Istraživanje podataka

Vežbe 4

12. März 2021

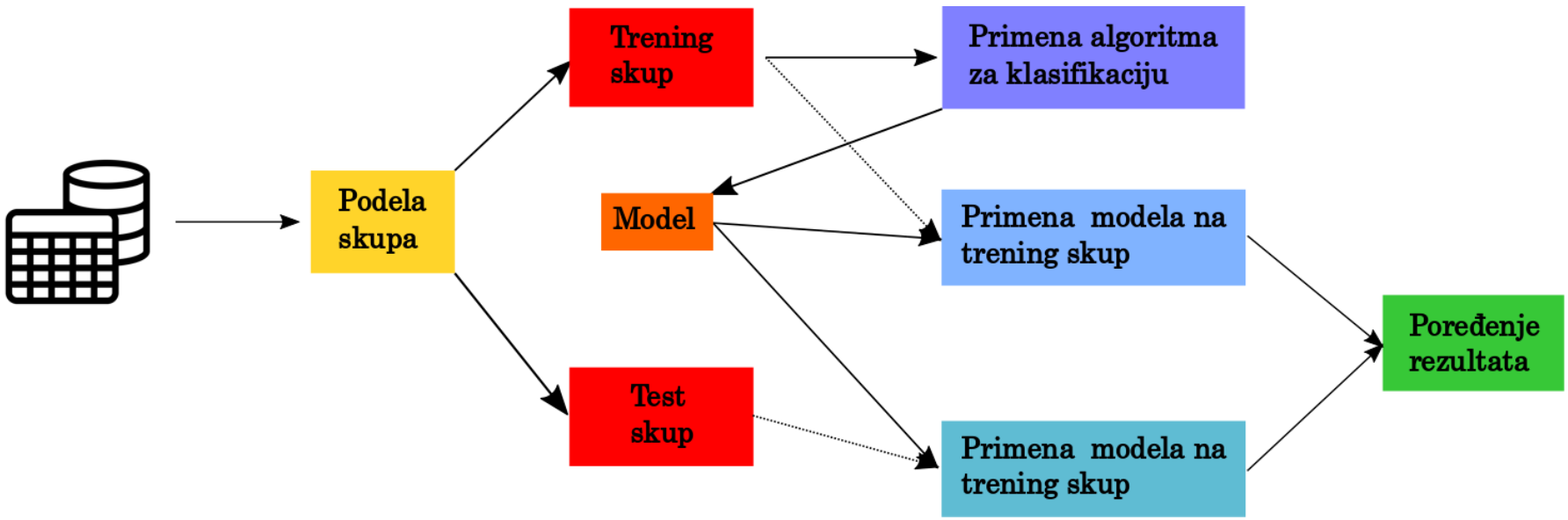
Outline

Outline

Klasifikacija

- Ulazni podaci: svaki slog (instanca) je oblika (x, y) gde je x skup (ulaznih) atributa, a y je ciljni atribut (klasa).
- Cilj klasifikacije: pronaći funkciju f (model klasifikacije) koja preslikava skup atributa x u jednu od predefinisanih oznaka klasa y.
- Podela skupa na trening i test skup.

