ITAP: prva domača naloga

Rok oddaje: ~~18.~~ 25. marec 2020

Rešitve stisnite v ZIP datoteko z imenom `ime-priimek.zip`, in jih oddajte preko spletne učilnice. Poročilo, s katerim opišete postopek reševanja, ni potrebno, saj bo ocenjevanje temeljilo na kvizu.

Priložite programe, s katerimi ste naloge rešili. Programi za vsako nalogo naj bodo v svoji mapi z imenom `naloga<zaporedna številka>`. Naloge naj bodo rešene v R, saj bodo kvizi pripravljeni v tem jeziku.

Za reševanje vsake od nalog uporabite predlogo, ki jo najtede na učilnici poleg datotek s podatki in tega besedila. S tem bomo preprečili nerodnosti pri uporabi psevdonaključnih števil.

Če uporabite kodo, ki ste jo našli, navedite vir. Če imate kakšno vprašanje o nalogah, se obrnite na asistenta ali profesorja, lahko pa objavite vprašanje kar na forumu.

1 Najbližji sosedje

V datoteki `podatki1.csv` se nahajajo podatki, ki jih opisuje 8 stolpcev:

- stolpci `x1`, ..., `x6` podajajo vrednosti vhodnih spremenljivk x_i, $1 \leq i \leq 6$,
- stolpec `y` podaja vrednost **kategorične** ciljne spremenljivke $y \in \{0, 1\}$,
- stolpec `z` podaja napovedi vrednosti y, ki jih interpretiramo kot ocene verjetnosti $P(y = 1|(x_1, \ldots, x_6))$.

Primeri, za katere je $y = 1$, so pozitivni, preostali so negativni.

1.1 Pravilno pozitivni

Preštej število pravilno pozitivnih primerov pri izbrani vrednosti odločitvenega praga $\phi = 0{,}6$.

1.2 Ploščina pod krivuljo

Izračunaj ploščino pod krivuljo ROC, ki jo razpenjanjo vse mogoče vrednosti odločitvenega pragu.

1.3 Pomembna značilka

Ena od značilk x_1, x_2 in x_3 že sama loči oba razreda primerov, tj. obstajajo $i \in \{1, 2, 3\}$, $y_0 \in \{0, 1\}$ in $\vartheta \in \mathbb{R}$, za katere velja $x_i \leq \vartheta \Rightarrow y = y_0$ in $x_i > \vartheta \Rightarrow y = 1 - y_0$. V resnici obstaja neskončno ustreznih odločitvenih pragov in lahko izberemo poljubnega iz (maksimalnega) intervala $I = [\vartheta_0, \vartheta_1)$. Koliko je dolžina intervala I? (Maksimalnost I pomeni, da je vsak interval primernih pragov vsebovan v I.)

1.4 Predelaj podatke

Ker je značilka x_6 kategorična (z vrednostma `a` in `b`) ter je naš vir podatkov ugotovil, da njene vrednosti ne moremo poznati, preden poznamo vrednosti y, naredimo novo ciljno spremenljivko $y' \in \{0, 1\} \times \{\mathtt{a}, \mathtt{b}\}$. Kakšna je velikost najmanjšega od novih razredov?

1.5 Končni model

Naredi model $y' = f(x_1, \ldots, x_5)$, pri čemer za f uporabiš metodo enega najbližjega soseda. Za učenje uporabi prvih 600 primerov v podatkovju. Ostale uporabi za testiranje. Koliko je mikro priklic tega modela na testni množici? Postopek za izračun mikro priklica podaja Algoritem 1.

Algoritem 1 Mikro priklic(možni razredi Y, dejanske vrednosti $\boldsymbol{y}$, napovedane vrednosti $\boldsymbol{z}$)

n = dolžina seznamov $\boldsymbol{y}$ in $\boldsymbol{z}$
$p = 0$ # število pozitivnih primerov
$pp = 0$ # število pravilno pozitivnih primerov
za vse $a \in Y$ **do**
    $p = p + |\{i \mid \boldsymbol{y}_i = a,\ 1 \leq i \leq n\}|$
    $pp = pp + |\{i \mid \boldsymbol{y}_i = a = \boldsymbol{z}_i,\ 1 \leq i \leq n\}|$
konec zanke
vrni pp/p

Vidimo torej, da se mikro priklic za ciljne spremenljivke z več kot dvema mogočima vrednostma izračuna podobno kot priklic za dvojiške ciljne spremenljivke. Opazimo lahko tudi, da na koncu algoritma velja $p = n$, torej lahko v algoritmu nekaj dela privarčujemo.

