Rozpoznávanie obrazcov
šk.r. 2016-17

Rozhodovacie stromy

Zuzana Berger Haladova

Bayesovský klasifikátor

- Príznakový vektor klasifikujeme do triedy, ktorá má najväčšiu aposteriórnu pravde-podobnosť - $\mathrm{d}(\mathbf{x}) = \omega_i \leftrightarrow P(\omega_i|\boldsymbol{x})$ je maximálne
- Ako určíme apriórne pravdepodobnosti $P(\omega_i)$ a $P(\boldsymbol{x}|\omega_i)$ a združenú $P(\boldsymbol{x})$?
- Naivný Bayesovský klasifikátor – príznaky sa navzájom neovplyvňujú

Bayesovský klasifikátor

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$$

Pre *kategorické atribúty*

$$P(\mathbf{x}\,|\,\omega_i) = \prod_{k=1}^{d} P(x_k\,|\,\omega_i)$$

Naivný klasifikátor

$$P(x_k\,|\,\omega_i) = \frac{N^{i,k}}{N^i}$$

Pravdepodobnosť triedy ω_i je $\frac{N^i}{N}$

N # všetkých príkladov

N^i # tých príkladov, ktoré patria do triedy ω_i

$N^{i,k}$ # tých príkladov, ktoré patria do triedy ω_i a príznak k má hodnotu x_k

Bayesovský klasifikátor

X =
(age <=30,
Income = medium,
Student = yes
Credit_rating = Fair)

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Bayesovský klasifikátor

```
load fisheriris

X = meas;

Y = species;

Mdl = fitcnb(X,Y);

Mdl.Prior
```

Rozhodovacie stromy

- Rozhodnutie o zaradení triedy a postup, ako sme k nemu prišli, sú usporiadané do stromu
- Uzly (vrcholy stromu)
  - opisujú testy hodnôt jednotlivých príznakov
  - z uzlov vychádza toľko vetiev, koľko rôznych hodnôt test nadobúda (väčšinou binárne)
- Listy
  - klasifikačné triedy

Rozhodovacie stromy II

1. vezmi všetky nepoužité príznaky, ohodnoť ich

2. do uzlu vezmi príznak s najlepším ohodnotením (pomocou entropie a vzájomnej informácie)

3. pre každú hodnotu príznaku vytvor podmnožinu dát

...

Rozhodovacie stromy - príklad

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Rozhodovacie stromy a MATLAB

- Opäť vezmeme Fischerovu databázu iris
- Má tri triedy a štyri príznaky SL, SW, PL, PW, pričom dáta sú kategorické
- tc = fitctree(X,y)

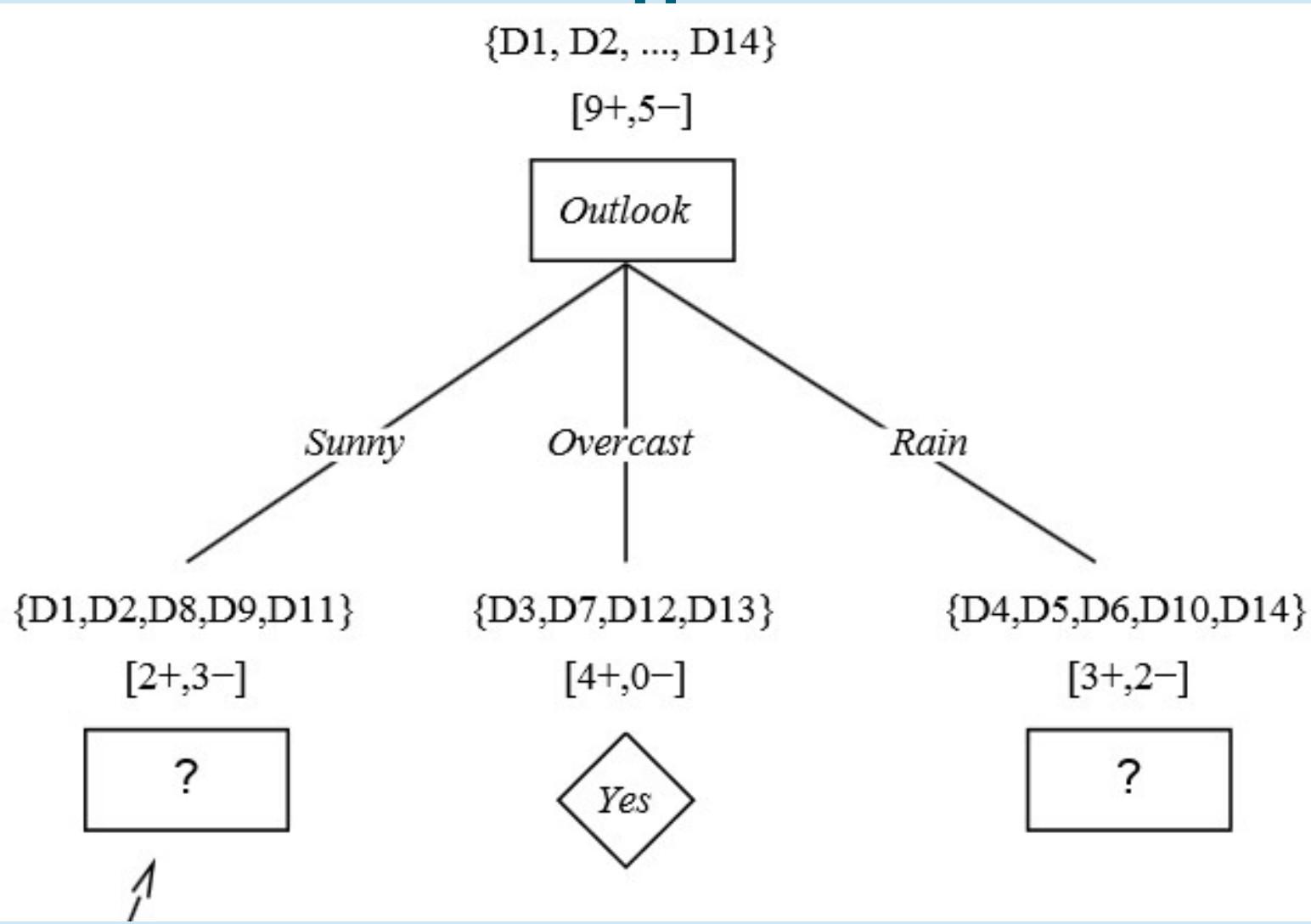
Rozhodovacie stromy a MATLAB

```
MdlDefault = fitctree(X,y);

view(MdlDefault,'Mode','graph')

label=predict(MdlDefault, New)
```

Rozhodovacie stromy – príklad



Rozhodovacie stromy a MATLAB

- Klasifikačné stromy dobre zodpovedajú trénovacej množine, ale nie vždy fungujú na testovacej množine
- Preto niekedy zmenšujeme strom pomocou príkazu prune
- Zadáme pruned = prune(t,'level',1)
- View(pruned) a vidíme výsledok

Porovnanie klasifikátorov

https://plot.ly/~jackp/16209/machine-learning-classifier-comparison.embed