	DA
8.	veliko	bliži	tamno	DA
9.	srednje	dalji	tamno	NE
10.	srednje	bliži	tamno	DA
12.	malo	bliži	tamno	DA

Skup ima 6 uzorka od kojih četiri pripadaju razredu DA i dva razredu NE. Stoga je $p(DA) = 4/6 = 0.6667$, a $p(NE) = 2/6 = 0.3333$. Skup ima entropiju:

$$Entropija(\text{Boja=svijetlo}) = -0.6667 * log_2(0.6667) - 0.3333 * log_2(0.3333) = 0.9183.$$

Prilikom izračuna informacijske dobiti entropiji početnog skupa oduzet ćemo ovu vrijednost skaliranu omjerom veličine ovog skupa (6) i početnog skupa (12), odnosno vrijednost 0.9183*6/12=0.4591.

Informacijska dobit ove podjele stoga je:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Boja}) = \; & \textit{Entropija}(\text{Početni skup}) \\ & - \frac{6}{12} \cdot \textit{Entropija}(\text{Boja=svijetlo}) \\ & - \frac{6}{12} \cdot \textit{Entropija}(\text{Boja=tamno}) \end{aligned}$$

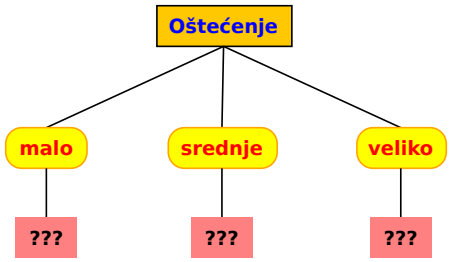
čime nakon uvrštavanja dobivamo:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Boja}) &= 0.9799 - \frac{6}{12} \cdot 1 - \frac{6}{12} \cdot 0.9183 \\ &= 0.9799 - 0.5 - 0.4591 \\ &= 0.0207. \end{aligned}$$

Koji ćemo od atributa uzeti za podjelu u korijenskom čvoru? Tablica u nastavku navodi utvrđene informacijske dobiti.

Podjela prema atributu	Informacijska dobit
Oštećenje	0.3761
Položaj	0.1957
Boja	0.0207

Najveća informacijska dobit (odnosno maksimalno smanjenje neuređenosti) postiže se za podjelu prema atributu *Oštećenje*. Stoga ćemo kao korijenski čvor stvoriti čvor koji analizira atribut *Oštećenje* i ima tri grane: malo, srednje i veliko; svaka od te tri grane analizirat će odgovarajući podskup uzoraka koji smo pripremili prilikom analize atributa *Oštećenje*. Djelomično stablo koje smo u ovom trenutku izgradili prikazano je u nastavku.



U lijevom čvoru ponavljamo postupak ID3, nad skupom uzoraka:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
3.	malo	dalji	svijetlo	NE
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

To su svi uzorci koji imaju "malo" kao vrijednost atributa *Oštećenje*. Broj uzoraka je 4; jedan pripada razredu DA, a tri pripadaju razredu NE. Stoga je $p(DA) = 1/4 = 0.25$, $p(NE) = 3/4 = 0.75$. Entropija ovog podskupa (koji za ovu granu predstavlja novi "početni skup") stoga je $-0.25 \cdot \log_2(0.25) - 0.75 \cdot log_2(0.75) = 0.8113$.

U ovoj grani trebamo razmotriti podjele ovog skupa samo prema atributima *Položaj* i *Boja*; vrijednost atributa *Oštećenje* svim je uzorcima fiksirana u roditeljskom čvoru pa se ovi uzorci po tom atributu ne razlikuju.

Razmotrimo podjelu prema atributu *Položaj*. Dobivamo dva podskupa; za *Položaj*=bliži imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

Veličina ovog podskupa je 2, a entropija 1. Za *Položaj*=dalji imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
3.	malo	dalji	svijetlo	NE

Veličina ovog podskupa je 2, a entropija 0. Informacijska dobit podjele prema atributu *Položaj* stoga je:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Položaj}) &= \textit{Entropija}(\text{Početni skup}) \\ &\quad - \frac{2}{4} \cdot \textit{Entropija}(\text{Položaj=bliži}) \\ &\quad - \frac{2}{4} \cdot \textit{Entropija}(\text{Položaj=dalji}) \\ &= 0.8113 - \frac{2}{4} \cdot 1 - \frac{2}{4} \cdot 0 \\ &= 0.8113 - 0.5 - 0 \\ &= 0.3113. \end{aligned}$$

