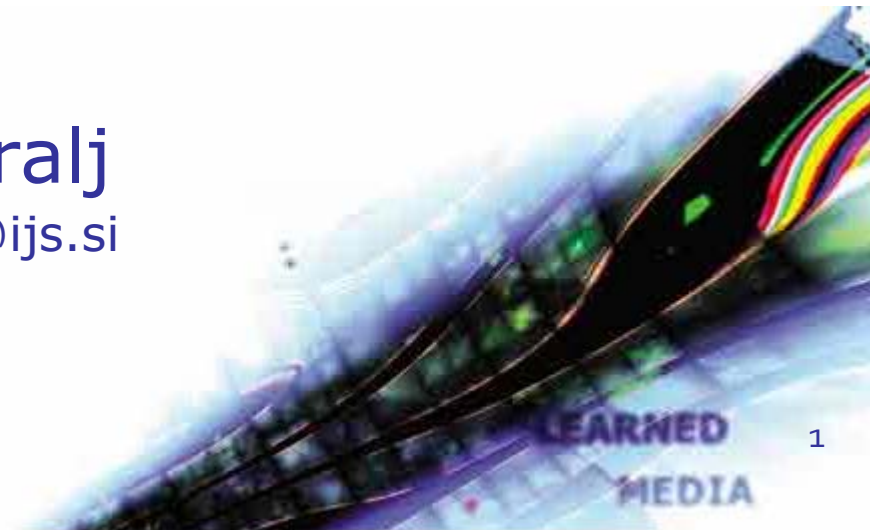
Vaje pri predmetu

# ODKRIVANJE ZAKONITOSTI V PODATKIH

Marec 2007

Petra Kralj
Petra.Kralj@ijs.si




DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

# Načrt vaj

- 16.3.2007: Napovedna indukcija
  - Odločitvena drevesa
  - Naivni Bayesov klasifikator
  - Evalvacija (kontingenčna tabela, klasifikacijska točnost)
  - Napovedna indukcija v Weki
- 23.3.2007: Opisna indukcija in ostalo
  - Asociacijska pravila
  - Opisna indukcija v Weki
  - Povzetek snovi in priprava na izpit
  - Dogovori o seminarskih nalogah

# Gradnja odločitvenih dreves (ID3)

Imamo:

Tabelarične podatke z nominalno ciljno spremenljivko

Določimo učno in testno množico

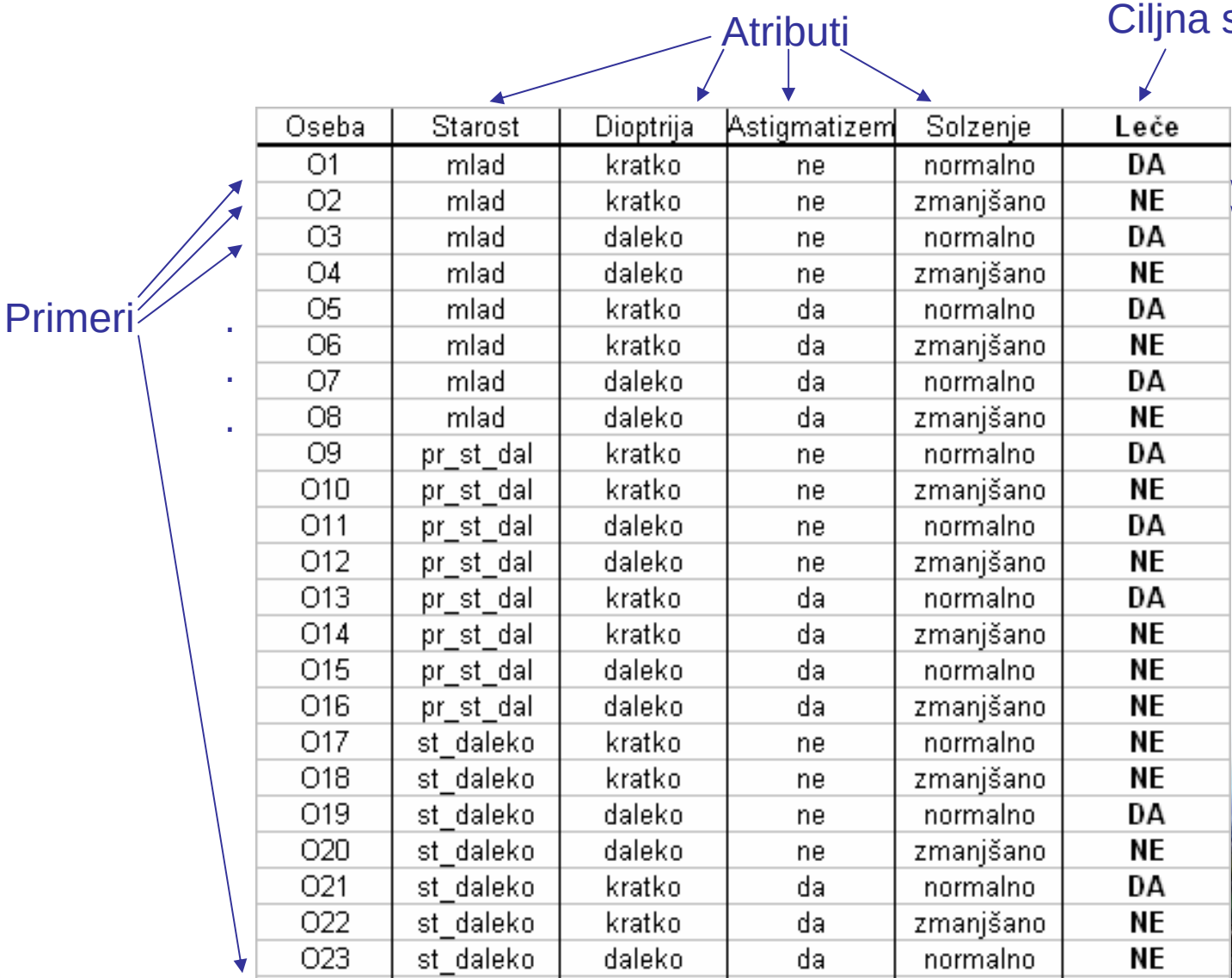
---

Gradimo drevo na učni množici S:

1. Izračunamo entropijo množice E(S)
2. **Če** E(S) = 0
3. Trenutno vozlišče je list, ki klasificira v večinski razred
4. **Če** E(S) > 0
5. Izračunaj informacijski pridobitek vsakega atributa Gain(S, A)
6. Atribut z največjim informacijskim pridobitkom A damo v koren drevesa
7. Množico S razdelimo na podmnožice $S_i$ glede na vrednosti atributa A
8. Na vsaki množici Si ponovimo korake 1-7

Testiramo drevo na testni množici

# Tabelarični podatki




| Oseba | Starost | Dioptrija | Astigmatizem | Solzenje | Leče |
|---|---|---|---|---|---|
| O1 | mlad | kratko | ne | normalno | DA |
| O2 | mlad | kratko | ne | zmanjšano | NE |
| O3 | mlad | daleko | ne | normalno | DA |
| O4 | mlad | daleko | ne | zmanjšano | NE |
| O5 | mlad | kratko | da | normalno | DA |
| O6 | mlad | kratko | da | zmanjšano | NE |
| O7 | mlad | daleko | da | normalno | DA |
| O8 | mlad | daleko | da | zmanjšano | NE |
| O9 | pr_st_dal | kratko | ne | normalno | DA |
| O10 | pr_st_dal | kratko | ne | zmanjšano | NE |
| O11 | pr_st_dal | daleko | ne | normalno | DA |
| O12 | pr_st_dal | daleko | ne | zmanjšano | NE |
| O13 | pr_st_dal | kratko | da | normalno | DA |
| O14 | pr_st_dal | kratko | da | zmanjšano | NE |
| O15 | pr_st_dal | daleko | da | normalno | NE |
| O16 | pr_st_dal | daleko | da | zmanjšano | NE |
| O17 | st_daleko | kratko | ne | normalno | NE |
| O18 | st_daleko | kratko | ne | zmanjšano | NE |
| O19 | st_daleko | daleko | ne | normalno | DA |
| O20 | st_daleko | daleko | ne | zmanjšano | NE |
| O21 | st_daleko | kratko | da | normalno | DA |
| O22 | st_daleko | kratko | da | zmanjšano | NE |
| O23 | st_daleko | daleko | da | normalno | NE |
| O24 | st_daleko | daleko | da | zmanjšano | NE |

