Mašinsko učenje
ID3 algoritam

Aleksandar Milosavljević
Vladan Mihajlović

Indukovanje stabla odluke

- ID3 (*Iterative Dihotomiser 3, J. Ross Quinlan, 1975*) je algoritam koji se koristi za indukovanje stabla odluke na osnovu primera tipa:
  - (***atribut1***, ***atribut2***, ..., ***atributN***, ***klasa***)
- Dobijeno stablo odluke se kasnije koristi za klasifikaciju novih uzoraka.
  - (***atribut1***, ***atribut2***, ..., ***atributN***)
  - ***klasa*** **= ?**

Stablo odluke

- Stablo odluke predstavlja struktura tipa stabla gde:
  - **čvorovi** odgovaraju atributima uzoraka,
  - **grane** ka drugim čvorovima odgovaraju vrednostima određenog atributa, a
  - **listovi** predstavljaju klase kojima pripadaju uzorci sa vrednostima atributa definisanim putem do korena.

Okamov brijač

- Indukovanje stabla odluke se zasniva na principu poznatom kao “Okamov brijač”.
- Ovaj princip glasi:
  - *Ako imamo dve podjednako verovatne teorije trebamo izabrati jednostavniju.*
- Primenjeno na indukciju stabla odluke, ovo pravilo znači da je najbolje stablo odluke najjednostavnije stablo odluke.

Entropija

- Kao meru jednostavnosti ID3 algoritam koristi **entropiju**.
- Entropija skupa **S** se računa kao:
  - $\mathbf{H(S) = \Sigma_{i=1..k}(- p(C_i) * \log_k(p(C_i)))}$
- gde su:
  - $\mathbf{p(C_i)}$ – procenat elemenata skupa **S** koji pripadaju klasi $\mathbf{C_i}$ (i = 1,...,k)
  - $\mathbf{\log_k}$ – je logaritam osnove k

Primer računanja entropije

- Ako je **S** skup koji sadrži **14** primera od kojih su **9** svrstani u klasu **C1**, a ostalih **5** u klasu **C2**, entropija skupa **S** je:
  - $H(S) = - (9/14) * \log_2(9/14) - (5/14) * \log_2(5/14)$
    $= - (9/14) * (\ln(9/14) / \ln2) - (5/14) * (\ln(5/14) / \ln2)$
    $= \mathbf{0,94}$

Značenje vrednosti entropije

- Vrednost entropije se kreće u intervalu **[0, 1]**.
- Vrednost entropije je **0** kada svi elementi skupa pripadaju jednoj istoj klasi.
  - Skup je perfektno klasifikovan
- Vrednost entropije je **1** kada svakoj klasi pripada podjednak broj elemenata.
  - Skup je potpuno stohastički

Izbor atributa

- Za koren stabla odluke bira se atribut koji nosi najviše informacija za ceo skup primera.
- Korist nekog atributa se računa po sledećoj formuli:
  - $\mathbf{G(S, A) = H(S) - \Sigma_{i=1..m}(|S_{Ai}|/|S| * H(S_{Ai}))}$
- gde je:
  - **H(X)** – entropija skupa X
  - **m** – broj različitih vrednosti atributa A
  - $\mathbf{S_{Ai}}$ – podskup skupa S gde atribut A ima vrednost A_i
  - **|X|** - broj elemenata skupa X

Opis ID3 algoritma

1. Ako svi primeri pripadaju istoj klasi:
   1. Kreiraj list sa vrednošću koja odgovara toj klasi.
2. U suprotnom:
   1. Nađi atribut sa najvećom dobiti.
   2. Dodaj granu za svaku vrednost tog atributa.
   3. Rasporedi primere u odgovarajuće podskupove.
   4. Za svaki podskup ponovi algoritam.

Zadatak

- Koristeći ID3 algoritam indukovati stablo koje će biti od pomoći pri donošenju odluke “Da li je vreme pogodno za igranje košarke?”
- Da bi se formirao skup primera za ID3 algoritam, u periodu od 14 dana praćeni su vremenski uslovi i aktivnost na obližnjem košarkaškom igralištu.