Razmotrimo podjelu prema atributu *Boja*. Dobivamo dva podskupa; za *Boja*=svijetlo imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
3.	malo	dalji	svijetlo	NE
11.	malo	bliži	svijetlo	NE

Veličina ovog podskupa je 2, a entropija 0. Za *Boja*=tamno imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
1.	malo	dalji	tamno	NE
12.	malo	bliži	tamno	DA

Veličina ovog podskupa je 2, a entropija 1. Informacijska dobit podjele prema atributu *Položaj*

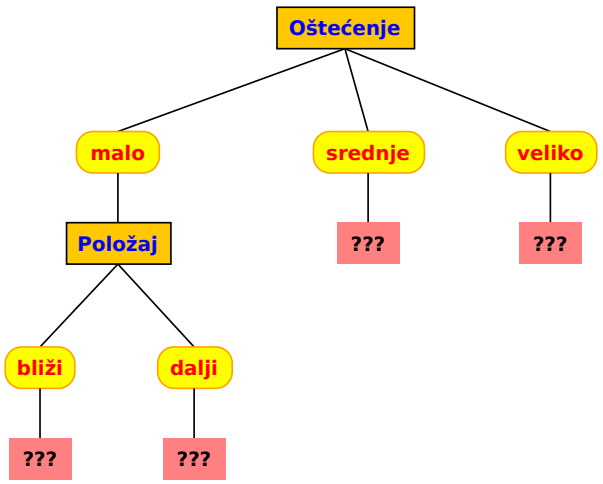
stoga je:

$$\begin{aligned} \textit{Informacijska Dobit}(\text{Boja}) &= \textit{Entropija}(\text{Početni skup}) \\ &\quad - \frac{2}{4} \cdot \textit{Entropija}(\text{Boja=svijetlo}) \\ &\quad - \frac{2}{4} \cdot \textit{Entropija}(\text{Boja=tamno}) \\ &= 0.8113 - \frac{2}{4} \cdot 0 - \frac{2}{4} \cdot 1 \\ &= 0.8113 - 0 - 0.5 \\ &= 0.3113. \end{aligned}$$

Koji ćemo sada atribut uzeti za podjelu u analiziranom čvoru? Tablica u nastavku navodi utvrđene informacijske dobiti.

Podjela prema atributu	Informacijska dobit
Položaj	0.3113
Boja	0.3113

Kako za oba atributa imamo jednaku informacijsku dobit, možemo uzeti bilo koji od njih. Mi ćemo uzeti atribut *Položaj*. Stoga u stablo dodajemo novi čvor koji uzorke razvrstava prema atributu *Položaj*:



U najnižem lijevom čvoru stabla sada analiziramo uzorke 11 i 12; primijetite da oni imaju fiksirano *Oštećenje*=malo i *Položaj*=bliži:

Redni broj	Oštećenje	Položaj	Boja	Hitno
11.	malo	bliži	svijetlo	NE
12.	malo	bliži	tamno	DA

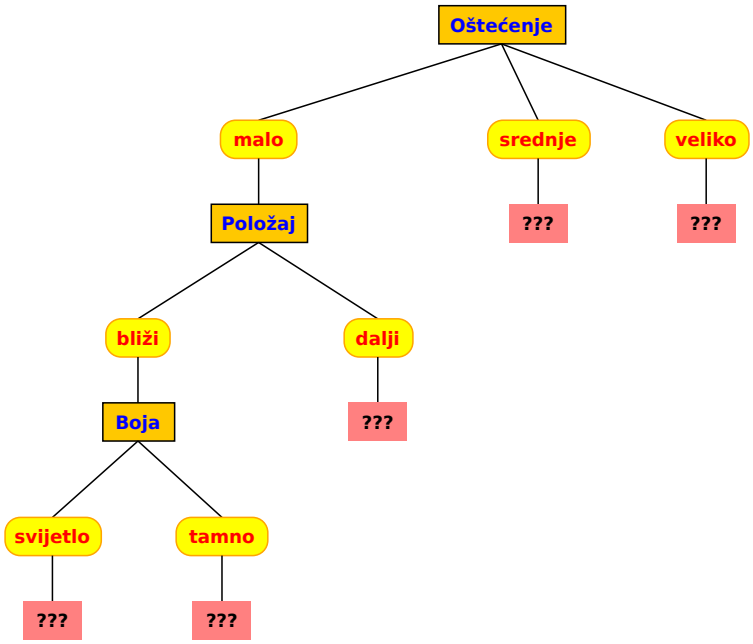
Veličina ovog podskupa je 2, a entropija 1. Preostalo nam je uzorke razvrstati po atributu *Boja*. Za *Boja*=svijetlo imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
11.	malo	bliži	svijetlo	NE

a za *Boja*=tamno imamo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
12.	malo	bliži	tamno	DA

