JMBAG | IME I PREZIME | BROJ BODOVA

STATISTIČKO UČENJE

Završni ispit – 21. veljače 2022.

- Dozvoljeno je koristiti samo pribor za pisanje i brisanje.
- Ispit ukupno nosi 50 bodova, a vrijeme pisanja je 2 sata.
- U svim zadacima koristimo iduće oznake. Y označava varijablu odziva; $Y \in \mathbb{R}$ označava problem regresije, a $Y \in S = \{0, 1, \ldots, K-1\}$ problem klasifikacije s K kategorija. $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ su kovarijate. Skup za učenje je $\tau = \{(x^{(i)}, y_i) : i = 1, \ldots, n\} \subseteq \mathbb{R}^p \times \mathbb{R}$. $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ je matrica u kojoj je i-ti redak jednak $(x^{(i)})^\tau$; pretpostavljamo da su stupci matrice $\boldsymbol{X}$ centrirani, da je $p < n$ te da je $\boldsymbol{X}$ punog ranga. $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$ je vektor kojemu je i-ti element jednak y_i.

Zadatak 1. Neka je $Y \in \mathbb{R}$. Za parametar $\lambda \geq 0$, koeficijenti dobiveni ridge regresijom su

$$\hat{\beta}^r := \arg\min_{\beta \in \mathbb{R}^p} RSS_\lambda(\beta) := \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - \beta^\tau x^{(i)})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

(a) (*3 boda*) Pokažite da je za sve $\lambda \geq 0$,

$$\hat{\beta}^r = (\boldsymbol{X}^\tau \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\tau \boldsymbol{y},$$

pri čemu je $I \in \mathbb{R}^{p \times p}$ identiteta. (*Napomena:* Možete pretpostaviti da je za svaki $\lambda \geq 0$, $\beta \mapsto RSS_\lambda(\beta)$ konveksna funkcija.)

(b) (*2 boda*) U slučaju da su kovarijate ortonormirane, tj. ako vrijedi $\boldsymbol{X}^\tau \boldsymbol{X} = I$, izrazite $\hat{\beta}^r$ u ovisnosti od λ i vektoru $\hat{\beta}^{ls} \in \mathbb{R}^p$ dobivenom metodom najmanjih kvadrata.

(c) (*2 boda*) Neka je $S = \frac{1}{n} \boldsymbol{X}^\tau \boldsymbol{X}$ uzoračka kovarijacijska matrica. Za $j = 1, \ldots, p$, definirajte vektor $\boldsymbol{z}_j \in \mathbb{R}^n$ koji predstavlja j-tu glavnu komponentu od $\boldsymbol{X}$.

(d) (*3 boda*) Odredite koeficijente vektora $\hat{\boldsymbol{y}}^{ls} := \boldsymbol{X}\hat{\beta}^{ls} \in \mathbb{R}^n$ i $\hat{\boldsymbol{y}}^r := \boldsymbol{X}\hat{\beta}^r \in \mathbb{R}^n$ u bazi $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. Za fiksan $\lambda > 0$, koristeći ta dva prikaza interpretirajte na koji način djeluje ridge regresija u odnosu na metodu najmanjih kvadrata? (*Napomena:* Ako vam pomaže, možete koristiti SVD dekompoziciju $\boldsymbol{X} = UDV^\tau$ pri čemu je $U \in \mathbb{R}^{n \times p}, V \in \mathbb{R}^{p \times p}, D \in \mathbb{R}^{p \times p}$, te $U^\tau U = V^\tau V = VV^\tau = I$ i D dijagonalna s nerastućim vrijednostima na dijagonali.)

JMBAG | IME I PREZIME | BROJ BODOVA

STATISTIČKO UČENJE

Završni ispit – 21. veljače 2022.

Zadatak 2. Neka je $Y \in S = \{0, 1\}$. Linearna diskriminacijska analiza (LDA), Kvadratna diskriminacijska analiza (QDA) i Naivni Bayes (NB) su metode koje za svaki $k \in S$, modeliraju uvjetnu razdiobu $f_k(x) = \mathbb{P}(X = x \mid Y = k)$, $x \in \mathbb{R}^p$.

(a) (*2 boda*) Ako znamo (tj. procijenimo) uvjetne razdiobe f_k, $k \in S$, i vjerojatnosti $\pi_k = \mathbb{P}(Y = k)$, kako definiramo klasifikator $\hat{f} : \mathbb{R}^p \to S$? (*Uputa:* Koristite Bayesov teorem.)

(b) (*3 boda*) Navedite pretpostavke na f_k, $k \in S$, za svaku od gore navedenih metoda.

(c) (*3 boda*) Od gore navedenih metoda, navedite one koje

(c1) generiraju linearnu granicu odluke: _____

(c2) prirodno dopuštaju kvalitativne kovarijate: _____

(*Napomena:* Nije potrebno obrazloženje.)

(e) (*2 boda*) Koji problem pri procjeni uvjetnih razdioba imaju LDA i QDA u slučaju kada je p relativno velik u odnosu na n?

JMBAG | IME I PREZIME | BROJ BODOVA

STATISTIČKO UČENJE

Završni ispit – 21. veljače 2022.

Zadatak 3. U ovom zadatku bavimo se unakrsnom validacijom. Pretpostavimo da je $Y \in \mathbb{R}$. Testnu grešku procjenitelja $g : \mathbb{R}^p \to \mathbb{R}$ definiramo kao $L(g) := \mathbb{E}[L(Y, g(X))]$ pri čemu $L : \mathbb{R}^2 \to [0, \infty)$ označava funkciju gubitka.

