Statističko učenje 23./24.
Druga domaća zadaća

Rok za predaju: **6. prosinca, 2023.**

Broj bodova: 10

Teorijski dio

Napomena: *Dovoljno je riješiti jedan teorijski zadatak po izboru. U tom slučaju preostali zadatak služi kao vježba za završni ispit.*

Zadatak 1 (Ridge regresija)

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 $(x^{(i)})^\tau$, a $\boldsymbol{y} \in \mathbb{R}^{n \times 1}$ je vektor kojemu je i-ti element jednak y_i. Za parametar $\lambda \geq 0$, koeficijenti dobiveni ridge regresijom su

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

Pokažite da je za sve $\lambda \geq 0$,

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

pri čemu je $I \in \mathbb{R}^{p \times p}$ identiteta. Objasnite zašto je za $\lambda > 0$ gornje rješenje uvijek dobro definirano (dakle, postoji i jedinstveno je).

Zadatak 2 (Težinski smoothing splajn) Neka je $p = 1$, a za sve $\lambda > 0$ i sve $f \in S := \{g : [a, b] \to \mathbb{R} : g \text{ dva puta neprekidno dfb.}\}$,

$$\mathrm{RSS}(f, \lambda) := \sum_{i=1}^{n} w_i (y_i - f(x^{(i)}))^2 + \lambda \int_a^b (f''(t))^2 \mathrm{d}t\,, \tag{1}$$

gdje su $w_1, \ldots, w_n \geq 0$ proizvoljne (ali fiksne) težine.

(a) Ako pretpostavimo je $x^{(i)} \neq x^{(j)}$, za sve $i \neq j$, pokažite da je $f \in S$ koja minimizira $\mathrm{RSS}(f, \lambda)$ za fiksni $\lambda > 0$, nužno prirodni kubični splajn s čvorovima u $x^{(1)}, \ldots, x^{(n)}$, te karakterizirajte to rješenje u ovisnosti o τ, λ, težinama w_i, te odabranoj bazi za prostor prirodnih kubičnih splajnova.

(b) Pretpostavimo sada da su težine $w_1 = \cdots = w_n = 1$, ali da među $x^{(i)}$-evima moguće ima jednakih vrijednosti. Objasnite kako se sada problem minimizacije $\mathrm{RSS}(f, \lambda)$ (čiji rezultat je točno smoothing splajn) može svesti na slučaj u (a) za prikladno odabrane podatke i težine. (*Uputa:* Grupirajte podatke s istom vrijednosti x-a, a odzive zamijenite s njihovim prosjekom.)

Praktični dio

Zadatak 1 (Ridge, lasso i elastic net regresija)

U `bstar.Rdata` nalazi se vektor `bstar` duljine $p = 2500$ koji predstavlja jednostavnu sliku dimenzija 50×50 – možete ju nacrtati koristeći funkciju `plot.image`. Vaš cilj je na temelju n slučajnih linearnih kombinacije piksela, pri čemu je n dosta manji od p, rekonstruirati originalnu sliku; ovdje je ključno što je slika `bstar` *rijetka* (engl. *sparse*).

Preciznije, zadan je vektor `y` koji sadrži $n = 1300$ skalarnih produkata vektora `bstar` sa vektorom $x^{(i)}$ (duljine p) sačinjenim od njd $\mathrm{N}(0,1)$ slučajnih varijabli, s tim da je svakoj linearnoj kombinaciji dodana slučajna greška ϵ_i koja ima $\mathrm{N}(0, 5^2)$ razdiobu – vidi `zad1.R`. Cilj zadatka je dobiti što bolju procjenu za vektor koeficijenata `bstar` koristeći ridge, lasso i elatic net regresiju, s tim da **nećete** uključivati slobodni član (engl. *intercept*); za elastic net metodu koristite parametar $\alpha = 1/2$.

(a) Provedite unakrsnu validaciju s 10 blokova za sve tri metode te nacrtajte CV procjenu testne grešku (koristite funkciju `cv.glmnet`). Čine li se vrijednosti odabrane za λ u redu; npr. bi li trebalo uključiti neke manje ili veće vrijednosti? Ako da, učinite to. Za sve tri metode, odredite λ koji minimizira CV grešku te onaj dobiven tzv. pravilom jedne standardne greške. Za koju je metodu minimalna CV greška najmanja?

(b) Koeficijenti dobiveni u (a) dijelu predstavljaju procjene vektora `bstar`. Nacrtajte slike koje odgovaraju tim procjenema za parametre iz (a) dijela (dakle, po dvije slike za svaku metodu). Komentirajte rezultate – koja se metoda čini najbolja?