# Določimo učno in testno množico

| Oseba | Starost | Dioptrija | Astigmatizem | Solzenje | Leče |
|---|---|---|---|---|---|
| O1 | mlad | kratko | ne | normalno | DA |
| O2 | mlad | kratko | ne | zmanjšano | NE |
| O3 | mlad | daleko | ne | normalno | DA |
| O4 | mlad | daleko | ne | zmanjšano | NE |
| O5 | mlad | kratko | da | normalno | DA |
| O6 | mlad | kratko | da | zmanjšano | NE |
| O7 | mlad | daleko | da | normalno | DA |
| O8 | mlad | daleko | da | zmanjšano | NE |
| O9 | pr_st_dal | kratko | ne | normalno | DA |
| O10 | pr_st_dal | kratko | ne | zmanjšano | NE |
| O11 | pr_st_dal | daleko | ne | normalno | DA |
| O12 | pr_st_dal | daleko | ne | zmanjšano | NE |
| O13 | pr_st_dal | kratko | da | normalno | DA |
| O14 | pr_st_dal | kratko | da | zmanjšano | NE |
| O15 | pr_st_dal | daleko | da | normalno | NE |
| O16 | pr_st_dal | daleko | da | zmanjšano | NE |
| O17 | st_daleko | kratko | ne | normalno | NE |
| O18 | st_daleko | kratko | ne | zmanjšano | NE |
| O19 | st_daleko | daleko | ne | normalno | DA |
| O20 | st_daleko | daleko | ne | zmanjšano | NE |
| O21 | st_daleko | kratko | da | normalno | DA |
| O22 | st_daleko | kratko | da | zmanjšano | NE |
| O23 | st_daleko | daleko | da | normalno | NE |
| O24 | st_daleko | daleko | da | zmanjšano | NE |

30% primerov damo v testno množico

# Testna množica

| Oseba | Starost | Dioptrija | Astigmatizem | Solzenje | Leče |
|---|---|---|---|---|---|
| O3 | mlad | daleko | ne | normalno | DA |
| O9 | pr_st_dal | kratko | ne | normalno | DA |
| O12 | pr_st_dal | daleko | ne | zmanjšano | NE |
| O13 | pr_st_dal | kratko | da | normalno | DA |
| O15 | pr_st_dal | daleko | da | normalno | NE |
| O16 | pr_st_dal | daleko | da | zmanjšano | NE |
| O23 | st_daleko | daleko | da | normalno | NE |

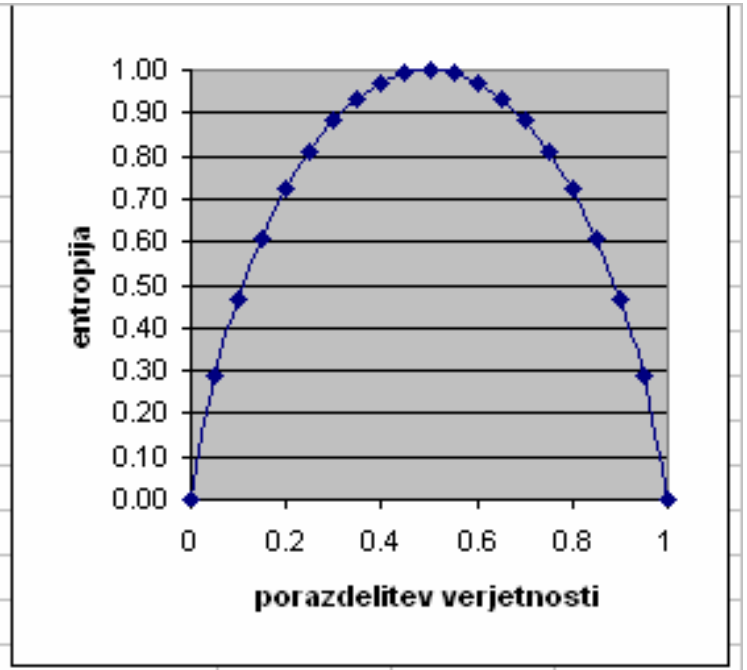
Te podatke damo k strani in jih med postopkom učenja ne gledamo !