Klasifikacija



Klasifikacija - mere za ocenu modela

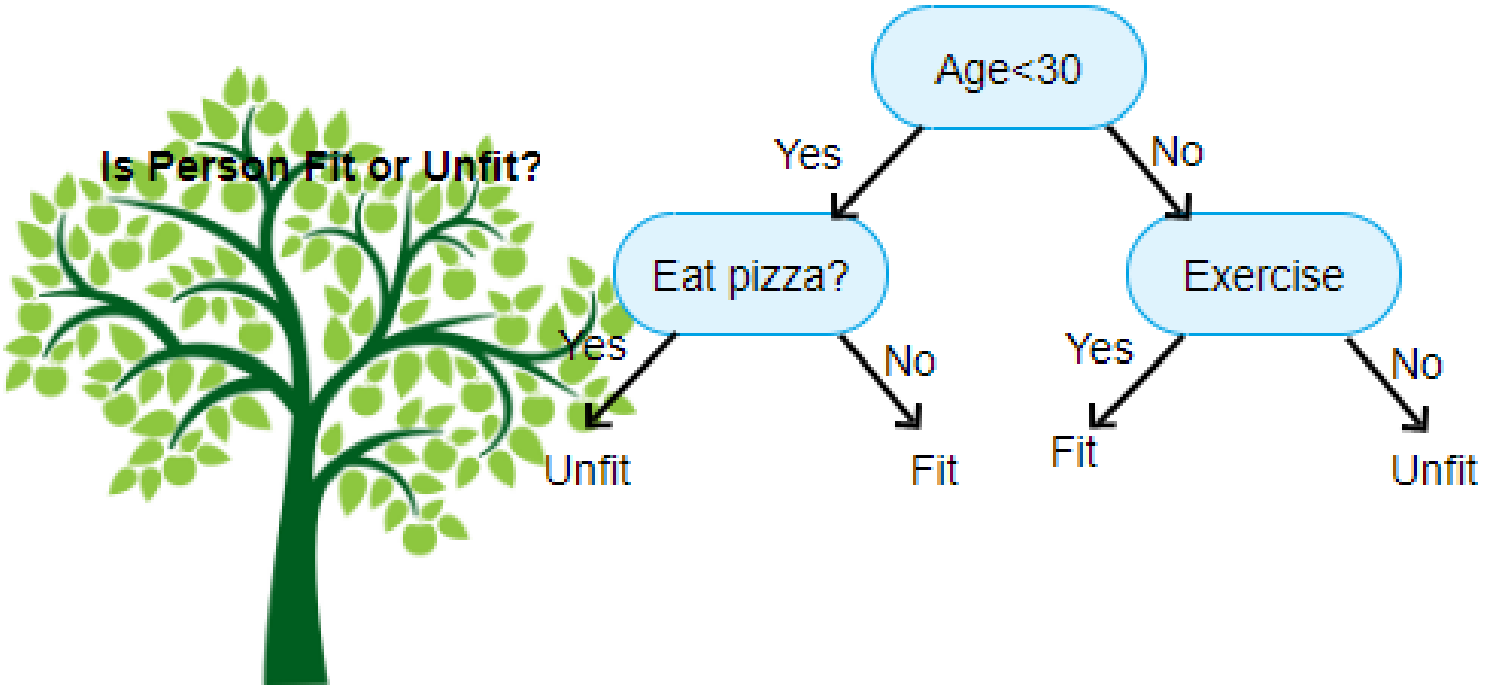
- *preciznost* $= \frac{\text{Broj slogova čija klasa je dobro predviđena modelom}}{\text{Ukupan broj slogova}}$ (eng. accuracy)
- stopa greške $= \frac{\text{Broj slogova čija klasa nije dobro predviđena modelom}}{\text{Ukupan broj slogova}}$ (eng. error rate)