V premislek: a) Kako bi definirali mikro natančnost? b) Če obstaja mikro-različica nečesa, potem mora gotovo obstajati tudi makro-različica. Kako bi torej lahko še povprečili priklic?

2 Linearna regresija

V datoteki `podatki2.csv` se nahajajo podatki, ki jih opisuje $m + 1 = 21$ numeričnih stolpcev:

- prvih m podaja vrednosti vhodnih spremenljivk x_i, $1 \leq i \leq m$,
- zadnji podaja vrednosti ciljne spremenljivke y.

2.1 Osnovna naloga

Uporabi kar vse podatke in zgradi napovedni model s pomočjo linearne regresije. Izračunaj napovedi. Naj bo e_m RMSE (koren srednjega kvadratnega odklona) modela na učni množici. Koliko je e_m?

2.2 Koristne značilke

Naj bo absolutna vrednost koeficienta za dano značilko ocena za njeno pomembnost pri napovedovanju ciljne spremenljivke. Z $x_{(i)}$ označimo i-to najpomembnejšo (pozor, v splošnem NE velja $x_1 = x_{(1)}$ ipd.). Naj bo X_k množica k najpomembnejših značilk, tj. $X_k = \{x_{(1)}, \ldots, x_{(k)}\}$, $1 \leq k \leq m$. Če imata dve značilki enako pomembnost (kar je zelo malo verjetno), naj ima prednost tista z nižjim indeksom.

Naj bo e_k RMSE modela, ki za učenje uporabi zgolj značilke iz množice X_k, in naj bo k_0 najmanjši indeks, pri katerem se zgodi, da je $e_k \leq 1{,}1 e_m$. Koliko je e_{k_0}?

2.3 Dodatne značilke

Posumimo, da bi se dalo model še izboljšati, če dodamo za vhodne spremenljivke še kvadrate značilk. Z linearno regresijo najdi torej model

$$y = f(x_1, \ldots x_m, x_{m+1}, \ldots, x_{2m}),$$

kjer je $x_{m+i} = x_i^2$. Koliko je pripadajoč RMSE?

2.4 Regularizacija

Začetni problem (z uporabo značilk x_i, $1 \leq i \leq m$) rešimo še z regularizacijo. Tokrat iščemo vektor parametrov $\boldsymbol{\beta}^*$, v katerem je dosežen minimum izraza

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 . \tag{1}$$

Reši zgornji problem s prevedbo na standardnega, tj. najdi taka $\boldsymbol{X}'$ in $\boldsymbol{Y}'$, da bo vrednost $\min_{\boldsymbol{\beta}} \|\boldsymbol{Y}' - \boldsymbol{X}'\boldsymbol{\beta}\|_2^2$ dosežena v isti točki kot vrednost (1).

Koliko je $\|\boldsymbol{\beta}^*\|_2^2$, če je $\lambda = 0{,}01$?

2.5 Kako velik λ?

Naj bo $\boldsymbol{\beta}^*(\lambda)$ rešitev optimizacijskega problema (1) in $\alpha = \|\boldsymbol{\beta}^*(0{,}01)\|_2^2$. Kateri je najmanjši λ_0, za katerega je $\|\boldsymbol{\beta}^*(\lambda_0)\|_2^2 < \alpha/10$? (Zadošča, da se tvoja ocena $\hat{\lambda}_0$ od prave vrednosti ne razlikuje za več kot 10^{-7}.)

3 Variance in še kaj

V datoteki `podatki3.csv` se nahajajo podatki s strukturo kot v 2. nalogi. Tokrat je vhodna spremenljivka ena sama.

3.1 Kdo se najbolj prilega?

Za vsak s med 1 in 10 najdi (glede na RMSE) optimalni polinom p_s stopnje s, s pomočjo katerega iz stolpca x napoveš y. Kolikšen je najmanjši RMSE? Naj bo dosežen pri $s = s_0$.

