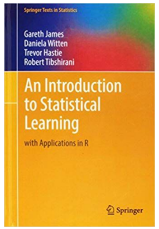
Klasifikacija

(odločitvena drevesa)

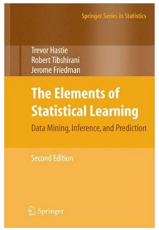
Vsebina

- Odločitvena drevesa
- Ocenjevanje verjetnosti
- Rezanje

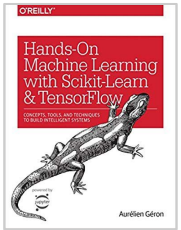
Literatura



Razdelek 8.1

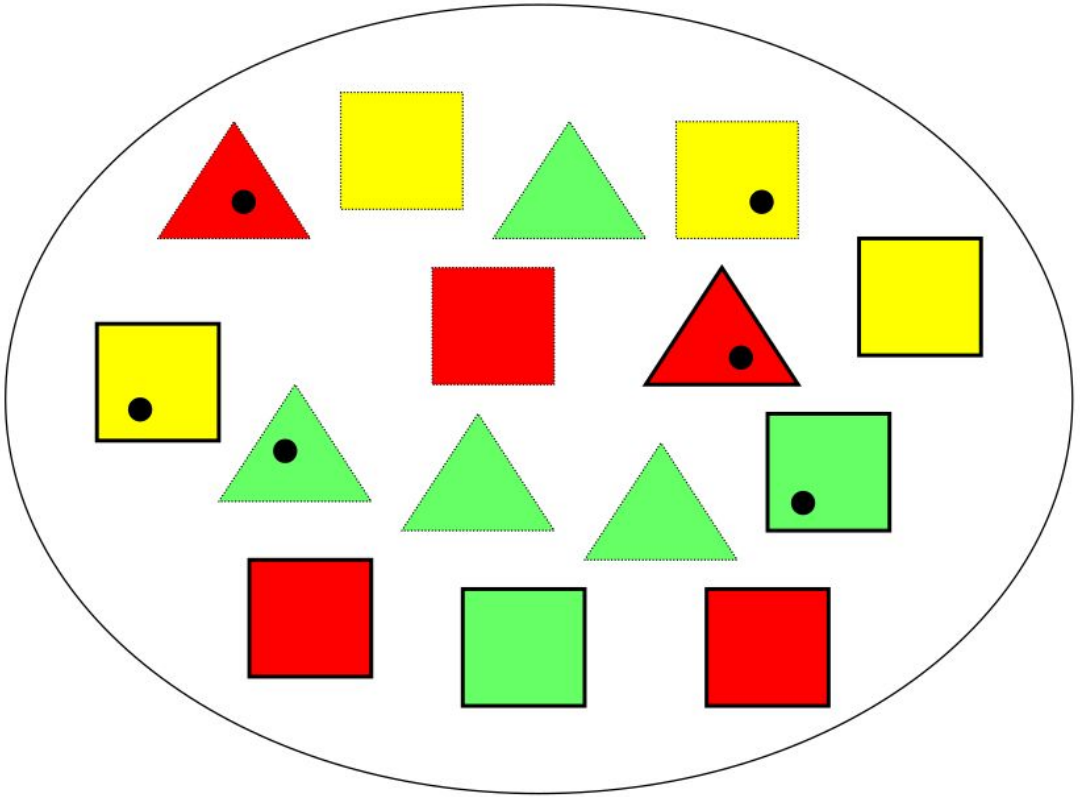


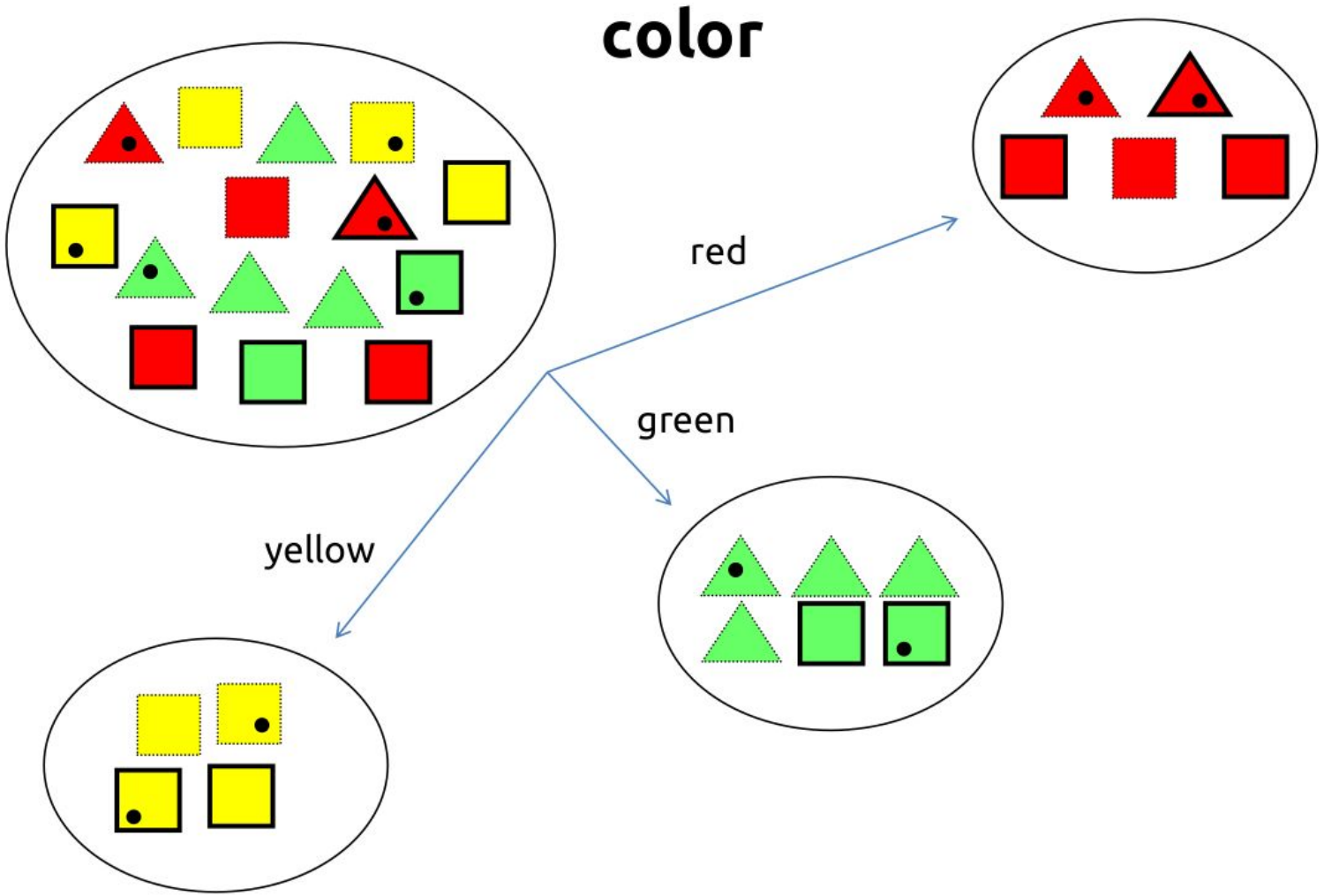
Razdelek 9.2.3



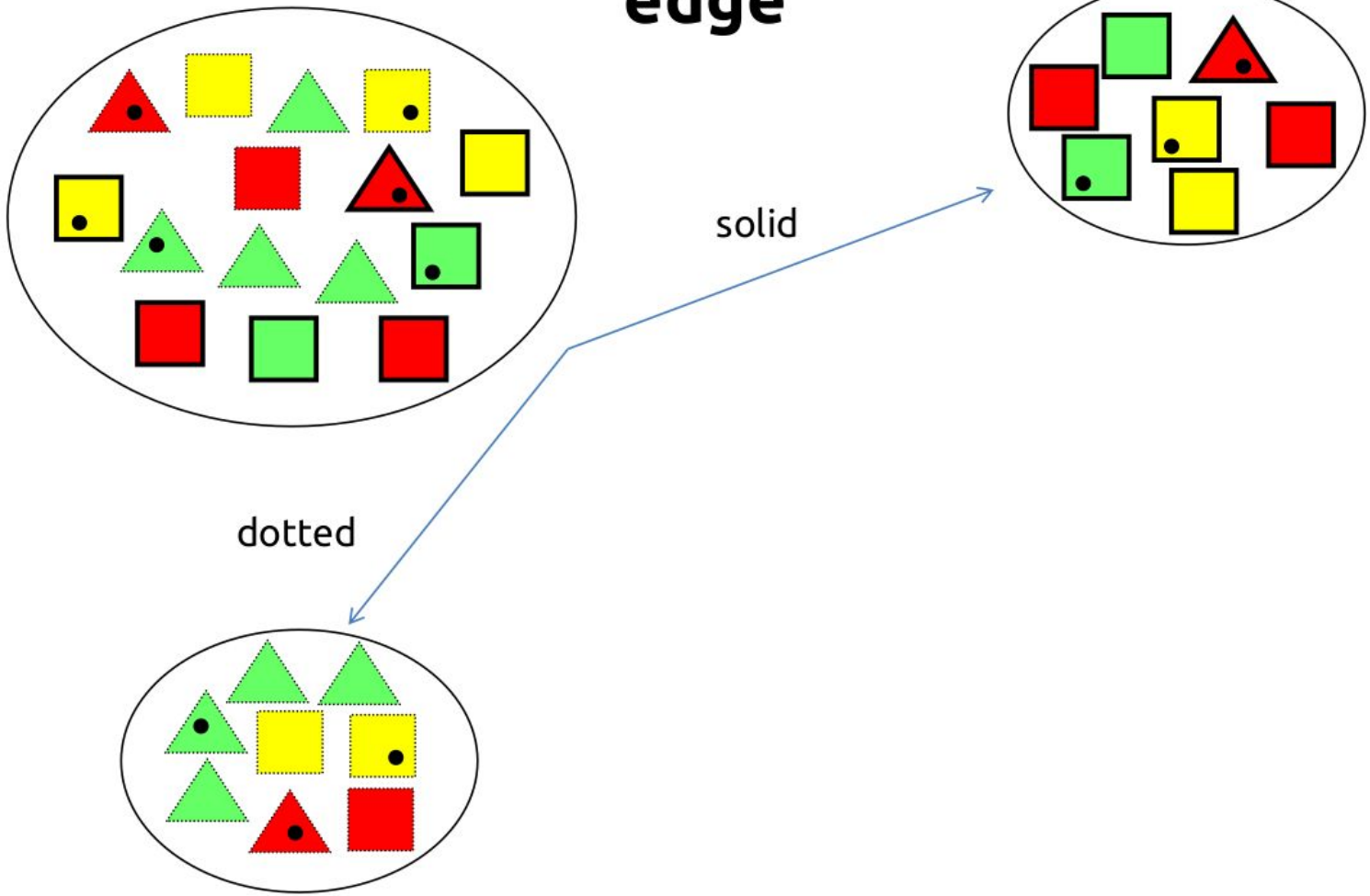
Strani: 162-168

Oblike likov

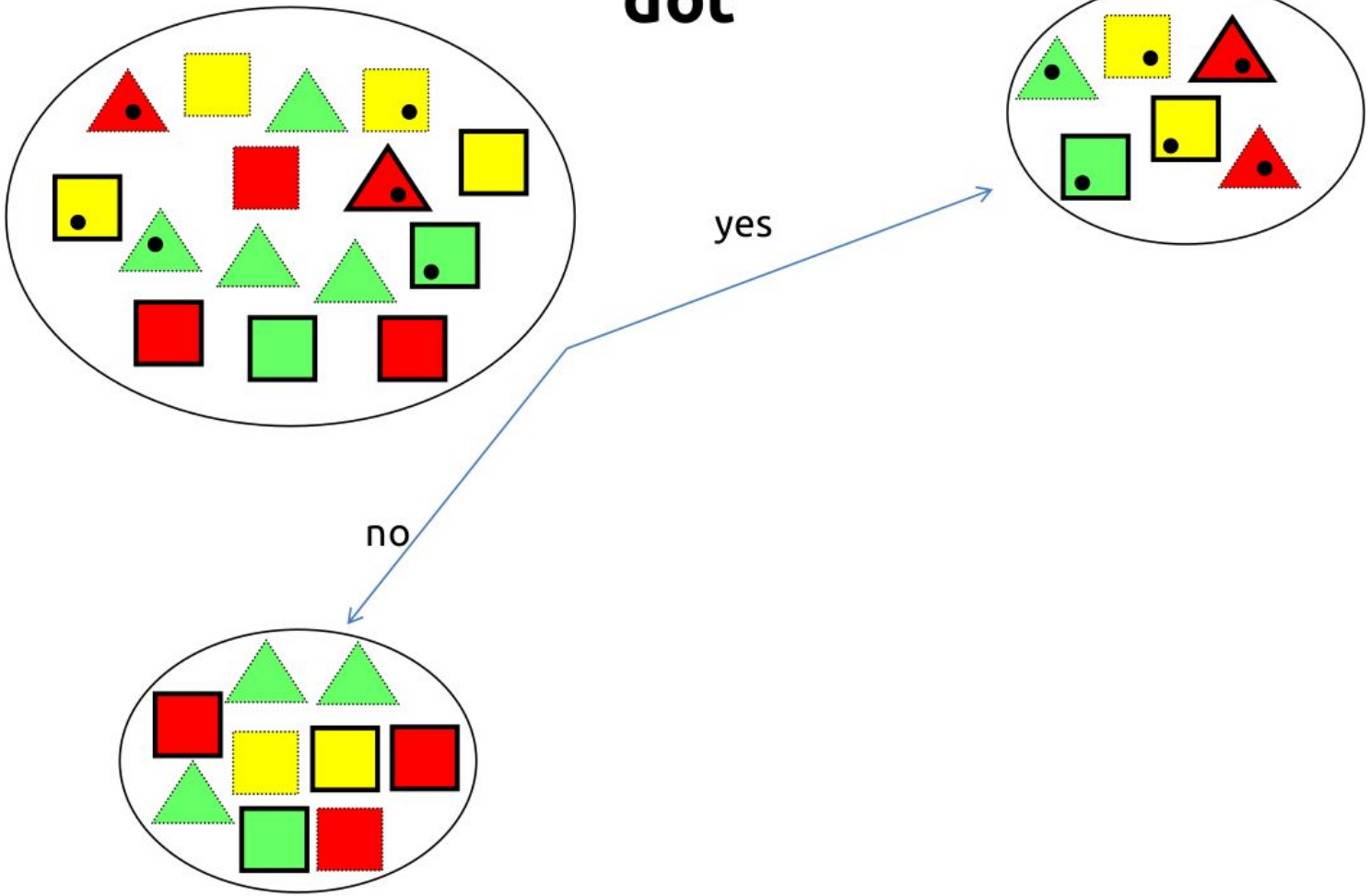


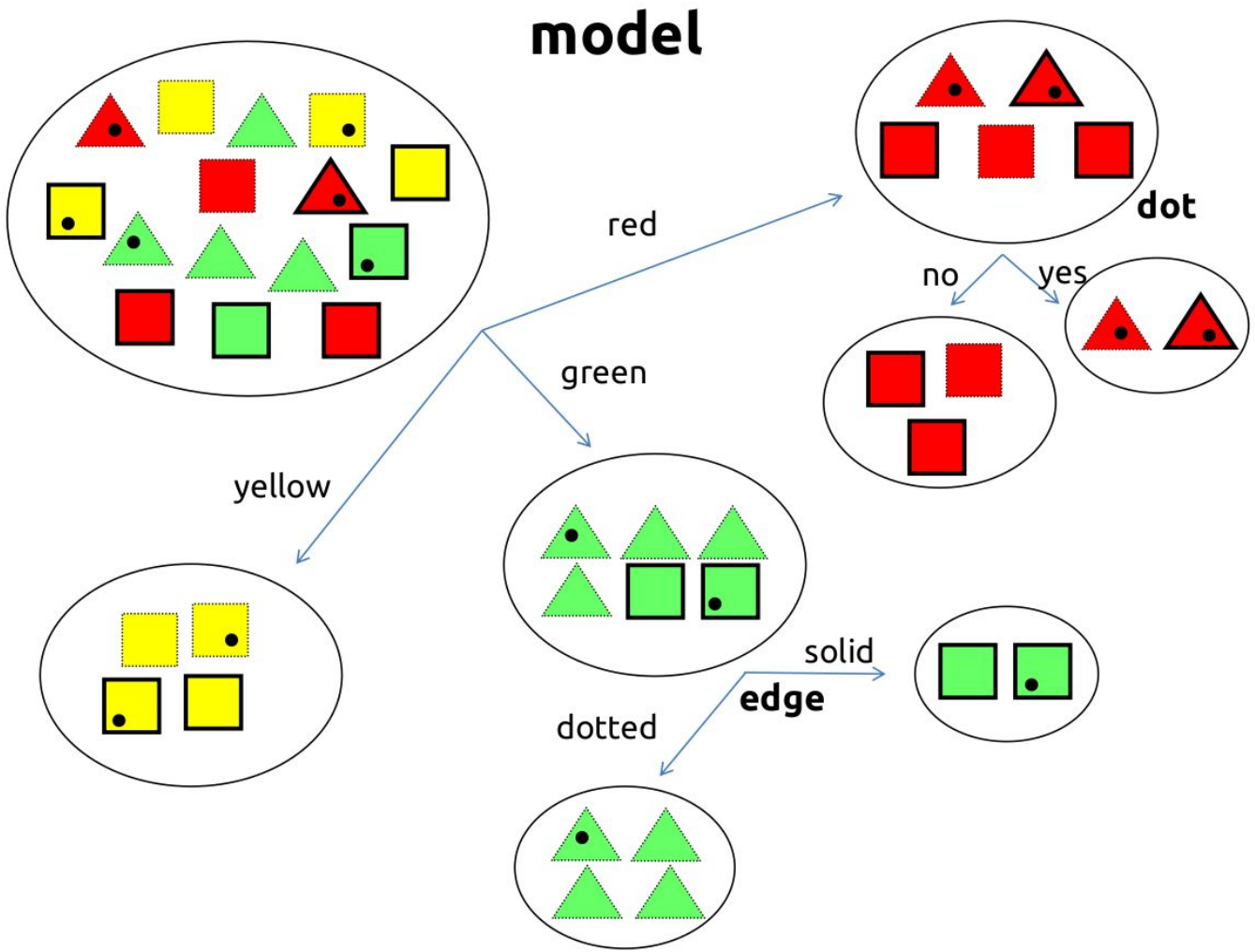

color
red
green
yellow