Zadatak

- Vreme je opisano sledećim atributima:
  - **Izgled vremena**, sa vrednostima:
    - **sunčano**, **oblačno** i **kiša**.
  - **Temperatura**, sa vrednostima:
    - **toplo**, **prijatno** i **hladno**.
  - **Vlažnost**, sa vrednostima:
    - **normalna** i **visoka**.
  - **Vetar**, sa vrednostima:
    - **slab** i **jak**.
- Primeri su klasifikovani u dve klase:
  - **pogodno** i **nepogodno** vreme za košarku.

Zadatak (tablica primera)

	Izgled vremena	Temperatura	Vlažnost	Vetar	KLASA
1.	sunčano	toplo	visoka	slab	nepogodno
2.	sunčano	toplo	visoka	jak	pogodno
3.	oblačno	toplo	visoka	slab	pogodno
4.	kiša	prijatno	visoka	slab	nepogodno
5.	kiša	hladno	normalna	slab	nepogodno
6.	kiša	hladno	normalna	jak	nepogodno
7.	oblačno	hladno	normalna	jak	nepogodno
8.	sunčano	prijatno	visoka	slab	pogodno
9.	sunčano	hladno	normalna	slab	pogodno
10.	kiša	prijatno	normalna	slab	nepogodno
11.	sunčano	prijatno	normalna	jak	pogodno
12.	oblačno	prijatno	visoka	jak	pogodno
13.	oblačno	toplo	normalna	slab	pogodno
14.	kiša	prijatno	visoka	jak	nepogodno

Zadatak (rešenje, 1. korak)

- Računamo entropiju skupa primera:
  - Imamo dve klase: pogodno i nepogodno.
  - Imamo ukupno 14 primera.
  - Klasi pogodno pripada 7 primera i klasi nepogodno pripada 7 primera.
- $H = - (7/14) * \log_2(7/14)$
  $- (7/14) * \log_2(7/14)$
  $= - \ln(0,5) / \ln(2) =$ **1**

Zadatak (rešenje, 1. korak)

- Računamo dobit za atribut **izgled vremena**:
  - Broj primera po vrednostima je:
    - **sunčano** -> 5, **oblačno** -> 4, **kiša** -> 5
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **sunčano** -> 4:1, **oblačno** -> 3:1, **kiša** -> 0:5
  - Entropija podskupa za vrednost **sunčano** je:
    - $H = - (4/5) * \log_2(4/5) - (1/5) * \log_2(1/5) =$ **0,72**
  - Entropija podskupa za vrednost **oblačno** je:
    - $H = - (3/4) * \log_2(3/4) - (1/4) * \log_2(1/4) =$ **0,81**
  - Entropija podskupa za vrednost **kiša** je:
    - H = **0**
- $G = 1 - 5/14 * 0,72 - 4/14 * 0,81 =$ **0,51**

Zadatak (rešenje, 1. korak)

- Računamo dobit za atribut **temperatura**:
  - Broj primera po vrednostima je:
    - **toplo** -> 4, **prijatno** -> 6, **hladno** -> 4
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 3:1, **prijatno** -> 3:3, **hladno** -> 1:3
  - Entropija podskupa za vrednost **toplo** je:
    - $H = - (3/4) * \log_2(3/4) - (1/4) * \log_2(1/4) =$ **0,81**
  - Entropija podskupa za vrednost **prijatno** je:
    - H = **1**
  - Entropija podskupa za vrednost **hladno** je:
    - $H = - (1/4) * \log_2(1/4) - (3/4) * \log_2(3/4) =$ **0,81**
- $G = 1 - 4/14*0,81 - 6/14*1 - 4/14*0,81 =$ **0,11**

Zadatak (rešenje, 1. korak)

- Računamo dobit za atribut **vlažnost**:
  - Broj primera po vrednostima je:
    - **normalna** -> 7, **visoka** -> 7
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **normalna** -> 3:4, **visoka** -> 4:3
  - Entropija podskupa za vrednost **normalna** je:
    - $H = - (3/7) * \log_2(3/7) - (4/7) * \log_2(4/7) =$ **0,98**
  - Entropija podskupa za vrednost **visoka** je:
    - $H = - (4/7) * \log_2(4/7) - (3/7) * \log_2(3/7) =$ **0,98**
- $G = 1 - 7/14 * 0,98 - 7/14 * 0,98 =$ **0,02**