Outline

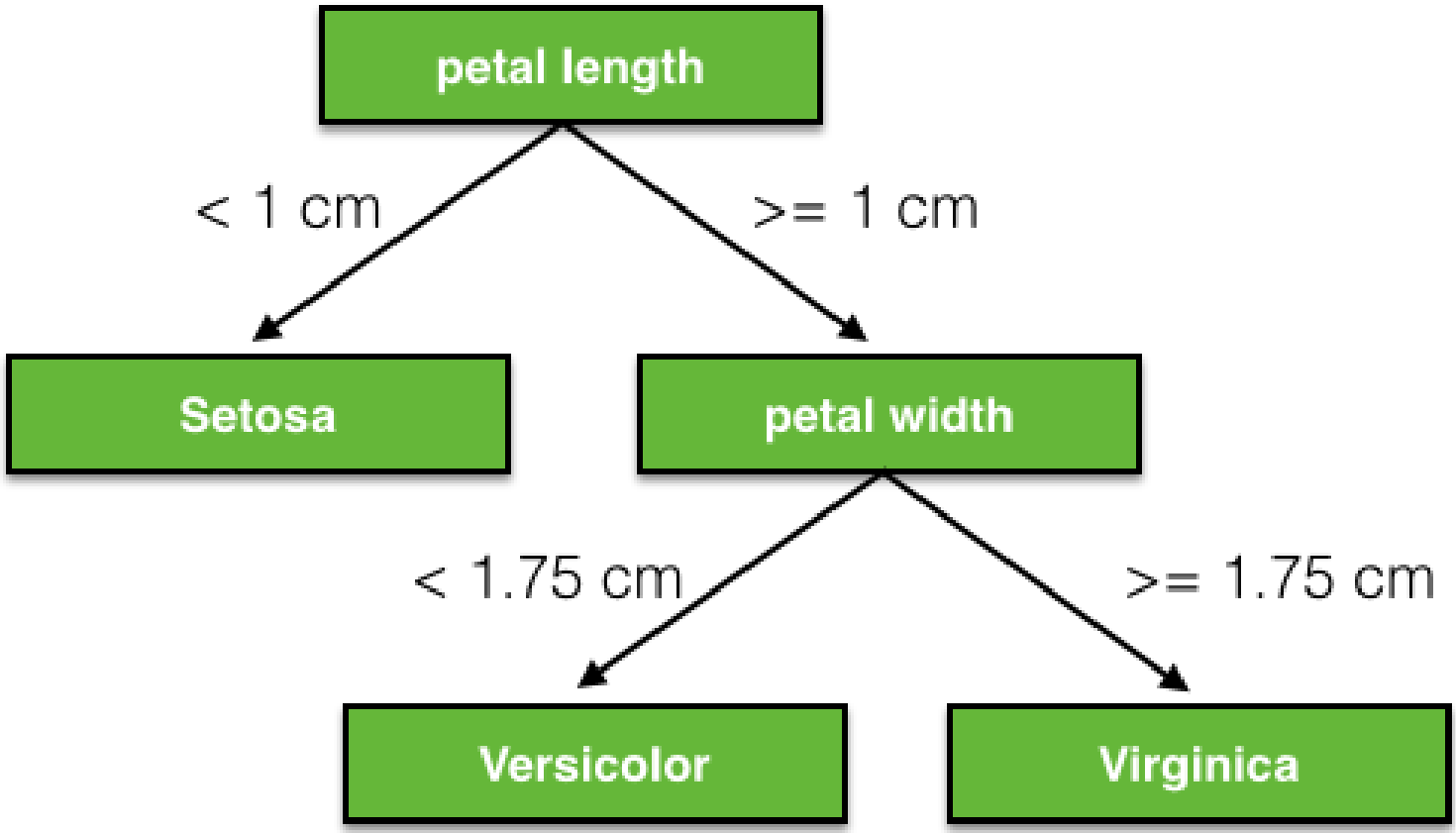
Drveta odlučivanja

- Model klasifikacije se predstavlja kao drvo odlučivanja koje ima
  - unutrašnje čvorove. Svaki unutrašnji čvor sadrži uslov nad test atributom koji služi za podelu slogova koji imaju različite karakteristike tako da se dobiju *čistije* grupe slogova. Grane koje izlaze iz unutrašnjeg čvora odgovaraju mogućim vrednostima test atributa.
  - listove. Svakom listu je dodeljena jedna klasa.

Primer drveta odlučivanja



Primer drveta odlučivanja



Drveta odlučivanja - klasifikacija sloga

Klasifikacija sloga: počevši od korena drveta odlučivanja, primenjuje se test uslov nad slogom i prati se grana koja odgovara dobijenom rezultatu. Ukoliko se pri spuštanju niz drvo odlučivanja naiđe na unutrašnji čvor, postupak se ponavlja (test uslov se primenjuje na slog i prati se grana koja odgovara rezultatu testa). Ako se naiđe na list, slogu se dodeljuje klasa koja je pridružena tom listu.

Drveta odlučivanja - pravljenje drveta odlučivanja

Opšti algoritam

1. Neka je D_t skup slogova za trening koji se nalaze u čvoru t, a $y = y_1, ..., y_c$ su oznake klasa
2. Ako D_t sadrži samo slogove koji pripadaju jednoj klasi y_t, tada je t list označen sa y_t
3. Ako D_t sadrži slogove koji se nalaze u više od jedne klase, tada se koristi test atribut radi podele podataka u manje podskupove. Na dobijene podskupove se zatim rekurzivno primenjuje kompletna procedura.