Entropija oba skupa je 0, a informacijska dobit razvrstavanja po ovom atributu je 1. Stoga na ovom mjestu dodajemo novi čvor koji uzorke razvrstava po atributu *Boja* te u njegovu lijevu granu šaljemo prvi podskup, a u desnu granu drugi. Slika u nastavku prikazuje trenutno stablo:



U najdonjem lijevom čvoru na analizu dolazi samo uzorak:

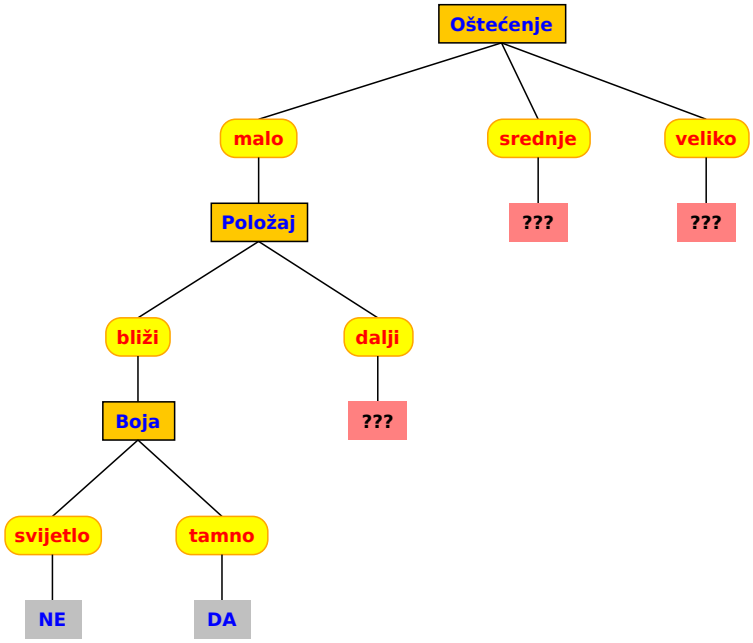
Redni broj	Oštećenje	Položaj	Boja	Hitno
11.	malo	bliži	svijetlo	NE

Kako ovdje imamo situaciju da svi uzorci pripadaju istom razredu, čvor pretvaramo u klasifikacijski čvor i zapisujemo da se uzorcima pridjeljuje razred NE.

U susjedni čvor na analizu dolazi samo uzorak:

Redni broj	Oštećenje	Položaj	Boja	Hitno
12.	malo	bliži	tamno	DA

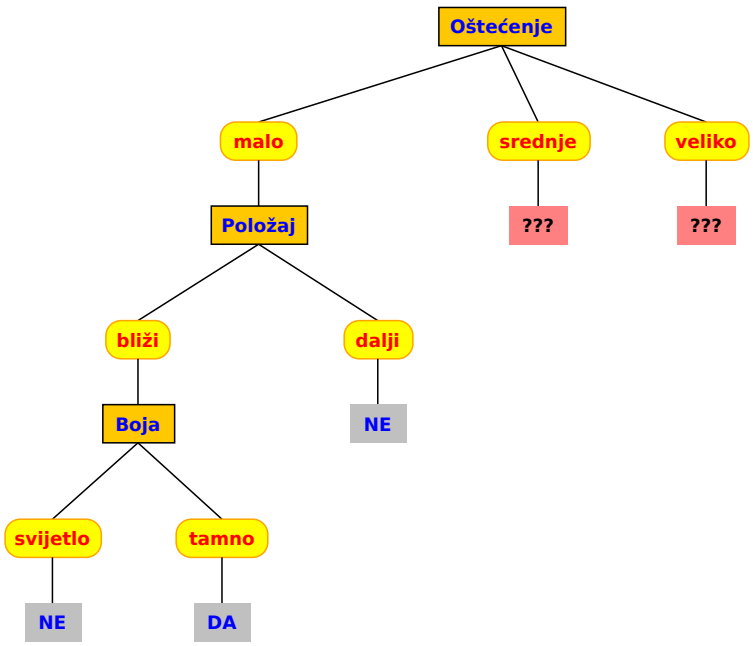
I ovdje imamo situaciju da svi uzorci pripadaju istom razredu. Čvor pretvaramo u klasifikacijski čvor i zapisujemo da se uzorcima pridjeljuje razred DA. Trenutno izgrađeno stablo prikazano je na slici u nastavku.