(c) Koristeći funkciju `truncate` modificirajte procjene iz (b) dijela tako da sadrže samo vrijednosti iz $[0, 1]$ te izračunajte njihove udaljenosti od vektora `bstar` u Euklidskoj normi. Koja metoda daje najmanju grešku?

(d) Ponovite korake (a)-(c) s tim da ćete vektor `y` generirati kao gore, ali će sada kovarijate $X_1, \ldots, X_p$ biti zavisne. Preciznije, neka $x^{(i)}$ bude realizacija slučajnog vektora $X = (X_1, \ldots, X_p)$ koji ima p-dimenzionalnu normalnu razdiobu uz $\mathbb{E}[X] = (0, \ldots, 0)$, a za kovarijacisku matricu gledajte dva slučaja

   (i) $\mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_p) = 1$ i $\rho(X_i, X_j) = 0.8^{|i-j|}$, za sve $i \neq j$.

   (ii) $\mathrm{Var}(X_1) = \cdots = \mathrm{Var}(X_p) = 1$ i $\rho(X_i, X_j) = 0.6$, za sve $i \neq j$.

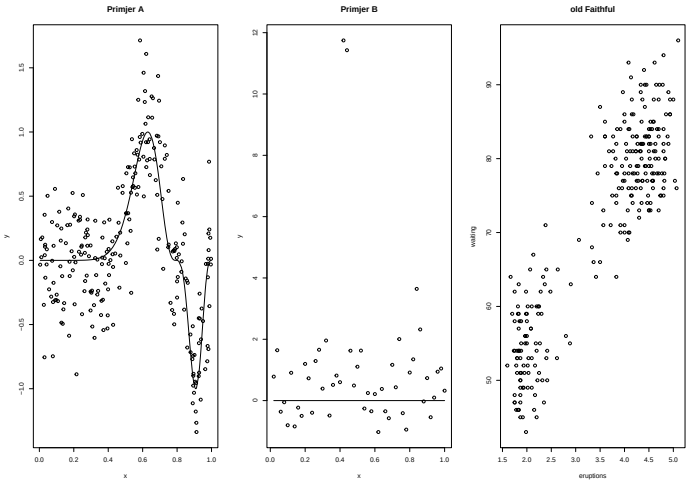
   Normalan slučajni vektor možete generirati pomoću funkcije `mvrnorm` iz paketa `MASS`; ukoliko je to jako sporo, možete potražiti na webu neke efikasnije načine za simuliranje koristeći dekompoziciju Choleskog.

(e) Provedite sljedeću simulacijsku studiju. Za sva tri slučaja (nezavisne kovarijate, te dva slučaja za zavisne kovarijate), $M = 30$ (po mogućnosti i više) puta ponovite

gornje korake, tj. M puta generirajte vektor `y` koristeći n linearnih kombinacija (s novim simulacijama vektora $x^{(i)}$, $i = 1, \ldots, n$, i greške ϵ_i) te izračunajte Euklidske udaljenosti između vektora `bstar` i dobivenih 6 procjene kao u (c) dijelu. Nacrtajte *boxplotove* dobivenih M vrijednosti za sve slučajeve (dakle, 3 puta po 6 *boxplotova*). Komentirajte rezultate.

Zadatak 2 (Nelinearna regresija za $p = 1$)

Promatramo tri skupa podataka, prva dva su simulirani s poznatom regresijskom funkcijom, a treći su stvarni podaci `faithful` (vidi `?faithful`):



Kod koji generira gornju sliku nalazi se u `zad2_web.R`.

Za svaki skup podataka prilagodite *smoothing* splajn te lokalnu kvadratnu regresiju za parametre koji minimiziraju GCV grešku. U slučaju lokalne regresije, sami napišite funkciju koja za niz parametara `span` vraća onaj koji minimizira GCV grešku; možete koristiti predložak dan u `zad2_web.R`.

U sva tri slučaja komentirajte rezultate. Čini li prilagodba dobra? Ukoliko ne, isprobajte i druge parametre za gornje modele. U slučaju podataka `faithful`, usporedite dobivene procjene s procjenom dobivenom linearnom regresiju (metodom najmanjih kvadrata) – kakva je razlika u predikciji ako promatramo dva "klastera" podatke, one za koje je kovarijata `eruption` između 1.5 i 2.5, odnosno 3.5 i 5?

Poruka zadatka: GCV je korisna metoda za odabir optimalnog parametra kompleksnosti, ali ne radi uvijek najbolje pa je tipično koristimo kao početni *prijedlog* pri odabiru.