Mere nečistoće

$p(j|t)$ je relativna frekvencija klase j u čvoru t

Ginijev indeks

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

Entropija

$$Entropy(t) = -\sum_j p(j|t) * \log_2 p(j|t)$$

Greška klasifikacije

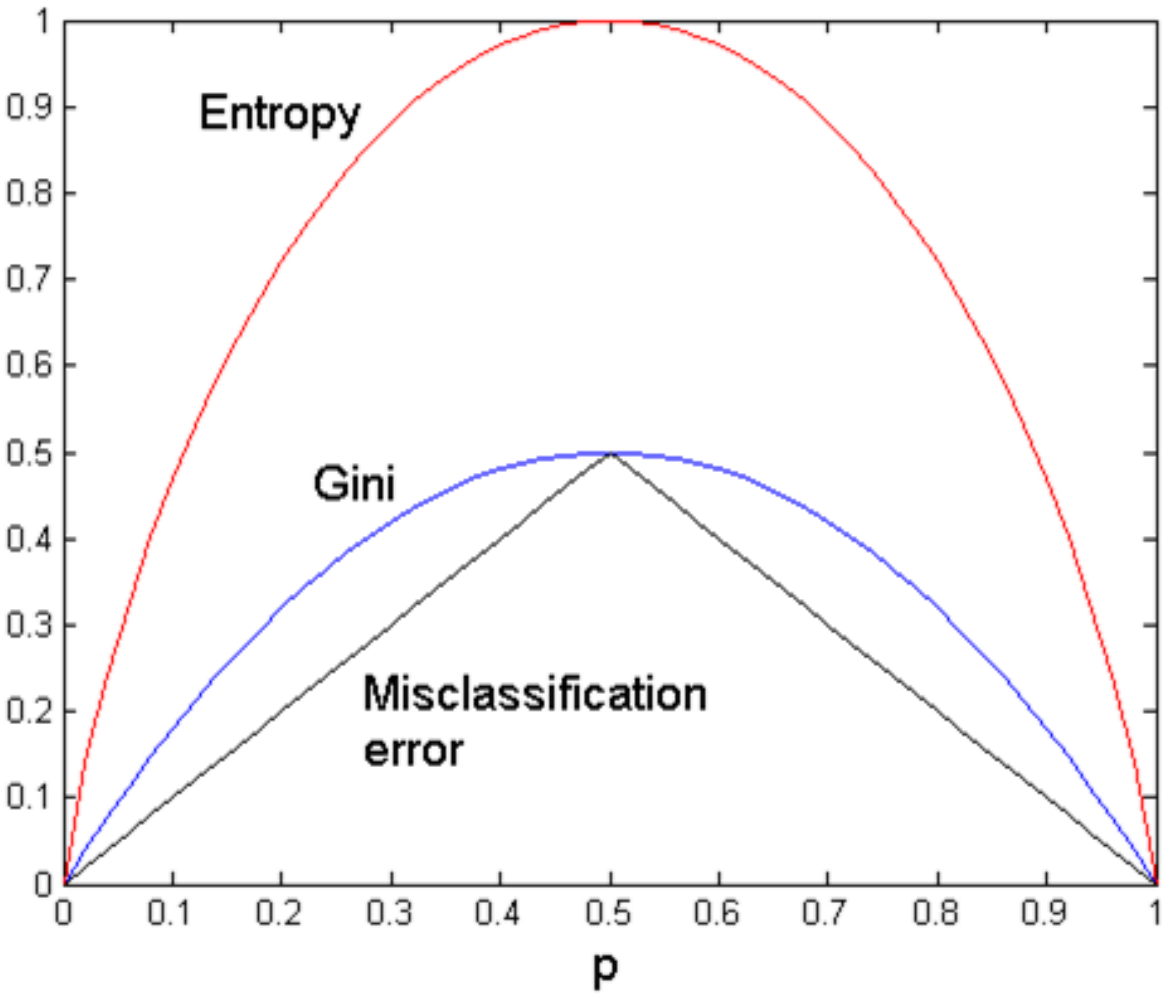
$$Error(t) = 1 - \max_j p(j|t)$$

Mere nečistoće

Dobit

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} * I(v_j)$$

Mere nečistoće



Drveta odlučivanja - podela prema tipu atributa

- Imenski atributi: binarna ili višestruka podela
- Redni atributi: binarna ili višestruka podela vodeći računa o uređenju
- Neprekidni atributi: potrebno je pronaći najbolju tačku/tačke prekida za binarna ili višestruku podelu

Outline

Zadatak 1

Dati su trening primeri za problem binarne klasifikacije.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Zadatak 1

- Kolika je entropija skupa trening podataka?