3.2 Prečno preverjanje prvič

Implementiraj funkcijo `razbij(podatki, k)`, ki sprejme množico podatkov in število k, vrne pa razbitje podatkov na k delov, ki jih podajajo množice indeksov vrstic P_j, $1 \leq j \leq k$. Če so podatki podani v n vrsticah, naj množica P_j vsebuje vse indekse $i = qk + j$, kjer je $q \in \mathbb{N}$ poljuben in je $1 \leq i \leq n$. Pri $n = 8$ in $k = 3$ tako npr. velja $P_1 = \{1, 4, 7\}$, $P_2 = \{2, 5, 8\}$ in $P_3 = \{3, 6\}$.

Če je $k = 4$, kolikšna je povprečna vrednost vhodne spremenljivke v množici primerov z indeksi iz P_1?

3.3 Prečno preverjanje drugič

Naj bo $k = 4$. Za vse s iz prvega podvprašanja ponovi sledeče:

- Zgeneriraj k parov (U_j, T_j) (učne in testne množice), pri čemer so v T_j primeri z indeksi iz množice P_j, v U_j pa PO VRSTI vsi ostali (za $j = 1$ je, če sledimo primeru iz prejšnje podnaloge tako vrstni red primerov v U_j enak $2, 3, 5, 6, 8$)

- Za vsako množico U_j najdi optimalen polinom $p_{s,j}$ ter izračunaj njegovo napako $e_{s,j}$ (RMSE) na T_j. Izračunaj $e_s = \sum_{j=1}^{k} e_{s,j}/k$.

Naj bo najmanjša povprečna napaka dosežena pri stopnji s_1. Koliko je e_{s_1}?

Vrstni red v učni množici sicer ne bi smel biti važen, a bodimo previdni . . .

Varianca p_{i_0}

Po metodi najmanjših kvadratov (ali ekvivalentno, po metodi največjega verjetja) smo našli oceno $\hat{\boldsymbol{a}}$ za vektor parametrov $\boldsymbol{a}$ pravega modela $y = \boldsymbol{x}^T\boldsymbol{a} + \varepsilon$, kjer je $\varepsilon \sim N(0, \sigma^2)$. (V ta okvir spadajo tudi polinomi zgoraj.) Ob predpostavki, da je šum neodvisen od primera do primera, izpelji formulo za varianco napovedi ocenjenega modela $\hat{y}(\boldsymbol{x}_*) = \boldsymbol{x}_*^T\hat{\boldsymbol{a}}$ oz. natančneje, izračunaj

$$\mathrm{Var}[\hat{y}(\boldsymbol{x}_*) \mid \mathcal{D}] \;=\; \mathbb{E}_{\boldsymbol{\varepsilon}}\left[(\hat{y}(\boldsymbol{x}_*) - \mathbb{E}_{\boldsymbol{\varepsilon}}[\hat{y}(\boldsymbol{x}_*) \mid \mathcal{D}])^2 \mid \mathcal{D}\right],$$

kjer smo z $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$ označili opažene podatke, tj. dejansko matriko $\boldsymbol{X}$ vrednosti vhodnih spremenljivk in vektor vrednosti $\boldsymbol{Y}$ ciljne spremenljivke.

Nasvet: izrazi $\hat{\boldsymbol{a}}$ z $\boldsymbol{X}$ in $\boldsymbol{Y}$ ter upoštevaj, da je $\hat{\boldsymbol{a}}$ nepristranska ocena za $\boldsymbol{a}$. Morda ti prav pride tudi zveza $b^2 = bb^T$ za vse skalarje b.

V razmislek: a) Varianca torej sploh ni odvisna od opaženih ciljnih vrednosti. Zanimivo, kajne? b) Smo predpostavko o normalnosti šuma potrebovali?

3.4 Varianca p_{s_0}

Denimo, da je $\sigma = 1.0$. Koliko je varianca napovedi $p_{s_0}(1)$ (polinom iz prve podnaloge)?

3.5 Varianca p_{s_1}

Denimo, da je $\sigma = 1.0$. Koliko je varianca napovedi $p_{s_1}(1)$ (polinom iz tretje podnaloge)?