23. Odabir značajki

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.1

1 Motivacija i pristupi

- Metode za **smanjenje dimenzionalnosti** ulaznog prostora:
  - **Odabir značajki** (*feature selection*) – odabir podskupa izvornih značajki
  - **Transformacija značajki** – izvođenje novih značajki iz izvornih značajki
- Svrha:
  - Uklanjanje **irelevantnih** i **redundatnih** značajki povećava točnost modela
  - Lakše razumijevanje i objašnjavanje modela
  - Pomoć u vizualizaciji podataka
- Odabir značajki čuva izvornu semantiku značajki ⇒ bolja tumačivost modela

2 Univarijatni filtar

- Procjena intrinsične vrijednosti (*merit*) svake značajke pa odabir po pragu ili rangu
- Prednosti: dobro skalira, računalno jednostavno, nezavisno od modela
- Nedostatci: nezavisno od modela, ne uzima u obzir interakciju između značajki
- Ideja: značajka x_k je **relevantna** $\Leftrightarrow$ postoji **zavisnost** između varijabli x_k i y
- **Uzajamna informacija** – zavisnost varijabli x i y kao odstupanje $P(x,y)$ od $P(x)P(y)$:

$$I(x,y) = D_{\mathrm{KL}}\big(P(x,y)||P(x)P(y)\big) = \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x)P(y)}$$

  ⇒ relevantnost značajke x_k za klasu y proporcionalna je sa $I(x_k, y)$
- **t-test** (primjenjivo za $K = 2$)
  - Test značajnosti razlike srednje vrijednosti od x_k za klase $y = 0$ i $y = 1$
  - Hipoteza H_0: srednje vrijednosti su jednake

– t-statistika (pod H_0 distribuirana po t-distribuciji):

$$t = \frac{\bar{x}_k^0 - \bar{x}_k^1}{\hat{\sigma}_i \sqrt{\frac{1}{N_0} + \frac{1}{N_1}}} \sim t(N_0 + N_1 - 2)$$

gdje $N_y = \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = y\}$ i $\bar{x}_k^y = \frac{1}{N_y} \sum_{i=1}^{N} x_k^{(i)} \mathbf{1}\{y^{(i)} = y\}$

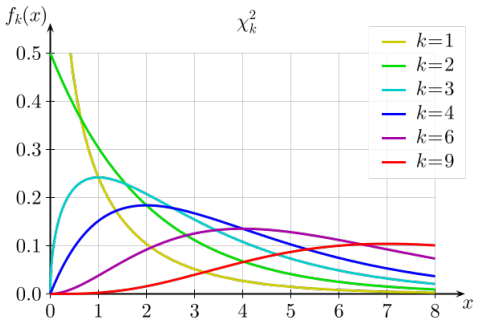
$\Rightarrow$ relevantnost značajke x_k obrnuto je proporcionalna **p-vrijednosti**

- **ANOVA** (za $K > 2$)
  - Testiranje razlika srednjih vrijednosti značajke x_k kroz K klasa
- χ^2**-test** (primjenjivo za kategoričke značajke)
  - Hipoteza H_0: varijable x_k i y su nezavisne ($x_k \perp y$)
  - N – broj primjera, K – broj klasa, K_k – broj vrijednosti varijable x_k
  - **Tablica kontingencije** dimenzije $K_k \times K$ sadrži opažene frekvencije $O_{i,j}$
  - Izračun očekivanih frekvencija ($E_{i,j}$) uz pretpostavku H_0:

$$P(x_k = i) = \sum_j P(x_k = i, y = j)$$
$$P(y = j) = \sum_i P(x_k = i, y = j)$$
$$E_{i,j} = N P(x_k = i) P(y = j)$$

  - χ^2-statistika (pod H_0 distribuirana po χ^2-distribuciji):

$$\chi^2 = \sum_{i=1}^{K_k} \sum_{j=1}^{K} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2\big((K_k - 1)(K - 1)\big)$$



$\Rightarrow$ relevantnost značajke x_k obrnuto je proporcionalna **p-vrijednosti**

- **p-mjera** – neparametarska usporedba srednjih vrijednosti x_k za klase $y = 0$ i $y = 1$:

$$\mathrm{p}(x_k, y) = \frac{\bar{x}_k^0 - \bar{x}_k^1}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

$\Rightarrow$ relevantnost značajke x_k proporcionalna je vrijednosti p-mjere

- **RELIEF** (Kira i Rendell, 1992) – neparametarska iterativna metoda (za $K = 2$)
  - Iterativno ugađanje vektora relevantnosti svih n značajki (vektor $\mathbf{w}$)
  - Slučajan odabir pivotnog primjera i primjera iste (*hit*) i različite klase (*miss*)
  - Relevantnost x_k pada ako primjeri istih klasa imaju različite vrijednosti
  - Relevantnost x_k raste ako primjeri različitih klasa imaju različite vrijednosti

Algoritam RELIEF

postavi $w_k \leftarrow 0$ za svaku značajku $k = 1, \ldots, n$
za $i = 1, \ldots, m$ **radi**: -- *m je broj iteracija*
  nasumično odaberi primjer $\mathbf{x} \in \mathcal{D}$
  pronađi najbliži pogodak $\mathbf{x}^h \in \mathcal{D}$ i promašaj $\mathbf{x}^m \in \mathcal{D}$ (po L2-normi)
  za $k = 1, \ldots, n$ **radi**:
    $w_k = w_k - \frac{1}{N}(x_k - x_k^h)^2 + \frac{1}{N}(x_k - x_k^m)^2$