Zadatak 1

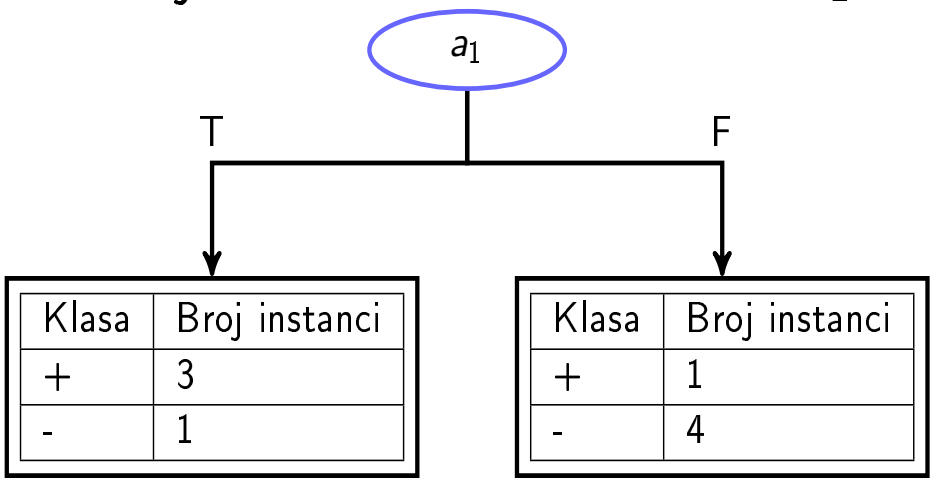
- Kolika je entropija skupa trening podataka?
  $p(+) = \frac{4}{9}$ $p(-) = \frac{5}{9}$
  $Entropy(root) = -\frac{4}{9} * \log_2 \frac{4}{9} - \frac{5}{9} * \log_2 \frac{5}{9} = 0,9911$

Zadatak 1

- Kolika je informaciona dobit za a_1 na ovim trening podacima?

Zadatak 1

- Kolika je informaciona dobit za a_1 na ovim trening podacima?



Klasa	Broj instanci
+	3
-	1

Klasa	Broj instanci
+	1
-	4

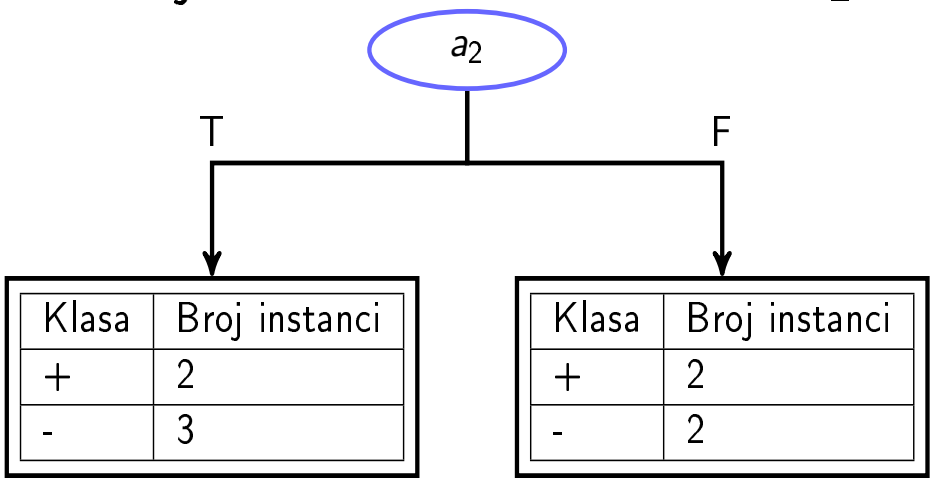
$$Entropy(a_1 = T) = -\frac{3}{4} * \log_2 \frac{3}{4} - \frac{1}{4} * \log_2 \frac{1}{4} = 0,8113$$
$$Entropy(a_1 = F) = -\frac{1}{5} * \log_2 \frac{1}{5} - \frac{4}{5} * \log_2 \frac{4}{5} = 0,7219$$
$$\Delta = 0,9911 - \frac{4}{9} * 0,8113 - \frac{5}{9} * 0,7219 = 0,2295$$

Zadatak 1

- Kolika je informaciona dobit za a_2 na ovim trening podacima?