U desno dijete čvora *Položaj* na analizu dolazi skup:

Redni broj	Oštećenje	Položaj	Boja	Hitno
3.	malo	dalji	svijetlo	NE
11.	malo	bliži	svijetlo	NE

Kako svi uzorci pripadaju istom razredu, čvor pretvaramo u klasifikacijski čvor i zapisujemo da se uzorcima pridjeljuje razred NE. Trenutno izgrađeno stablo prikazano je na slici u nastavku.



Analizu nastavljamo sa srednjim djetetom korijenskog čvora. U njega smo poslali na analizu uzorke 2, 7, 9 i 10:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
7.	srednje	dalji	svijetlo	NE
9.	srednje	dalji	tamno	NE
10.	srednje	bliži	tamno	DA

Veličina ovog skupa je 4, a entropija 1. Skup možemo razdijeliti prema vrijednostima atributa *Položaj* odnosno prema vrijednostima atributa *Boja*.

Podjela prema vrijednostima atributa *Položaj* daje za *Položaj*=bliži:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
10.	srednje	bliži	tamno	DA

odnosno za *Položaj*=dalji:

Redni broj	Oštećenje	Položaj	Boja	Hitno
7.	srednje	dalji	svijetlo	NE
9.	srednje	dalji	tamno	NE

Entropija oba podskupa je 0 pa je informacijska dobit podjele prema atributu *Položaj* jednaka 1.

Podjela prema vrijednostima atributa *Boja* daje za *Boja*=svijetlo:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
7.	srednje	dalji	svijetlo	NE

odnosno za *Boja*=tamno:

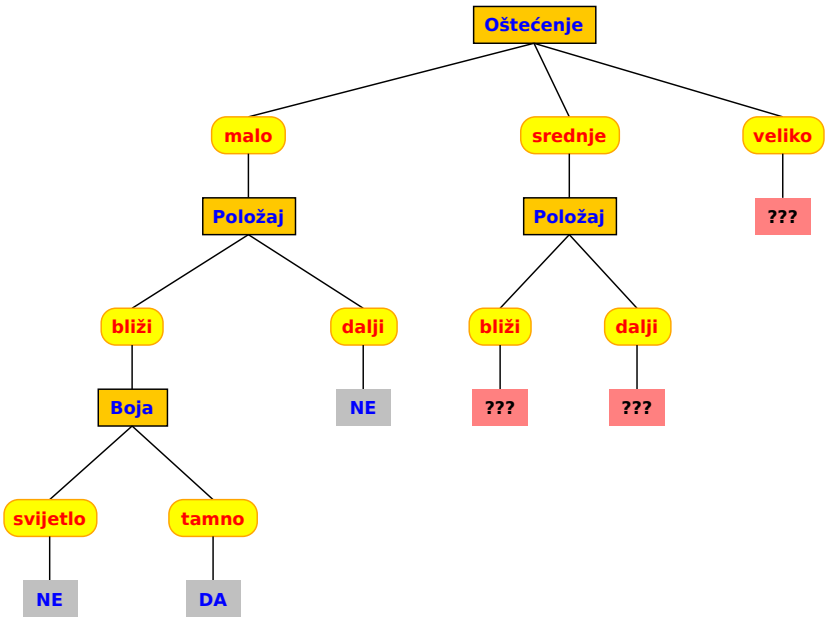
Redni broj	Oštećenje	Položaj	Boja	Hitno
9.	srednje	dalji	tamno	NE
10.	srednje	bliži	tamno	DA

Entropija oba podskupa je 1 pa je informacijska dobit podjele prema atributu *Boja* jednaka 0.

Koji ćemo sada atribut uzeti za podjelu u analiziranom čvoru? Tablica u nastavku navodi utvrđene informacijske dobiti.

Podjela prema atributu	Informacijska dobit
Položaj	1
Boja	0

Biramo atribut *Položaj*, te stvaramo čvor koji obavlja razvrstavanje prema njemu. Trenutno stanje stabla prikazano je na slici u nastavku.