3 Multivarijatni filtar

- Univarijatne metode ocjenjuju relevantnost, neovisno o redundanciji značajki
- Multivarijatne metode ocjenjuju relevantnost i redundantnost skupa značajki
- **Uklanjanje značajki faktorom inflacije varijance (VIF)** (*variance inflation factor*)
  - Ideja: x_k je redundantna $\Leftrightarrow$ može ju se dobro predvidjeti iz drugih varijabli
  - Model linearne regresije sa x_k kao zavisnom varijablom:

$$h_k(x_k; \mathbf{w}) = w_1 x_1 + \cdots + w_{k-1} x_{k-1} + w_{k+1} x_{k+1} + \cdots + w_n x_n$$

  - VIF varijable x_k:

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2} \in [1, \infty)$$

    gdje je R_k^2 **koeficijent determinacije** za h_k (v. odjeljak 5.1.3 dodatka skripti)
  - U praksi, značajke za koje VIF $\geqslant 10$ smatraju se redundantnima
  - Iterativno uklanjanje redundantnih značajki i ažuriranja VIF vrijednosti
  - VIF uklanja isključivo redundante značajke (ne odabire relevantne značajke)

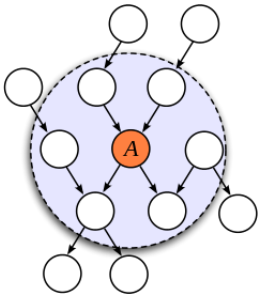
Postepeno (stepwise) uklanjanje značajki VIF-om

1: $S \leftarrow \{1, \ldots, n\}$
2: **za** $k \in S$ **radi:**
3: izračunaj VIF_k sa $S \backslash \{k\}$ kao nezavisnim varijablama
4: $m \leftarrow \text{argmax}_{k \in S} \text{VIF}_k$
5: **dok** $\text{VIF}_m \geqslant 10$ **radi**:
6: $S \leftarrow S \backslash \{m\}$
7: **za** $k \in S$ **radi:**
8: izračunaj VIF_k sa $S \backslash \{k\}$ kao nezavisnim varijablama
9: $m \leftarrow \text{argmax}_{k \in S} \text{VIF}_k$

- **Correlation feature selection (CFS)** – nalazi relevantne i neredundantne značajke
  - Ocjena vrijednosti **podskupa značajki** S koji sadrži k značajki:

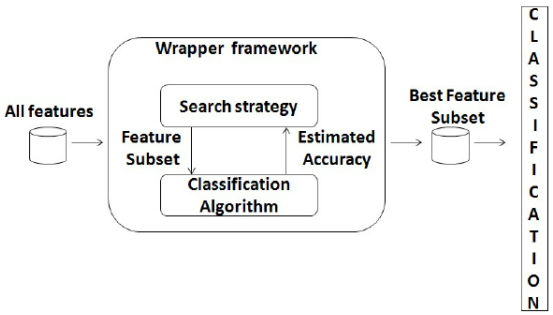
$$\text{Merit}_S = \frac{k\bar{r}_{x,y}}{\sqrt{k + k(k-1)\bar{r}_{x,x}}}$$

  - $\bar{r}_{x,y}$ – prosječna korelacija (npr. Pearsonova) između varijabli iz S i varijable y
  - $\bar{r}_{x,x}$ – prosječna korelacija između svih k varijabli iz S
  - **Unaprijedno pretraživanje** prostora od 2^n podskupova metodom **najbolji prvi**
- **Markovljev pokrivač** (*Markov blanket*) – izravan odabir značajki za PGM-ove
  - Markovljev pokrivač od x_k: roditelji od x_k, njegova djeca i roditelji djece
  - Vrijednost varijable x_k u Bayesovoj mreži ovisi samo o Markovljevom pokrivaču

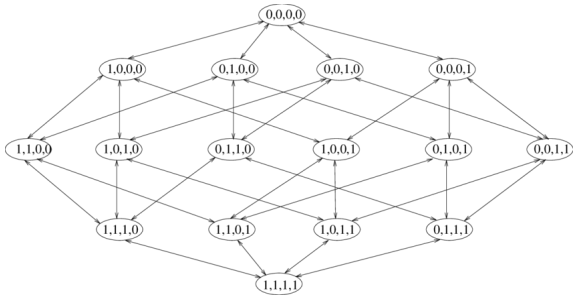


4 Metoda omotača

- Pretraživanje prostora od 2^n podskupova značajki i provjera točnosti modela



- Kriterijska funkcija:
  - **Točnost modela** procijenjena unakrsnom provjerom
  - Mjera **prikladnosti modela** (*goodness of fit*) (npr. F-test)
- Pretraživanje:
  - **Unaprijedni odabir** – kreće od praznog skupa i dodaje značajke
  - **Unatražni odabir** – kreće od svih značajki i uklanja značajke
  - **Stepenast odabir** (*stepwise*) – unaprijedan odabir s unatražnim uklanjanjem



- Prednost: prilagođenost konkretnom modelu; nedostatak: računalna složenost

5 Ugrađene metode

- Neki algoritmi odabiru značajke pri postupku učenju: **stabla odluke**, **slučajne šume**
- Svaki algoritam s **L1-regularizacijom** implicitno provodi odabir značajki