Zadatak 1

- Kolika je informaciona dobit za a_2 na ovim trening podacima?



Klasa	Broj instanci
+	2
-	3

Klasa	Broj instanci
+	2
-	2

$$Entropy(a_2 = T) = -\frac{2}{5} * \log_2 \frac{2}{5} - \frac{3}{5} * \log_2 \frac{3}{5} = 0,971$$
$$Entropy(a_2 = F) = 1$$
$$\Delta = 0,9911 - \frac{5}{9} * 0,971 - \frac{4}{9} * 1 = 0,0072$$

Zadatak 1

- Za a_3, koji je neprekidan atribut, izračunati informacionu dobit za svaku moguću podelu.

Zadatak 1

- Za a_3, koji je neprekidan atribut, izračunati informacionu dobit za svaku moguću podelu.

Vrednosti iz kolone	1	3	4	5	6	7	8
Tačka podele	2	3,5	4,5	5,5	6,5	7,5	
Uslovi grana	$\leq$ | $>$	$\leq$ | $>$	$\leq$ | $>$	$\leq$ | $>$	$\leq$ | $>$	$\leq$ | $>$	
Klasa +	1 | 3	1 | 3	2 | 2	2 | 2	3 | 1	4 | 0	
Klasa -	0 | 5	1 | 4	1 | 4	3 | 2	3 | 2	4 | 1	
Δ	0,1427	0,0026	0,0728	0,0072	0,0183	0,1022	

Zadatak 1

- Koji atribut je najbolji za podelu (između a_1, a_2 i a_3) prema informacionoj dobiti?

Zadatak 1

- Koji atribut je najbolji za podelu (između a_1, a_2 i a_3) prema informacionoj dobiti?

  a_1

Zadatak 1

- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?

Zadatak 1

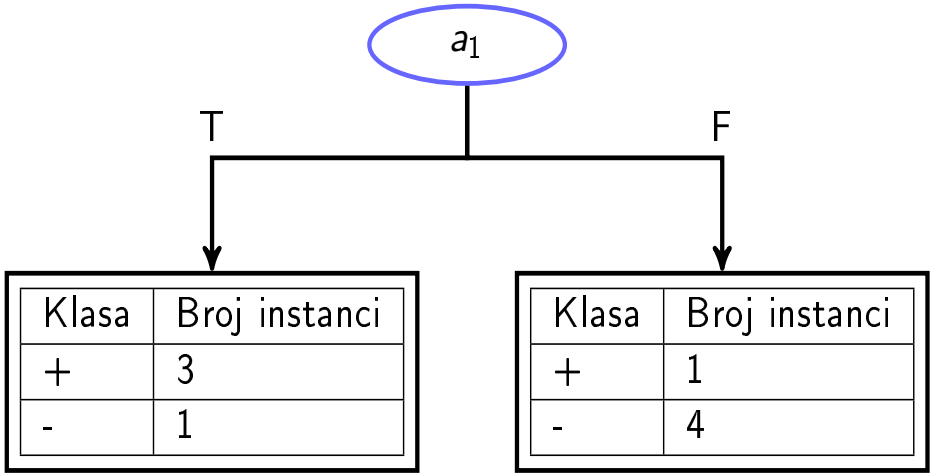
- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?
  $p(+) = \frac{4}{9}$ $p(-) = \frac{5}{9}$
  $CError(root) = 1 - \max(\frac{4}{9}, \frac{5}{9}) = \frac{4}{9}$

Zadatak 1

- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?

Zadatak 1

- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?



