RDD - 11

03 May 2021 14:50

- na stránke - skúškové otázky

---

dotaz: extrakcia premenných → skonštruovať q nových Y z pôvodných k X-ov (1. a 2.)

(3.) Metódy selekcie premenných (feature selection)

- už nie konštrukcia nových, ale: výber podmnožiny $\{1, \dots, k\}$

→ pre potreby nejakého modelu

tu: lineárna regresia: $y = X\beta + \varepsilon$ (rozmery: $n \times 1$, $n \times k$, $k \times 1$, $n \times 1$)

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

i-ty riadok: $y_i = x_i^T \beta + \varepsilon_i$

$$X = \begin{pmatrix} 1 & \dots \\ \vdots & \\ 1 & \end{pmatrix} = \begin{pmatrix} - & x_1^T & - \\ & \vdots & \\ - & x_n^T & - \end{pmatrix}_{n \times k} \quad n \times k$$

$x_i^T = (1 \; x_{i1} \dots x_{ik})$: riadky : pozorovaniam

X_j : stĺpce, premenné → vyberáme

$j = 1, \dots, k-1$

$$= \left(1_n, X_1, \dots, X_{k-1}\right) \quad n \times k$$

Kombinatorická selekcia:

- výber $q \leq k$ premenných, ktoré najlepšie popisujú dáta

→ intercept a $(q-1)$ X-ov

prečo nie všetky?

kvalita predikcie: $x_{new} \in \mathbb{R}^q$: nemáme

y_{new} : nemáme hodnota z rozdelenia $y = f(x, \beta) + \varepsilon$

$f(x, \beta)$ neznáme; dúfame, že je to $x^T\beta$

expected prediction error:

$$\Rightarrow EPE = E\left[(y_{new} - x_{new}^T \hat{\beta})^2 \right] = \text{var} + (\text{Bias})^2$$

$$\underbrace{\text{Var}(y_{new} - x_{new}^T \hat{\beta})}_{\text{rozptyl : var}} = \underbrace{E\left[(y_{new} - x_{new}^T \hat{\beta})^2\right]}_{EPE} - \underbrace{\left(E\left[y_{new} - x_{new}^T \hat{\beta}\right]\right)^2}_{\text{Bias : vychýlenie}}$$

- ak mám málo X_j v modeli $\Rightarrow x^T\beta$ dobre popisuje $f(x, \beta)$: nízky Bias

↳ $\Rightarrow$ typicky je ale vtedy vysoká variancia predikcií;

↳ ⇒ typicky je ale vtedy vysoká variancia predikcií: vysoké Var

→ celková chyba predikcie { vychýlenie: $(\text{Bias}(x_{new}^T \hat{\beta}))^2$; rozptýlenosť: $\text{var}(y_{new} - x_{new}^T \hat{\beta})$

– bias-variance trade-off

Pre: umelý príklad: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (M1)

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ (M2)

⋮

$y_i = \beta_0 + \dots + \beta_k x_i^k + \varepsilon_i$ (Mk)

– ako merať, či sú premenné dobré?

$RSS = \|y - X\beta\| = \sum (y_i - x_i^T \beta)^2 \to \min_{\beta}$: "chyba" regresie : súčet štvorcov reziduí

• pridávanie premenných ⇒ zmenšenie RSS : čím viac premenných dať ✗

→ pre modely s rovnakým # prem.: môžem porovnať pomocou RSS – ten s nižším je lepší

↘ pre rôzne # prem. (q) → iné kritériom

↳ s penalizáciou za # premenných

↳ informačné kritériá (information criteria) (IC)

IC → min

• $AIC = \ln\left(\frac{RSS}{n}\right) + \frac{2q}{n}$: Akaike inf. crit. (An inf. crit.)

• $BIC = \ln\left(\frac{RSS}{n}\right) + \frac{q}{n}\ln(n)$: Bayes inf. crit. (Schwarz inf. crit.)