Zadatak (rešenje, 1. korak)

- Računamo dobit za atribut **vetar**:
  - Broj primera po vrednostima je:
    - **slab** -> 8, **jak** -> 6
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **slab** -> 4:4, **jak** -> 3:3
  - Entropija podskupa za vrednost **slab** je:
    - H = **1**
  - Entropija podskupa za vrednost **jak** je:
    - H = **1**
- G = 1 - 8/14 * 1 - 6/14 * 1 = **0**

Zadatak (rešenje, 1. korak)

- Izračunate dobiti za pojedine atribute su:
  - Izgled vremena -> 0,51
  - Temperatura -> 0,11
  - Vlažnost -> 0,02
  - Vetar -> 0
- Najbolji atribut za klasifikaciju je:
  - **Izgled vremena** jer ima najveću dobit.
  - Pravimo tri nova podskupa za vrednosti atributa **sunčano**, **oblačno** i **kiša**.

Zadatak (podskup "sunčano")

	Temperatura	Vlažnost	Vetar	KLASA
1.	toplo	visoka	slab	nepogodno
2.	toplo	visoka	jak	pogodno
8.	prijatno	visoka	slab	pogodno
9.	hladno	normalna	slab	pogodno
11.	prijatno	normalna	jak	pogodno

$H = - (4/5) * \log_2(4/5) – (1/5) * \log_2(1/5) =$ **0,72**

Zadatak (podskup "oblačno")

	Temperatura	Vlažnost	Vetar	KLASA
3.	toplo	visoka	slab	pogodno
7.	hladno	normalna	jak	nepogodno
12.	prijatno	visoka	jak	pogodno
13.	toplo	normalna	slab	pogodno

$H = - (3/4) * \log_2(3/4) – (1/4) * \log_2(1/4) =$ **0,81**

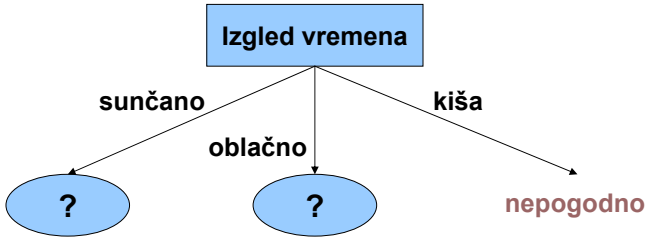
Zadatak (podskup “kiša”)

	Temperatura	Vlažnost	Vetar	KLASA
4.	prijatno	visoka	slab	nepogodno
5.	hladno	normalna	slab	nepogodno
6.	hladno	normalna	jak	nepogodno
10.	prijatno	normalna	slab	nepogodno
14.	prijatno	visoka	jak	nepogodno

H = **0**, pošto svi primeri ovog podskupa pripadaju klasi **nepogodno**, pravi se list stabla sa ovom vrednošću.

Zadatak (stablo odluke, 1. korak)

- Nakon 1. koraka formirano je sledeće stablo odluke:



- U koraku 2. obrađuje se podskup “sunčano”.

Zadatak (rešenje, 2. korak)

- Računamo dobit za atribut **temperatura**:
  - Broj primera po vrednostima je:
    - **toplo** -> 2, **prijatno** -> 2, **hladno** -> 1
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 1:1, **prijatno** -> 2:0, **hladno** -> 1:0
  - Entropija podskupa za vrednost **toplo** je:
    - H = **1**
  - Entropija podskupa za vrednost **prijatno** je:
    - H = **0**
  - Entropija podskupa za vrednost **hladno** je:
    - H = **0**
- G = 0,72 - 2/5 * 1 = **0,32**

Zadatak (rešenje, 2. korak)

- Računamo dobit za atribut **vlažnost**:
  - Broj primera po vrednostima je:
    - **normalna** -> 2, **visoka** -> 3
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **normalna** -> 2:0, **visoka** -> 2:1
  - Entropija podskupa za vrednost **normalna** je:
    - H = **0**
  - Entropija podskupa za vrednost **visoka** je:
    - H = - (2/3) * $\log_2$(2/3) - (1/3) * $\log_2$(1/3) = **0,92**
- G = 0,72 - 3/5 * 0,92 = **0,168**

