PDV - 16.10.2026

piatok 16. októbra 2026 10:58

Lin. regresia (dodatok)

- výhody: – jednoduchá ⇒ spätkliváš
  ⇒ relat. stabilné (lineárne) správanie
  ⇒ dobre interpretovateľná
  ⇒ dobre preskúmaná
  - vie zachytiť nelinearitu v x-ocl
- nevýhody - nevie nelinearitu v parametroch (v β)
  - nemusí predikovať tak dobre, ako zložitejšie modely

Kvalita modelov (pokr.)

$y \approx X\beta$ ... $y_i \approx x_i^T \beta$, $i = 1, \dots, n$

celk. chyba: $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- kvalita predikcie → niečo iné
  (ako dobre generalizuje na nové dáta)

$(x_1, y_1), \dots, (x_n, y_n)$: trénovacia vzorka
: $\{1, \dots, n\} = Train$
: $n_{train} := n$

⇒

↳ vypočítame $\hat{\beta}$, trénovacia chyba $= \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} (y_i - \hat{y}_i)^2 = MSE_{train}$
mean squared error

alterm.: $RMSE = \sqrt{MSE}$
↑ root MSE

testovacia vzorka: (x_i, y_i) pre $i \in Test$
: # pozorovaní: n_{test}

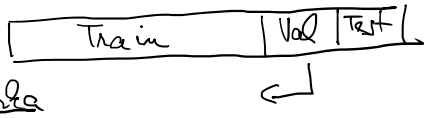
↳ kvalita modelu $= MSE_{test} = \frac{1}{n_{test}} \sum_{i \in Test} (y_i - \hat{y}_i)^2$
skut. váha ↑ ↑ predikovaná váha $\hat{y}_i = x_i^T \hat{\beta}$

(napr.: 80% - 20% train-test)

- testovacia vzorka → na finálne posúdenie kvality
  → model ju predtým nesmie vidieť

výber modelu:

1.) validácia: imitujeme Test
⇒ na 3 časti (napr. 70 - 20 - 10 %)

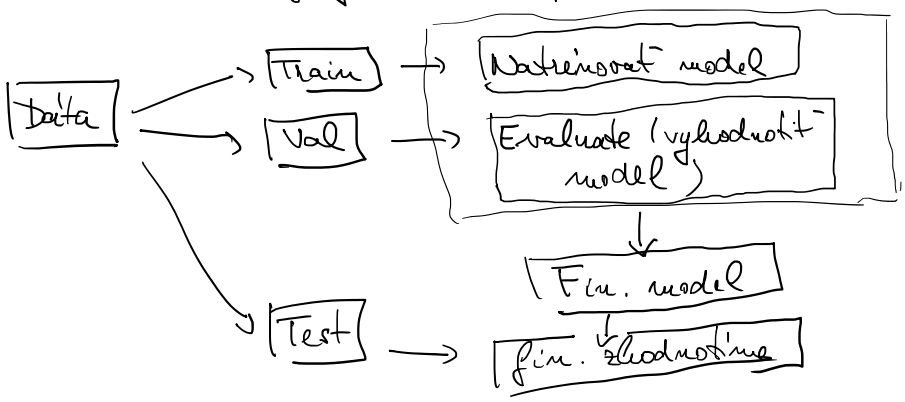


↳ validačná vzorka

Pre každý model: (i) na trénovacej: vypočítame $\hat{\beta}$
(ii) na validačnej: vypočítame $MSE_{val} = \frac{1}{n_{val}} \sum_{i \in Val} (y_i - \hat{y}_i)^2$

Tak každý model: (i) na trénovacej: vypočítame |>
(ii) na validačnej: vypočítame $MSE_{val} = \frac{1}{n_{val}} \sum_{i \in Val} (y_i - \hat{y}_i)^2$

→ vyberieme ten s najmenšou MSE_{val}
→ na testovacej vyhodnotíme jeho kvalitu (MSE_{test}, $RMSE_{test}$)

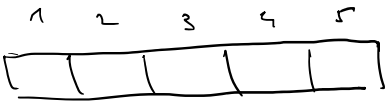


Načo je Test?