↳ tiež: SIC

• $C_p = \frac{RSS}{\hat{\sigma}^2} - n + 2q$: Mallow's C_p

↑

↳ v orig. článku P = množina indexov $\in \{0, \dots, k-1\}$, kt. sme vybrali do modelu

$\hat{\sigma}^2 = S^2_{Full} = \frac{RSS_{Full}}{n-k}$

↳ z plného modelu

↳ asymptoticky je to ekvivalentné s AIC

všeobecne: $AIC = -2 \log lik + 2q$

všeobecne: $AIC = -2\log lik + 2q$

$BIC = -2\log lik + q\log(n)$

Ako nájsť model s min IC?

i) všetky podmnožiny: všetky komb. prem. z $\{1, \dots, k-1\}$ $\Rightarrow$ 2^{k-1} modelov

↳ M_1^*: najlepší model pre $q=1$ (len intercept)

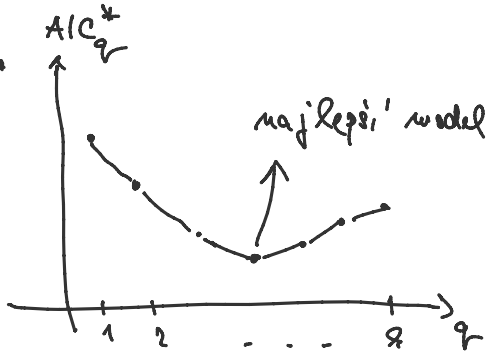
M_2^*: -"- $q=2$ (intercept + jeden člen)

⋮

M_k^*: -"- $q=k$ (plný model)

} Stačí RSS

- na základe IC: vyberiem najlepší z $\{M_1^*, \dots, M_k^*\}$



ii) krokové metódy (stepwise selection)

- 2^{k-1} je príliš veľa pre veľké "k"

→ vyskúšať iba niektoré

- dopredná selekcia:
  - začnem s úbohým modelom ($y_i = \beta_0 + \varepsilon_i$)
  - v každom kroku pridám prem., ktr. spôsobí najväčší pokles IC
  - → opakujem až kým sa nestane: pridanie žiad. prem. spôsobí nárast IC

(~~- najviac k^2 modelov~~) → posledný model = výsledok selekcie

1 ↓ 1,3 ↓ 1,3,7 ↓ (1,3,7,2) ↓ zastavím, lebo neviem zlepšiť IC

- spätná selekcia: začnem s plným (backward)
  - : v každom kroku odoberiem tú, ktorá spôsobí najväčší pokles IC
  - : až kým sa nestane: odobratie žiad. prem. zvýši IC

1,2,3,4,5 ↓ 1,2,3,5 ↓ (2,3,5) ↓ X

- obojstranná selekcia: môžem pridávať aj odoberať

Poznámky:

- krokové metódy - rýchlejšie
  - nemusia nájsť min IC
- klasická inferencia (testy, int. spoľahlivosti) nie je korektné po použití selekcie prem.
  → výberom premenných sa dopúšťam [illegible]

- klasická inferencia (testy, int. spoľahlivosti) nie je korektné po použití selekcie prem.
  → výberom premenných sa deformuje rozdelenie
- ak máme skupinu prem., ktoré nechceme oddeliť (napr. kategorická ⇒ dummy),
  tak ich pridávame / odhadujeme naraz (a tomu prispôsobíme výpočty)
- ∃ triky pre rýchle počítanie selekcie - nemusíme vždy znovu počítať $\hat{\beta}$, RSS, ...

<u>všeobecne</u>: veľa modelov môže byť dosť dobrých
⇒ ten vybraný by sa nemal používať na príliš silnú interpretáciu
(na predikciu je ale vhodný)

<u>Kvalita modelu</u> (pre predikciu)

- aký je dobrý model na predikciu: EPE = ?

→ vieme ju odhadnúť: pomocou <u>testovacej vzorky</u>
↑ EPE

- na začiatku: rozdelíme dáta na 2 časti
  - tréningová : Train : odhadujeme modely, vyberáme najlepší
  - testovacia : Test : posúdime kvalitu vybraného modelu

$y = X\beta + \varepsilon$

$\begin{pmatrix} y_{train} \\ y_{test} \end{pmatrix}$ $\begin{pmatrix} X_{train} \\ X_{test} \end{pmatrix}$

- y_{train}, X_{train} → finálny model → $\hat{\beta}$

testovacia chyba $MSE_{test} = \frac{1}{n_{test}} \sum_{i \in Test} (y_i - x_i^T \hat{\beta})^2 = \frac{1}{n_{test}} \| y_{test} - X_{test} \hat{\beta} \|^2$

mean squared error ← veľkosť test. vzorky

: priemerná chyba predikcia

→ môžem porovnať s inými modelmi, určiť relatívnu chybu predikcie

- napr.: tréningová : 90 % dát
  testovacia : 10 % dát

<u>R&o</u>: Online news - články z mashable.com
- dáta o nich
- modelujem # zdieľaní článku (shares)

a) iba niektoré premenné → výber i vyskúšanie všetkých podmnožín (manuálne)
→ R-ovská funkcia: regsubsets

[illegible] funzcia : regsubsets