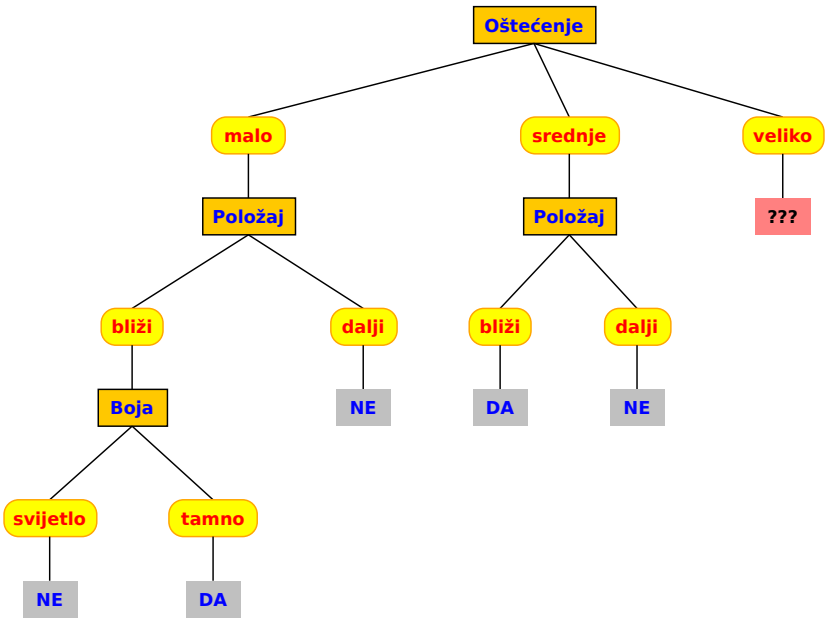
U lijevo dijete novostvorenog čvora šaljemo podskup:

Redni broj	Oštećenje	Položaj	Boja	Hitno
2.	srednje	bliži	svijetlo	DA
10.	srednje	bliži	tamno	DA

a u desno *Položaj*=dalji:

Redni broj	Oštećenje	Položaj	Boja	Hitno
7.	srednje	dalji	svijetlo	NE
9.	srednje	dalji	tamno	NE

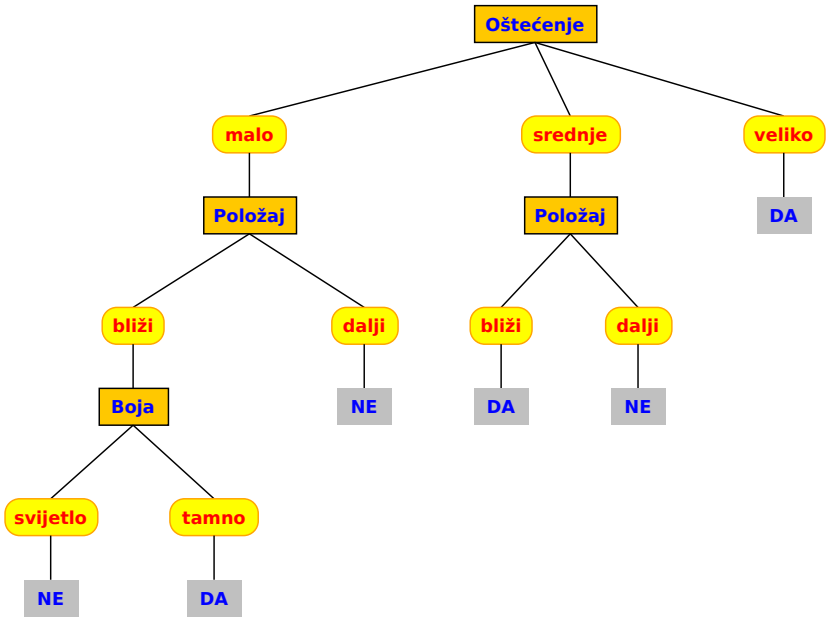
i rekurzivno provodimo algoritam. U oba slučaja utvrdit ćemo da svi uzorci imaju konzistentnu klasificikaciju (svi unutar podskupa pripadaju istom razredu), pa ćemo dodati čvorove koji obavljaju klasifikaciju; u prvom slučaju DA, a u drugom NE.



Konačno, u najdesnije dijete korijenskog čvora na analizu smo poslali podskup:

Redni broj	Oštećenje	Položaj	Boja	Hitno
4.	veliko	bliži	svijetlo	DA
5.	veliko	dalji	tamno	DA
6.	veliko	dalji	svijetlo	DA
8.	veliko	bliži	tamno	DA

Vidimo da svi uzorci pripadaju istom razredu; stoga ćemo na ovom mjestu napraviti čvor koji obavlja klasifikaciju u razred DA, što je prikazano na konačnoj slici u nastavku.



Isprobajte

Algoritam izgradnje stabla možete isprobati i samostalno. U naredbenom retku zadajte:
`java -cp book-ml.jar gui.id3.MainBuilder hitno.txt`
čime ćete pokrenuti program koji će Vam omogućiti da interaktivno gradite stablo odluke. Klikom na neizgrađeni čvor u donjem dijelu prozora prikazat će se kompletna analiza s objašnjenjima kao i linkovi kojima ćete moći odabrati kako želite dalje nastaviti. U slučajevima da u podskupu svi uzorci pripadaju istom razredu, samo ćete moći dodati klasifikacijski čvor. U suprotnom, moći ćete odabrati po kojem od atributa maksimalne informacijske dobiti (ako ih je više) želite napraviti podjelu.