Klasa	Broj instanci
+	3
-	1

Klasa	Broj instanci
+	1
-	4

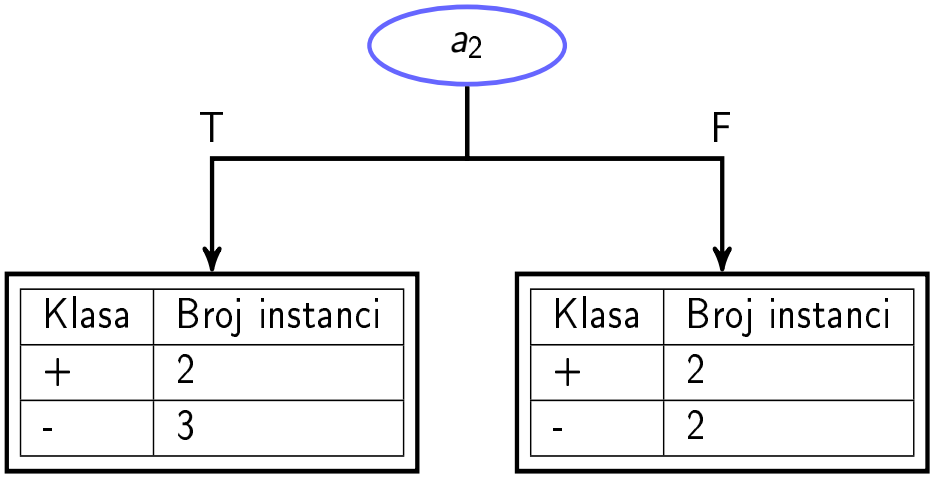
$$CError(a_1 = T) = 1 - \max(\tfrac{3}{4}, \tfrac{1}{4}) = \tfrac{1}{4}$$
$$CError(a_1 = F) = 1 \max(-\tfrac{1}{5}, \tfrac{4}{5}) = \tfrac{1}{5}$$
$$\Delta = \tfrac{4}{9} - (\tfrac{4}{9} * \tfrac{1}{4} + \tfrac{5}{9} * \tfrac{1}{5}) = \tfrac{2}{9}$$

Zadatak 1

- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?

Zadatak 1

- Koja je najbolja podela (između a_1 i a_2) ako se kao mera nečistoće koristi greška klasifikacije?



Klasa	Broj instanci
+	2
-	3

Klasa	Broj instanci
+	2
-	2

$CError(a_1 = T) = 1 - \max(\frac{2}{5}, \frac{3}{5}) = \frac{2}{5}$
$CError(a_1 = F) = 1 \max(-\frac{2}{4}, \frac{2}{4}) = \frac{1}{2}$
$\Delta = \frac{4}{9} - (\frac{5}{9} * \frac{2}{5} + \frac{4}{9} * \frac{2}{4}) = 0$
Zaključak: atribut a_1 je bolji za podelu.

Zadatak 2

Na osnovu datih podataka o životinjama iz trening skupa proceniti da li je životinja osobinama (*Velika*, *Biljke*, *Da*) opasna ili ne korišćenjem stabla odlučivanja dubine 2 uz korišćenje Ginijevog indeksa.

Veličina	Ishrana	Otrovnost	Opasna
Velika	Meso	Ne	Da
Mala	Meso	Ne	Ne
Mala	Biljke	Ne	Ne
Velika	Meso	Da	Da
Mala	Meso	Da	Da
Mala	Biljke	Ne	Ne
Mala	Biljke	Da	Da
Velika	Biljke	Ne	Da

Outline

C5.0

- koristi informacionu dobit (mera nečistoće entropija)
- binarna podela kada se numerički atribut koristi za test
- za kategoričke atribute podrazumevana podela - jedna vrednost jedna grana, a vrednosti mogu i da se grupišu

Opis nekih opcija

- korišćenje podeljenog skupa (trening i test skup)
- grupisanje kategoričkih podataka
- *boosting* - pravljenje više modela u nizu radi povećanja preciznosti. Prvi model se pravi na uobičajen način, a svaki sledeći se fokusira na instance koje su pogrešno klasifikovane prethodnim modelom. Za klasifikaciju instance se primenjuju svi modeli i koristi se sistem glasanja.
- *unakrsna-validacija* - pravljenje modela nad podskupovima radi procene preciznosti modela napravljenim nad celim skupom

Opis nekih opcija

- opcija za naklonost ka preciznosti ili uopštenosti modela
- očekivan procenat instanci sa greškom u trening skupu
- *strogost pri potkresivanju* - povećanjem vrednosti dobija se manje stablo
- minimalan broj instanci koji mora da bude u dete-čvoru nakon podele da bi se izvršila podela
- *winnow atributes* - izračunavanje važnosti atributa pre pravljenja modela
- matrica cene pogrešne klasifikacije

Outline

Zadatak

Primeniti klasifikaciju nad skupom *bank.csv* korišćenjem C5.0. Ciljni atribut je oročena štednja.

- Koji atributi su korišćeni pri pravljenju modela?
- Komentarisati dobijen model. Dati predlog za poboljšanje.