(a) (*3 boda*) Neka je $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ procjenitelj dobiven na temelju skupa za učenje τ koristeći neku metodu. Za fiksan $k \in \{2, \ldots, n\}$ (prepotstavite da je $n/k =: r \in \mathbb{N}$), precizno definirajte kako procjenjujemo $L(\hat{f})$ koristeći unakrsnu validaciju s k blokova (tzv. *k-fold cross validation*); u nastavku ćemo tu procjenu označiti s $CV^{(k)} = CV^{(k)}(\hat{f})$.

(b) (*2 boda*) Poznato je da $CV^{(k)}$ tipično dobro procjenjuje samo $\mathbb{E}_T[L(\hat{f}(T))]$ (očekivanje je s obzirom na slučajnost koja dolazi od skupa za učenje T). Objasnite intuitivno zašto je tipično $\mathbb{E}_T(CV^{(k)}) \geq \mathbb{E}_T[L(\hat{f}(T))]$. Što se tipično događa s razlikom $\mathbb{E}_T(CV^{(k)}) - \mathbb{E}_T[L(\hat{f}(T))]$ kako k raste prema n?

(c) (*3 boda*) Pretpostavimo da je $p = 1$ i $n = 5$, te $\tau = \{(1,1), (2,2), (4,2), (5.5,4), (6,1)\}$. Izračunajte $CV^{(n)}(\hat{f})$ ako je za sve $x \in \mathbb{R}$, $\hat{f}(x) := y_m$, pri čemu je $m \in \{1, \ldots, n\}$ takav da je $|x - x^{(m)}| = \min_{i \in \{1,\ldots,n\}} |x - x^{(i)}|$.

(d) (*2 boda*) Pretpostavimo da za modeliranje odziva koristimo linearnu regresiju (koeficijente procjenjujemo npr. metodom najmanjih kvadrata). Ipak, broj kovarijata p je jako velik pa tražimo manji podskup kovarijata koje imaju najveći utjecaj na odziv. U *best subset selection* metodi za svaki $k \in \{1, \ldots, p\}$ pronađemo model $\mathcal{M}_k$ s onih k kovarijata koje minimiziraju grešku na skupu za učenje τ između svih mogućih modela s točno k kovarijata. Objasnite kako koristeću unakrsnu validaciju odabiremo "optimalan" model između modela $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

JMBAG | IME I PREZIME | BROJ BODOVA

STATISTIČKO UČENJE

Završni ispit – 21. veljače 2022.

Zadatak 4. Pretpostavite da je $Y \in \mathbb{R}$ i $X \in D \subseteq \mathbb{R}^p$.

(a) (*2 boda*) Ako je T binarno stablo koje odgovara jednoj rekurzivnoj binarnoj particiji skupa D, listovi od T odgovaraju elementima te particije. Definirajte procjenitelj $\hat{f}_T : D \to \mathbb{R}$ dobiven iz T na temelju skupa za učenje τ.

(b) (*2 boda*) Definirajte bootstrap uzorak $T_{boot} = \{Z_1^*, \ldots, Z_n^*\}$ za τ.

(c) (*2 boda*) Definirajte *bagging* procjenitelj $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$, $x \in D$ za $B \in \mathbb{N}$. Je li glavna ideja smanjiti pristranost ili varijancu procjenitelja baziranog samo na jednom stablu?

(d) (*2 boda*) Kako se modificira *bagging* procjenitelj tako da se dobije tzv. *random forest* procjenitelj $\hat{f}_{rf}$? Koja je glavna ideja?

(e) (*2 boda*) Definirajte tzv. *Out-of-Bag* procjenitelj testne greške od $\hat{f}_{bag}$ (i $\hat{f}_{rf}$).

JMBAG IME I PREZIME BROJ BODOVA

STATISTIČKO UČENJE

Završni ispit – 21. veljače 2022.

Zadatak 5. (*10 bodova*) Svaki ispravno zaokružen odgovor nosi 1 bod, svaki krivo zaokružen nosi -1 bod, a svaki nezaokružen odgovor nosi 0 bodova. Ako je ukupna suma bodova na ovom zadatku negativna, dobit ćete 0 bodova. Vaše odabire nije potrebno obrazlagati.

(a) Koji od sljedećih pristupa tipično pomažu pri smanjivanju varijance rezultirajućeg procjenitelja?

(a1) Povećanje parametra λ u ridge regresiji.

Točno Netočno

(a2) Povećanje parametra k u *k-nearest neigbhors* regresiji.

Točno Netočno

(a3) Veći broj listova u regresijskom stablu.

Točno Netočno

(a4) Dodavanje novih kovarijata u model (npr. kod linearne regresije).

Točno Netočno

(a5) Veći broj elemenata u skupu za učenje (npr. kod linearne regresije).

Točno Netočno

(b) U ridge regresiji, broj stupnjeva slobode je barem onoliki koliki je broj kovarijata.

Točno Netočno

(c) U usporedbi s lasso regresijom, ridge regresija rezultira s većim brojem koeficijanata koji su jednaki 0.

Točno Netočno

(d) Kod *bagging* metode, prevelika vrijednost parametra B uzrokuje *overfitting*.

Točno Netočno

(d) Kod *gradient boosting* metode, tipično koristimo mala stabla (npr. manje od 10 listova) kao bazne procjenitelje.

Točno Netočno

(e) U svakom koraku *gradient boosting* algoritmu, prilagođavamo regresijsko stablo tzv. pseduo rezidualima trenutnog modela. Provjerite točnost sljedeće tvrdnje: ako u slučaju regresije koristimo L_1 gubitak, pseudo reziduali poprimaju vrijednosti 0 ili 1.

Točno Netočno