Zadatak (rešenje, 2. korak)

- Računamo dobit za atribut **vetar**:
  - Broj primera po vrednostima je:
    - **slab** -> 3, **jak** -> 2
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **slab** -> 2:1, **jak** -> 2:0
  - Entropija podskupa za vrednost **slab** je:
    - H = - (2/3) * $\log_2$(2/3) - (1/3) * $\log_2$(1/3) = **0,92**
  - Entropija podskupa za vrednost **jak** je:
    - H = **0**
- G = 0,72 - 3/5 * 0,92 = **0,168**

Zadatak 1. (rešenje, 2. korak)

- Izračunate dobiti za pojedine atribute su:
  - Temperatura -> 0,32
  - Vlažnost -> 0,168
  - Vetar -> 0,168
- Najbolji atribut za dalju klasifikaciju je:
  - **Temperatura** jer ima najveću dobit.
  - Pravimo tri nova podskupa za vrednosti atributa **toplo**, **prijatno** i **hladno**.

Zadatak (rešenje, 2. korak)

Podskup "sunčano-toplo"
H = 1

	Vlažnost	Vetar	KLASA
1.	visoka	slab	nepogodno
2.	visoka	jak	pogodno

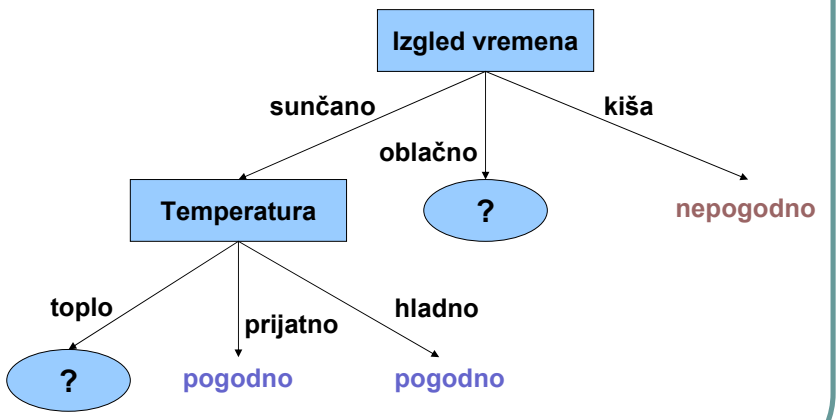
Podskup "sunčano-prijatno"
H = 0

	Vlažnost	Vetar	KLASA
8.	visoka	slab	pogodno
11.	normalna	jak	pogodno

Podskup "sunčano-hladno"
H = 0

	Vlažnost	Vetar	KLASA
9.	normalna	slab	pogodno

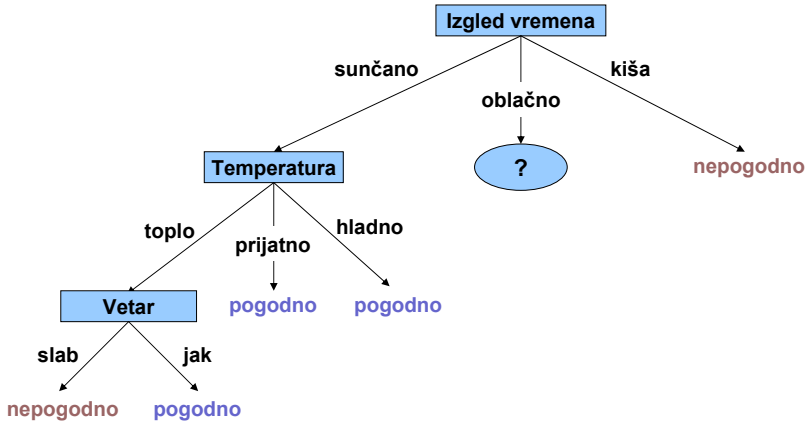
Zadatak (stablo odluke, 2. korak)



Zadatak (rešenje, 3. korak)