# Učna množica

| Oseba | Starost | Dioptrija | Astigmatizem | Solzenje | Leče |
|---|---|---|---|---|---|
| O1 | mlad | kratko | ne | normalno | DA |
| O2 | mlad | kratko | ne | zmanjšano | NE |
| O4 | mlad | daleko | ne | zmanjšano | NE |
| O5 | mlad | kratko | da | normalno | DA |
| O6 | mlad | kratko | da | zmanjšano | NE |
| O7 | mlad | daleko | da | normalno | DA |
| O8 | mlad | daleko | da | zmanjšano | NE |
| O10 | pr_st_dal | kratko | ne | zmanjšano | NE |
| O11 | pr_st_dal | daleko | ne | normalno | DA |
| O14 | pr_st_dal | kratko | da | zmanjšano | NE |
| O17 | st_daleko | kratko | ne | normalno | NE |
| O18 | st_daleko | kratko | ne | zmanjšano | NE |
| O19 | st_daleko | daleko | ne | normalno | DA |
| O20 | st_daleko | daleko | ne | zmanjšano | NE |
| O21 | st_daleko | kratko | da | normalno | DA |
| O22 | st_daleko | kratko | da | zmanjšano | NE |
| O24 | st_daleko | daleko | da | zmanjšano | NE |

# Entropija in informacijski pridobitek

| verjetnost razreda 1 | verjetnost razreda 2 | **entropija $E(p_1, p_2)$ =** |
|---|---|---|
| $p_1$ | $p_2 = 1-p_1$ | $-p_1*\log_2(p_1) - p_2*\log_2(p_2)$ |
| 0 | 1 | 0.00 |
| 0.05 | 0.95 | 0.29 |
| 0.10 | 0.90 | 0.47 |
| 0.15 | 0.85 | 0.61 |
| 0.20 | 0.80 | 0.72 |
| 0.25 | 0.75 | 0.81 |
| 0.30 | 0.70 | 0.88 |
| 0.35 | 0.65 | 0.93 |
| 0.40 | 0.60 | 0.97 |
| 0.45 | 0.55 | 0.99 |
| 0.50 | 0.50 | 1.00 |
| 0.55 | 0.45 | 0.99 |
| 0.60 | 0.40 | 0.97 |
| 0.65 | 0.35 | 0.93 |
| 0.70 | 0.30 | 0.88 |
| 0.75 | 0.25 | 0.81 |
| 0.80 | 0.20 | 0.72 |
| 0.85 | 0.15 | 0.61 |
| 0.90 | 0.10 | 0.47 |
| 0.95 | 0.05 | 0.29 |
| 1 | 0 | 0.00 |