edge



dot




model
red
dot
no
yes
green
yellow
solid
edge
dotted

Mere nečistoče

Klasifikacijska napaka

Gini indeks

Entropija

Delež primerov z razredom k v vozlišču m

Delež večinskega razreda v vozlišču m

$$\frac{1}{N_m} \sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$$

$$\sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

$$-\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

Informacijski prispevek

I=H(C) ... Entropija pred delitvijo po vrednostih atributa (v vozlišču n)

I_res=Sum p_vi H(C | vi)

InfoGain(A) = I - I_res(A)

najbolj informativen atribut ima max InfoGain

Informacijski prispevek

precenjuje kakovost večvrednostnih atributov; možne rešitve:

- relativni InfoGain (delimo ga z entropijo atributa)

- binarizacija večvrednostnih atributov

- uporaba alternativnih mer

Težave pri učenju dreves

- manjkajoče vrednosti: v splošnem imputacija (npr. manjkajoče vrednosti nadomestimo s povprečjem prisotnih vrednosti atributa). Lahko vpeljemo vrednost "manjkajoč", ki nam morda pomaga razložiti, kaj se dogaja s primeri, kjer meritev atributa manjka.
- binarna delitev boljša kot večvrednostna, ki preveč drobi na majhne podmnožice
- kratkovidnost požrešnega algoritma (XOR)
- šumni podatki...

Rezanje dreves

- Nepopolni podatki, (merske) napake v podatkih
- Učenje šuma, namesto učenja dejanske funkcije, ki generira podatke
- Slaba razumljivost dreves
- pretirano prilagajanje => nižja klasifikacijska točnost na testnih podatkih

Rezanje naprej

- omejevanje št. primerov v vozlišču
- ustavljanje gradnje pri doseženi želeni točnosti v vozlišču

Rezanje nazaj

Postopek MEP (Minimal Error Pruning)

Cilj: poreži drevo tako, da bo ocenjena klasifikacijska točnost maksimalna

Za vsako vozlišče v izračunamo:

- statično napako
- vzvratno napako

Režemo pod v, če je statična napaka manjša od vzvratne.

Ocenjevanje verjetnosti

Točnost T = verjetnost pravilne klasifikacije.

Napaka = 1 - T

N ... število vseh primerov, n ... število uspešnih poskusov

- relativna frekvenca: p = n/N

- m-ocena: p = (n + pa*m)/(N+m)

    ekspert zaupa v pa => velik m, sicer majhen m (tipično m=2)

- Laplace: p = (n+1)/(N+k)