- Posmatramo podskup “sunčano-toplo”.
- Dobitak atributa **vlažnost** je:
  - G = 1 - 1 = **0**
- Dobitak atributa **vetar** je:
  - G = 1 - 0 = **1**
- Znači biramo atribut **vetar**:
  - za vrednost **slab** pravimo list **nepogodno**
  - za vrednost **jak** pravimo list **pogodno**

Zadatak (stablo odluke, 3. korak)



Zadatak (podskup “oblačno”)

- U koraku 4. razmatramo podskup “oblačno”:

	Temperatura	Vlažnost	Vetar	KLASA
3.	toplo	visoka	slab	pogodno
7.	hladno	normalna	jak	nepogodno
12.	prijatno	visoka	jak	pogodno
13.	toplo	normalna	slab	pogodno

$$H = - (3/4) * \log_2(3/4) – (1/4) * \log_2(1/4) = \mathbf{0,81}$$

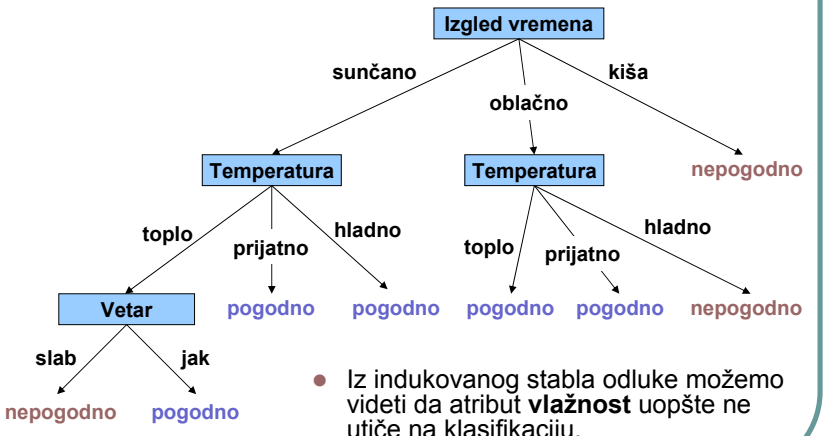
Zadatak (rešenje, 4. korak)

- Računamo dobit za atribut **temperatura**:
  - Broj primera po vrednostima je:
    - **toplo** -> 2, **prijatno** -> 1, **hladno** -> 1
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 2:0, **prijatno** -> 1:0, **hladno** -> 0:1
  - Entropija podskupa za vrednost **toplo** je:
    - H = **0**
  - Entropija podskupa za vrednost **prijatno** je:
    - H = **0**
  - Entropija podskupa za vrednost **hladno** je:
    - H = **0**
- G = 0,81 - 0 = **0,81**
  - Ovo je sigurno najveća dobit pa dalje nećemo da računamo.

Zadatak (rešenje, 4. korak)

- Za vrednost **toplo** pravimo list
  - **pogodno**
- Za vrednost **prijatno** pravimo list
  - **pogodno**
- Za vrednost **hladno** pravimo list
  - **nepogodno**
- **Nakon ovog koraka dobijamo konačno stablo odluke.**

Zadatak (konačno stablo odluke)



- Iz indukovanog stabla odluke možemo videti da atribut **vlažnost** uopšte ne utiče na klasifikaciju.

Indukcija pravila

- Stablo traženja se može predstaviti i u obliku pravila:

1. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = toplo* ***AND*** *vetar = slab* ***THEN*** *vreme-za-košarku = nepogodno*
2. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = prijatno* ***THEN*** *vreme-za-košarku = pogodno*
3. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = hladno* ***THEN*** *vreme-za-košarku = pogodno*
4. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = toplo* ***THEN*** *vreme-za-košarku = pogodno*
5. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = prijatno* ***THEN*** *vreme-za-košarku = pogodno*
6. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = hladno* ***THEN*** *vreme-za-košarku = nepogodno*
7. ***IF*** *izgled-vremena = kiša* ***THEN*** *vreme-za-košarku = nepogodno*

Primena ID3 algoritma

- Primarna namena ID3 algoritma je zamena eksperta u procesu ručnog pravljenja stabla odluke.
- ID3 algoritam se pokazao vrlo efikasnim:
  - Ugrađen je u više komercijalnih sistema za indukciju pravila.