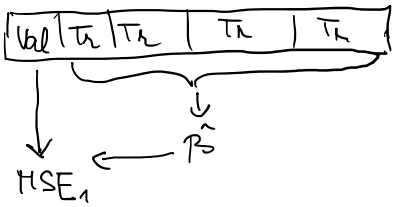
↳ nie je povinný – ale nevyhnutne vedieť kvalitu fin. modelu
↳ iba na posúdenie fin. kvality – lebo validačná chyba fin. modelu je typicky nižšia ako reálna chybovosť (lebo minimalizoval val. chybu – tak sme ho vybrali)

- niekedy: Val-Test: značia naopak
- problémy validácie – nestabilné
  – stratíme časť dát.

2.) krosvalidácia (cross-validation)
– opakovaná validácia
– už sme odčlenili Test, ostáva Train
– K-fold crossvalidation: rozdelíme Train na K častí (K=5)



časť – ona bude raz Validačná a ostatné trénovacie
– vypočítame MSE_{val}

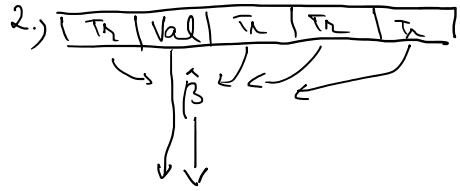
1.)


$výška_i = \beta_0 + \beta_1 výška_i$

$\Rightarrow MSE_{11}, MSE_{21}, \ldots, MSE_K$

$\Rightarrow$ Fin. chyba $= \frac{1}{K}(MSE_1 + \ldots + MSE_K)$

↳ vyberieme model s najmenšou chybou

2.)


Modely: 1) $y_i \approx \beta_0 + \beta_1 x_i$ (x_i: výška) → CV: chyba (1)
2) $y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ → CV: chyba (2)
3) $y_i \approx \beta_0 + \beta_1 z_i$ → CV: chyba (3)

} vyberieme najmenšiu

(1) $y_i \approx \beta_0 + \beta_1 x_i + \beta_2 x_i^2$ —> CV : chyba (2) } vyznením najmenšia

3) $y_i \approx \beta_0 + \beta_1 z_i$ —> CV : chyba (3)
z ... druhá päta

— Fin. model → $\hat{\beta}$ natrénujem na celej vzorke (Train Test)
→ nestrácame dáta → numericky náročnejšia
→ stabilnejšie : každé dáta bolo dvakrát vo validácii

3.) Špeciálne pre lin. reg. : namiesto Val / CV

- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
↑ residual / total sum of squares
→ $\in [0, 1]$, čím bližšie k 1, tým lepšie
nech sa použiť na výber modelu — vyberá vždy komplexnejší!
- adjusted R^2 — zohľadňuje aj # premenných
(adjusted coefficient of determination)
- na výber premenných (správneho lin. modelu) častejšie : informačné kritériá (IC)

Akaike inf. criterion = $AIC = \ln\left(\frac{1}{n} \cdot RSS\right) + \frac{2p}{n}$ ← # premenných, ← # pozorovaní

Bayes inf. cr. = $BIC = \ln\left(\frac{1}{n} RSS\right) + \frac{p \ln(n)}{n}$

$\frac{1}{n} RSS = MSE_{train}$ ↑ ako nedobrý je model ↑ ako komplexný je model (penalta za komplexnosť)

→ veľmi ľahko sa počítajú, netreba odkladať dáta
→ AIC, BIC : čo najmenšie
→ nevýhoda : nepočítajú priamo predikčnú chybu
→ fungujú pre lub. model s likelihoodom

- CV pre $K = n$: leave-one-out cross validation