Datoteka `hitno.txt` kao i datoteke `sport.txt` i `sport2.txt` su ugrađene datoteke. Možete i sami pripremiti svoje primjere: napravite tekstovnu datoteku na disku u kojoj prvi redak sadrži nazive atributa (posljednji atribut se smatra ciljnim) a svaki preostali redak po jedan uzorak. Kao separator koristite tabulator.

Želite li sami birati prema kojem atributu želite raditi podjelu (bez ograničenja da to mora biti jedan od onih koji daju maksimalnu informacijsku dobit), možete pokrenuti:
`java -cp book-ml.jar gui.id3.MainFreeBuilder hitno.txt`

Ako samo želite pogledati gotovo stablo, bez potrebe da ga gradite čvor po čvor, zadajte:
`java -cp book-ml.jar gui.id3.Main hitno.txt`

3.1 Formalna definicija algoritma ID3

Algoritam ID3 je rekurzivan pohlepni algoritam za izgradnju stabla odluke. Algoritam u rekurzivnom pozivu analizira trenutni skup uzoraka - označimo ga sa *S*.

1. Ako je skup uzoraka *S* prazan, tada se stvara klasifikacijski čvor koji uzorke klasificira u razred koji je najčešću u skupu uzoraka koji je razmatrao neposredni roditelj.
2. Ako svi uzorci u promatranom skupu *S* pripadaju istom razredu (označimo taj razred kao *r*), algoritam stvara klasifikacijski čvor koji uzorcima dodjeljuje razred *r* i tu rekurzija staje.
3. Ako više nema atributa koji nisu razmatrani u nekom od roditelja (sve do korijena), stvara se klasifikacijski čvor koji uzorcima pridjeljuje razred koji je najčešći u analiziranom skupu *S*.
4. U suprotnom, algoritam razmatra podjele skupova u podskupove prema svim atributima koji nisu u roditeljskim čvorovima. Algoritam odabire atribut koji daje maksimalnu informacijsku dobit, stvara čvor koji razmatra vrijednost tog atributa te za svaku različitu vrijednost atributa dodaje po jedno dijete koje gradi rekurzivnim pozivom uz podskup skupa *S* u kojem se nalaze samo oni uzorci kojima je odabrani atribut postavljen na razmatranu vrijednost.

Slučaj pod brojem 1 može se dogoditi u dubljim koracima rekurzije, s obzirom da malo po malo smanjujemo skup uzoraka koji analiziramo, da dođemo u situaciju da kada razmatrani skup razvrstavamo u podskupove prema vrijednostima nekog atributa, da je neki od tih podskupova prazan (primjerice, analizom smo došli u situaciju da razvrstavamo prema atributu *Oštećenje*, a svi uzorci u podskupu imaju *Oštećenje*=veliko ili *Oštećenje*=malo; tada bismo u granu koja odgovara *Oštećenje*=srednje poslali prazan skup koji bi potom bio razriješen kao slučaj 1. Evo i konkretnog primjera. Razmotrite skup podataka prikazan tablicom u nastavku.

Redni broj	Atribut1	Atribut2	Atribut3	Cilj
1.	a1	b1	c1	DA
2.	a1	b2	c1	DA
3.	a2	b1	c1	DA
4.	a2	b2	c1	DA
5.	a3	b1	c1	DA
6.	a3	b2	c1	DA
7.	a1	b1	c2	DA
8.	a1	b2	c2	DA
9.	a1	b3	c2	NE
10.	a3	b1	c2	NE
11.	a3	b2	c2	NE
12.	a3	b3	c2	NE

U korijenskom čvoru otkrit ćemo da se najveća informacijska dobit postiže podjelom po atributu *Atribut3*. Stoga ćemo napraviti korijenski čvor koji će uzorke razvrstavati prema atributu *Atribut3* i koji će imati dva djeteta: lijevo za *Atribut3*=c1 u koje ćemo dalje poslati uzorke:

Redni broj	Atribut1	Atribut2	Atribut3	Cilj
1.	a1	b1	c1	DA
2.	a1	b2	c1	DA
3.	a2	b1	c1	DA
4.	a2	b2	c1	DA
5.	a3	b1	c1	DA
6.	a3	b2	c1	DA

i desno za *Atribut3*=c2 u koje ćemo dalje poslati uzorke:

