Statističko učenje 21./22.
Drugi praktični zadatak (seminar)

Rok za predaju: **1. ožujka, 2022.**

Cilj praktičnog zadatka je izrada (1) `R` *koda koji rješava donje zadatke, te (2) kratke prezentacije ili dokumenta u kojem su prezentirani glavni rezultati (slike, tablice itd.) te zaključci i komentari. Maksimalan broj bodova koji se može dobiti je 25. Diskusija među studentima je naravno dopuštena (i preporučena), ali očekuje se da svaki student potpuno samostalno napiše svoje rješenje. Pripazite da se vaši rezultati mogu reproducirati (koristite funkciju* `set.seed`*). Ukoliko vam se učini zanimljivo, možete (dapače, poželjno je) napraviti i više nego što se od vas traži u zadacima. U slučaju bilo kakvih nejasnoća, pitajte!*

Zadatak 1

U ovom zadatku primijenit ćete generalizirani aditivan model (GAM) za regresiju na podatke `chicago` iz paketa `gamair`. Podaci sadrže informacije o dnevnom broju umrlih osoba u gradu Chicagu (`death`) (odziv), a potencijalne kovarijate su: razine ozona (`o3median`), sumporovog dioksida (`so2median`) i određenih štetnih čestica (`pm10median`) u zraku, te prosječna dnevna temperatura (`tmpd`) i dan (`time`); dan 0 je 31. prosinca 1993. Tipično se modelira logaritam broja smrti, pa ćemo tako i mi. U GAM-ovima u nastavku zadatka vezu između odziva i svake od kovarijate modelirajte koristeći *smoothing spline*; ako broj stupnjeva slobode nije eksplicitno zadan, koristite zadanu (*default*) vrijednost (to je 4). Temperatura je zadana u stupnjevima Fahrenheita – pretvorite ih u stupnjeve Celzijeve.

(a) Prilagodite smoothing spline za odziv `log(death)` u ovisnosti o `time`; stupnjeve slobode odredite koristeći unakrsnu validaciju. Nacrtajte dobiveni smoothing spline zajedno sa stvarnim vrijednostima za dnevni broj smrti (dakle ne logaritmirane vrijednosti). Trebali biste primijetiti 4 uzastopna dana s neobično visokim brojem smrti. Kojim datumima odgovaraju te vrijednosti (koristite funkciju `as.Date`)?

(b) Prilagodite GAM za odziv `log(death)` u ovisnosti o svim dostupnim kovarijatama; za kovarijatu `time`, ovdje i u nastavku, koristite isti broj stupnjeva slobode kao u (a) dijelu.

  (b1) Nacrtajte procijenjene funkcije za svaku kovarijata (tzv. parcijalne funkcije). Interpretirajte grafove (npr. kako svaka od kovariajta utječe na odziv, jesu li veze približno linearne, koja se čini da ima veći/manji utjecaj na odziv, što znači kada je funkcija pozitivna/negativna itd.).

  (b2) Nacrtajte vrijednosti dnevnih smrti dobivene iz modela u ovisnosti o kovarijati `time` (to je analogno grafu iz (a) dijela zadatka). Kako su se procjene promijenile u odnosu na model iz (a) dijela? Jesu li 4 "outliera" još uvijek tu, predviđa li ih ovaj model bolje?

(c) Sada ćete umjesto kovarijate o dnevnoj temperaturi te kovarijata koje opisuju količinu plinova/čestica u zraku u jednom danu, koristiti prosjek vrijednosti trenutnog i 3 prethodna dana. Zašto to ima smisla? Ponovite korake u (b) s transformiranim kovarijatama i komentirajte rezultate. Što se promijenilo? Jesu li outlieri još uvijek tu?

(d) Sada dodajemo interakcije u model, tj. utjecaj (transformiranih) kovarijata `tmpd` i `o3median` modelirat ćemo zajedno umjesto zasebno. Ponovite korake te odgovorite na pitanja kao u (b) i (c) dijelu zadatka; za dvodimenzionalan smoothing spline koji modelira utjecaj kovarijata `tmpd` i `o3median` isprobajte nekoliko različitih vrijednosti (npr. iz intervala $[4, 60]$) za broj stupnjeva slobode – za koju ste se vrijednost odlučili? Je li dodavanje ove interakcije značajno utjecalo na predviđanje naših outliera?

Zadatak 2

Cilj zadatka je simulacijama ilustrirati kako promjena parametara u *gradient boosting* algoritmu utječe na rezultirajući procjenitelj. Neka su kovarijate $X_1, ..., X_p$ njd slučajne varijable s standardnom normalnom razdiobom, te neka je

$$Y' := \begin{cases} 1\,, & \text{ako je } \sum_{j=1}^{p} X_j^2 > m_p\,, \\ 0\,, & \text{inače}\,, \end{cases}$$

pri čemu je m_p =`qchisq(0.5, p)` medijan χ^2-razdiobe s p stupnjeva slobode; dakle, vrijedi $\mathbb{P}(Y' = 1) = 1/2$. Sada stavimo da je odziv $Y \in \{0, 1\}$ s vjerojatnosti 0.8 jednak Y', a s vjerojatnosti 0.2 jednak $1 - Y'$ (tj. zamijenimo klasu). Provedite iduće korake za slučajeve $p = 2$ i $p = 10$:

1. Simulirajte uzorak (skup za učenje) duljine $n = 500$ iz razdiobe od (X, Y).[1]

2. Primijenite gradient boosting algoritam na dobiveni uzorak uz log gubitak koristeći $M = 5000$ stabala dubine (`interaction.depth`) $1, 6$, i 20 i vrijednost `shrinkage` parametra $1, 0.1$ i 0.01 (dakle, ukupno 9 modela za svaki p). Koristite `bag.fraction` =1.[2]

3. Na testnom skupu duljine $m = 2000$ procijenite testnu grešku uz 0-1 gubitak za svaki od 9 dobivenih modela iz prethodnog dijela, a u ovisnosti o broju stabala korištenih u procjeni (dakle, od 1 do M). Prikažite dobivene rezultate te naznačite vrijednost Bayesove greške, tj. minimalne moguće testne greške koju neki procjenitelj može postići.

Komentirajte i usporedite rezultate. Mogu li se intuitivno objasniti?

[1] U slučaju $p = 2$, podaci se mogu i grafički prikazati.

[2] Zadana vrijednost je 0.5, a u tom slučaju se svako stablo generira koristeći samo pola slučajno odabranih podataka iz skupa za učenje – to je tzv. *stochastic* gradient boosting algoritam.