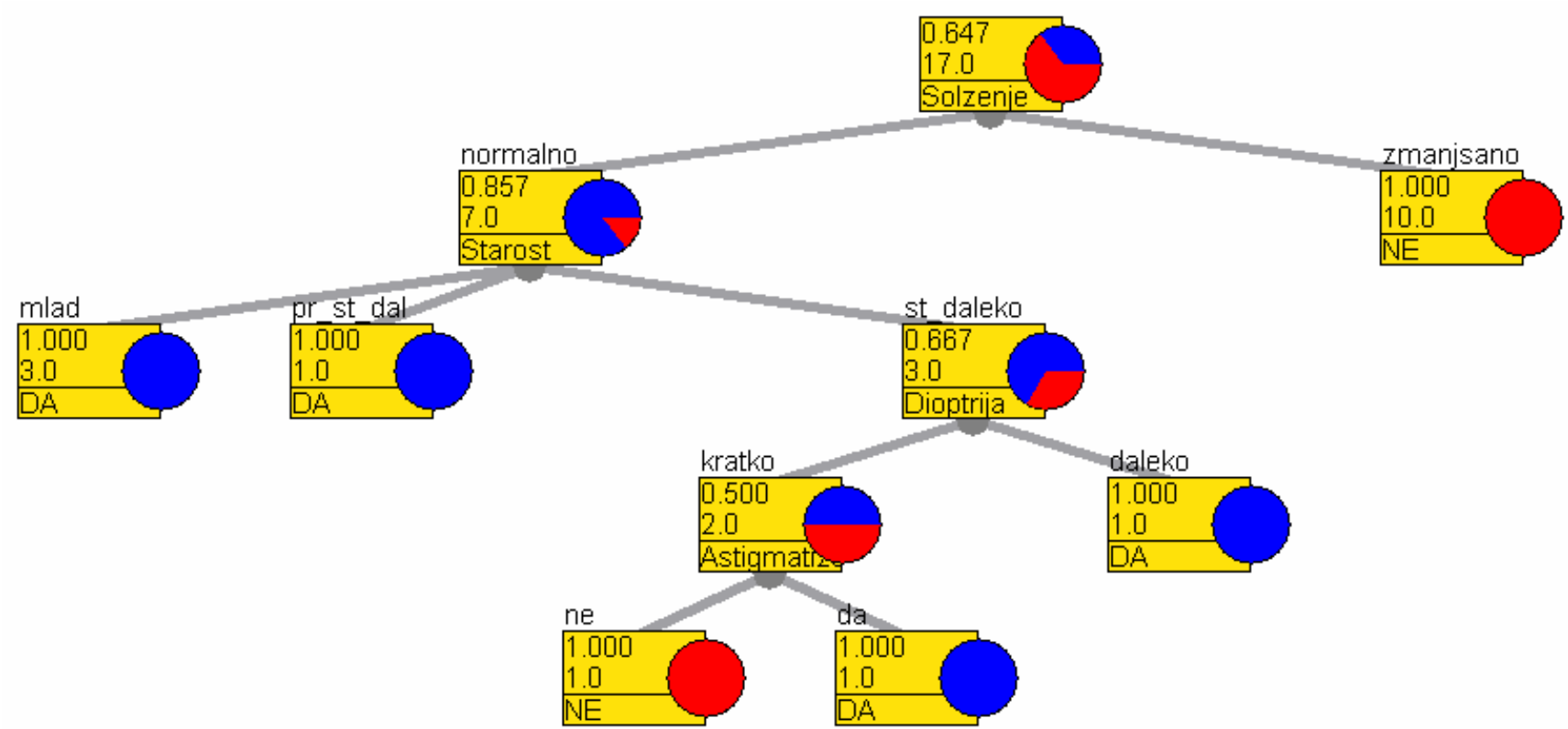



število primerov v podmnožici
verjetnost "veje"
atribut

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

množica
število primerov v množici

# Odločitveno drevo



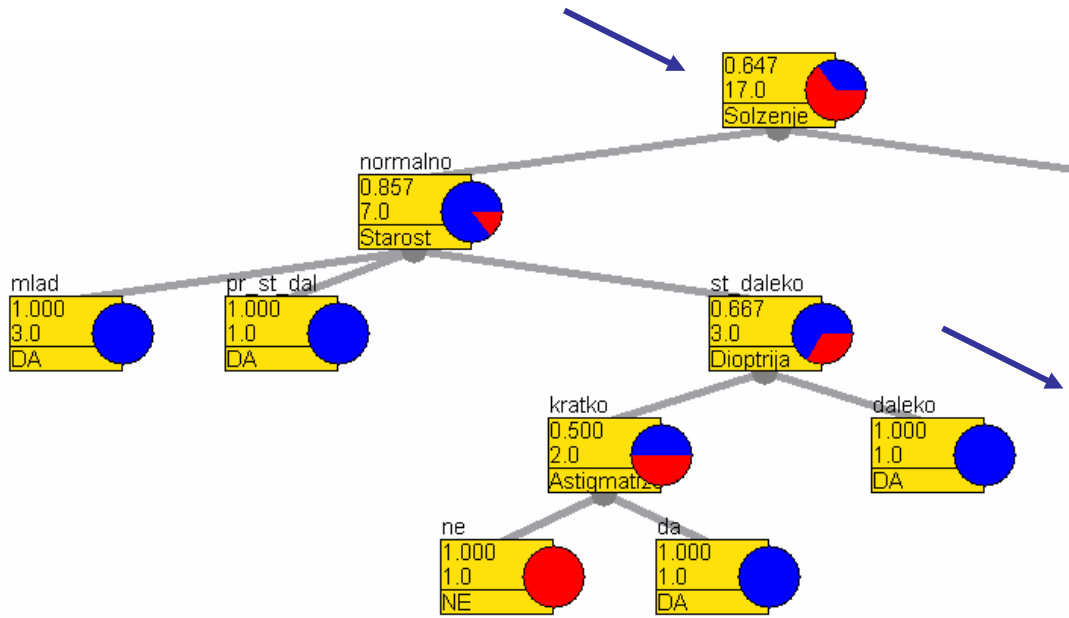
# Kontingenčna tabela

| actual \ predicted | Predicted positive | Predicted negative |
|---|---|---|
| Actual positive | TP | FN |
| Actual negative | FP | TN |