Redni broj	Atribut1	Atribut2	Atribut3	Cilj
7.	a1	b1	c2	DA
8.	a1	b2	c2	DA
9.	a1	b3	c2	NE
10.	a3	b1	c2	NE
11.	a3	b2	c2	NE
12.	a3	b3	c2	NE

Analizom ovog drugog slučaja utvrdit ćemo da se najveća informacijska dobit postiže podjelom po atributu *Atribut1*. Stoga ćemo stvoriti čvor koji obavlja razvrstavanje prema tom atributu i ima tri djeteta. U prvo ćemo poslati podskup uzoraka kod kojeg uzorci imaju *Atribut1*=a1:

Redni broj	Atribut1	Atribut2	Atribut3	Cilj
7.	a1	b1	c2	DA
8.	a1	b2	c2	DA
9.	a1	b3	c2	NE

srednjem djetetu ćemo poslati podskup uzoraka kod kojeg uzorci imaju *Atribut1*=a2:

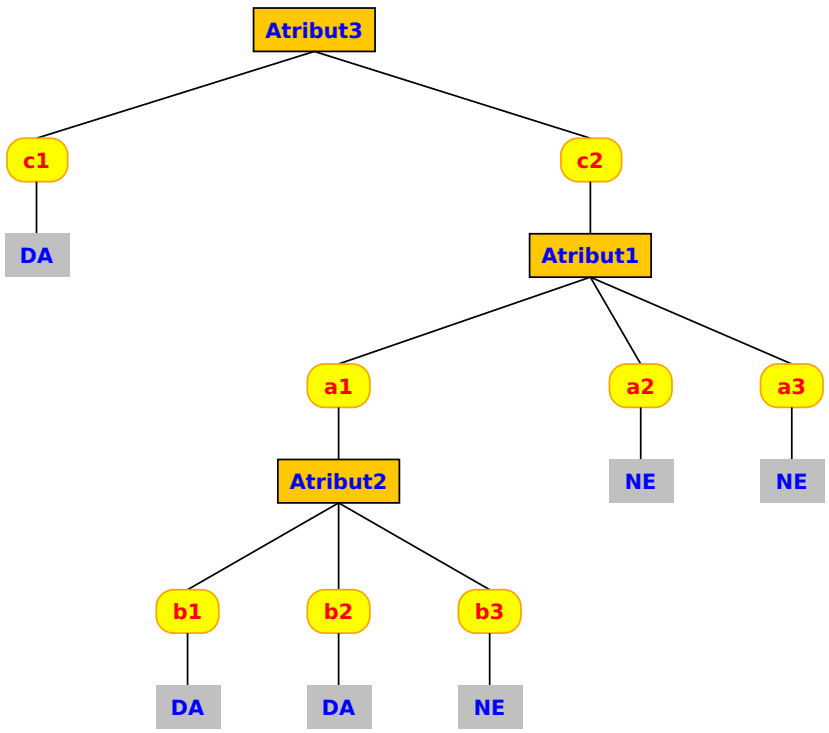
Redni broj	Atribut1	Atribut2	Atribut3	Cilj
-	-	-	-	-

i posljednjem djetetu ćemo poslati podskup uzoraka kod kojeg uzorci imaju *Atribut1*=a3:

Redni broj	Atribut1	Atribut2	Atribut3	Cilj
10.	a3	b1	c2	NE
11.	a3	b2	c2	NE
12.	a3	b3	c2	NE

Primijetite što se dogodilo sa srednjim djetetom: na analizu je dobilo prazan skup uzoraka. Obrada tog slučaja pogledat će uzorke koje je analizirao roditelj i ustanoviti isti imao 2 uzorka razreda DA i 4 uzorka razreda NE; posljedično, napravit će se terminalni klasifikacijski čvor koji će uzorcima dodijeljivati razred NE.

Konačan izgled stabla odluke nakon što su razriješeni svi čvorovi prikazan je na slici u nastavku.



Isprobajte

Ovo možete isprobati i samostalno. U naredbenom retku zadajte:

```
java -cp book-ml.jar gui.id3.MainBuilder slucaj1.txt
```