- Kontingenčna tabela primerja napovedane in resnične vrednosti
- Iz kontingenčne tabele lahko izračunamo klasifikacijsko točnost
- Klasifikacijska točnost
  = pravilno napovedani primeri / vsi primeri
  = (TP+TN) / (TP+TN+FP+FN)

# Testiranje klasifikatorja

| Oseba | Starost | Dioptrija | Astigmatizem | Solzenje | Leče |
|---|---|---|---|---|---|
| O3 | mlad | daleko | ne | normalno | DA |
| O9 | pr_st_dal | kratko | ne | normalno | DA |
| O12 | pr_st_dal | daleko | ne | zmanjšano | NE |
| O13 | pr_st_dal | kratko | da | normalno | DA |
| O15 | pr_st_dal | daleko | da | normalno | NE |
| O16 | pr_st_dal | daleko | da | zmanjšano | NE |
| O23 | st_daleko | daleko | da | normalno | NE |




Ca = (3+2)/ (3+2+2+0) = 0,71%

| | Predicted positive | Predicted negative |
|---|---|---|
| Actual positive | TP=3 | FN |
| Actual negative | FP=2 | TN=2 |

# Naivni Bayesov klasifikator

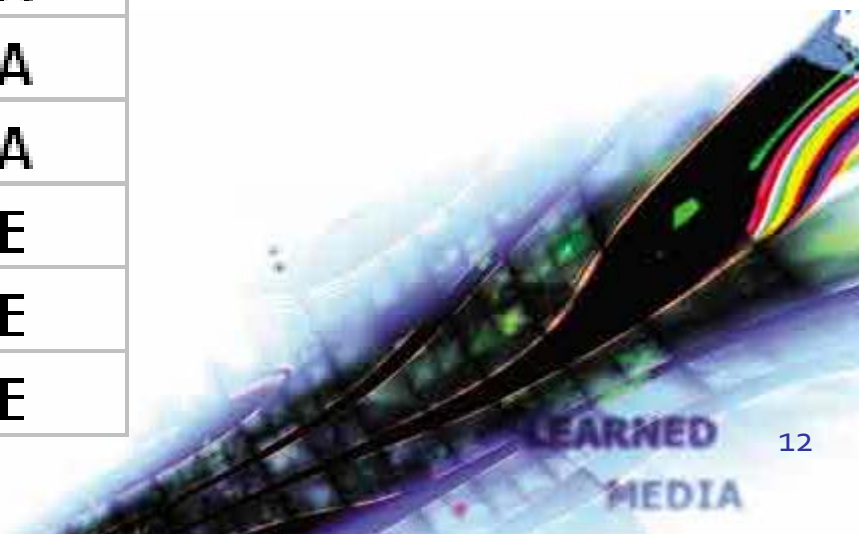
Predpostavlja neodvisnost atributov pri danem razredu

$$P(c \mid a_1, a_2, \ldots . a_n) = P(c) \prod_i \frac{P(c \mid a_i)}{P(c)}$$

Ali bo pajek ujel ti dve mravlji?

- Barva = bela, Čas = noč
- Barva = črna, Velikost = velika, Čas = dan

| Barva | Velikost | Čas | Ujel? |
|---|---|---|---|
| črna | velika | dan | DA |
| bela | mala | noč | DA |
| črna | mala | dan | DA |
| rdeča | velika | noč | NE |
| črna | velika | noč | NE |
| bela | velika | noč | NE |



# V razmislek

- Katere načine za testiranje klasifikatorja poznaš?
- Kako bi računal entropijo za trorazredno ciljno spremenljivko Leče = {trde=4, mehke=5, ne=13}
- Kako bi izračunal informacijski pridobitek zveznega atributa?
- V opisu algoritma ID3 smo kot ustavitveni kriterij uporabili entropija E(S)=0. Kateri kriteriji bi bili še smiselni?
- Kakšna bi bila klasifikacijska točnost drevesa, če bi ga porezali vozlišču *Dioptrija*?
- Kolikšna bi bila cene napačne klasifikacije, če je cena napake FP=5, FN=2
- Primerjajte odločitvena drevesa in naivnega Bayesovega klasifikatorja glede obravnavanja manjkajočih vrednosti.