3.1.1 Svojstva i nadogradnje algoritma

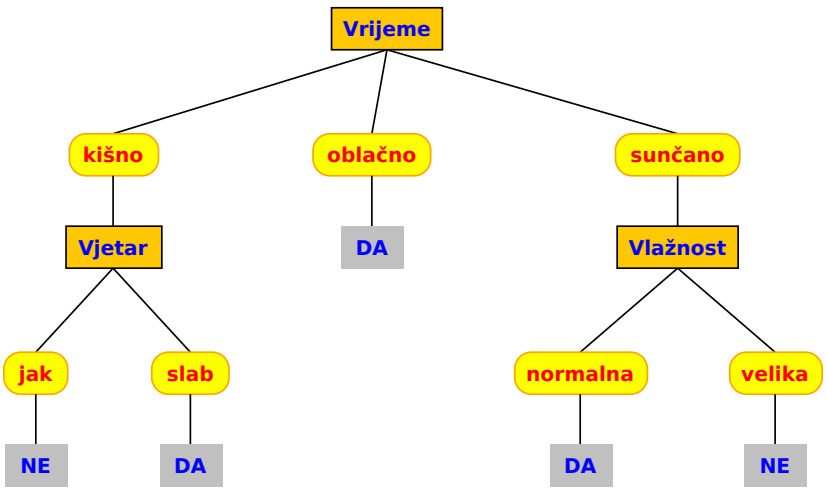
Algoritam ID3 je pohlepan algoritam jer u svakom koraku donosi lokalno optimalnu odluku. Time algoritam može upasti u lokalni minimum i izgraditi stablo koje nije doista minimalno moguće.

Daljnja nadogradnja ovog algoritma su algoritmi C4.5 te C5 koje ovdje nećemo razmatrati. Spomenimo da postoje i daljnja proširenja ideje stabala odluke, a jedan od primjera su takozvane *Slučajne šume* (engl. *Random forest*): kod tog pristupa na temelju različitih podskupova atributa uzorka gradi se niz klasifikacijskih stabala; prilikom uporabe, svako od stabala obavlja klasifikaciju te se na temelju njihovih rezultata mogu odrediti vjerojatnosti da uzorak pripada svakom od razreda (primjerice, razmotrite slučaj gdje šuma ima 100 stabala, i 90 od njih uzorak klasificira kao DA, a 10 kao NE). Također, spomenimo da se prilikom odabira atributa ne mora uvijek gledati informacijska dobit već postoje i druge mjere kvalitete podjele (primjer je *gini index* – ostavljamo zainteresiranom čitatelju da istraži o čemu se radi).

U prisustvu stršećih vrijednosti, algoritam se može prenaučiti. Posljedica toga jest staranje stabla koje ima puno više čvorova no što je to potrebno, te koje loše generalizira. Jedna od tehnika popravljanja generalizacijskih sposobnosti stabla jest postupak podrezivanja koji se najčešće provodi nakon što je (potencijalno prenaučeno) stablo izgrađeno. Kreće se od najdubljih čvorova stabla i putuje prema korijenu. Razmatra se cijena uklanjanja podstabla kojem je razmatrani čvor korijen te zamjena tog čitavog podstabla jednim klasifikacijskim čvorom koji bi uzorcima pridruživao razred koji je bio najčešći među uzorcima koje je čvor analizirao prilikom izgradnje stabla. Ako je ta cijena prihvatljiva, provodi se uklanjanje podstabla i zamjena klasifikacijskim čvorom.

3.2 Skup uzoraka *Dan za sport*

Jedan od čestih primjera na kojem se analizira izgradnja stabla odluke jest skup uzoraka *Dan za sport*. Klasifikacijsko stablo koje obavlja klasifikaciju uzoraka prikazano je na slici u nastavku.



Skup uzoraka *Dan za sport* prikazan je tablicom u nastavku.

Redni broj	Vrijeme	Temperatura	Vlažnost	Vjetar	Igra
1.	sunčano	vruće	velika	slab	NE
2.	sunčano	vruće	velika	jak	NE
3.	oblačno	vruće	velika	slab	DA
4.	kišno	ugodno	velika	slab	DA
5.	kišno	hladno	normalna	slab	DA
6.	kišno	hladno	normalna	jak	NE
7.	oblačno	hladno	normalna	jak	DA
8.	sunčano	ugodno	velika	slab	NE
9.	sunčano	hladno	normalna	slab	DA
10.	kišno	ugodno	normalna	slab	DA
11.	sunčano	ugodno	normalna	jak	DA
12.	oblačno	ugodno	velika	jak	DA
13.	oblačno	vruće	normalna	slab	DA
14.	kišno	ugodno	velika	jak	NE

Isprobajte

Čitav postupak izgradnje ovog stabla odluke možete isprobati i samostalno. U naredbenom retku zadajte:

```
java -cp book-ml.jar gui.id3.MainBuilder sport.txt
```

pa kliknite na korijenski čvor. U donjem dijelu prozora prikazat će se sva objašnjenja i izračuni i tu ćete moći odabrati željenu akciju. Ponavljanjem postupka malo po malo moći ćete izgraditi čitavo stablo.

Bibliografija

Knjige

Članci

Konferencijski radovi i ostalo