

Uvod u umjetnu inteligenciju

UNIZG FER, ak. god. 2021./2022.

Vježbe, v2

10 Strojno učenje

- 1** (T) Razvijamo model strojnog učenja za predviđanje broja gledatelja na kinoprojekcijama. U obzir smo uzeli tri značajke: dan u tjednu, žanr filma i cijena produkcije filma. **Koji bi algoritam strojnog učenja bilo prikladno upotrijebiti za ovaj problem, i zašto?**
- ☐ A Naivan Bayesov klasifikator, jer predviđamo diskretne vrijednosti (cijeli brojevi)
- ☐ B Neuronsku mrežu, jer predviđamo brojčanu vrijednost
- ☐ C Stablo odluke, jer imamo tri značajke s podjednakom informacijskom dobiti
- ☐ D Neuronsku mrežu, jer je cijena produkcije filma brojčana značajka
- 2** (T) Naivan Bayesov klasifikator nazivamo “naivnim” jer model pretpostavlja uvjetnu nezavisnost značajki x_j unutar klase y . Uz tu pretpostavku, izglednost klase $P(x_1, \dots, x_n|y)$ možemo zamijeniti umnoškom $\prod_{j=1}^n P(x_j|y)$. **Koja je motivacija za uvođenje pretpostavke uvjetne nezavisnosti?**
- ☐ A Veća točnost modela na skupu za učenje
- ☐ B Mogućnost korištenja značajki koje nisu binarne
- ☐ C Mogućnost generalizacije na neviđene primjerke
- ☐ D Sprječavanje podljeva pri množenju vjerojatnosti
- 3** (R) Mali je Ivica svakog svakog ljeta u zadnjih sedam godina naučio jedan novi programski jezik. Svoja vrijedna iskustva sažeo je u listu “*Programski jezik koji mi se sviđa*”, gdje je svaki jezik opisao četirima značajkama, te je naznačio je li mu se dotični jezik svidio ($y = 1$) ili nije ($y = 0$). Ta lista izgleda ovako:

i	Evalucija	Izvođenje	Paradigma	Provjera tipova	y
1	lijena	kompajler	imperativna	statička	0
2	striktna	interpreter	deklarativna	dinamička	0
3	lijena	kompajler	imperativna	dinamička	0
4	lijena	interpreter	hibridna	statička	0
5	striktna	interpreter	imperativna	statička	1
6	lijena	kompajler	hibridna	dinamička	1
7	striktna	kompajler	hibridna	dinamička	1

Ovog ljeta Mali Ivica želi puno jesti i spavati te opet naučiti novi programski jezik. U užem je izboru jezik \mathbf{x} sa sljedećim karakteristikama: $\mathbf{x} = (\text{striktna}, \text{interpreter}, \text{hibridna}, \text{dinamička})$. Međutim, ovog puta Mali bi Ivica volio unaprijed znati hoće li mu se dotični programski jezik svidjeti, tako da ne gubi cijelo ljeto bezveze. Pomozite Malom Ivici te na gornji skup primjera primjenite naivan Bayesov klasifikator s Laplaceovim zaglađivanjem “dodaj jedan”. **Koliko iznosi vjerojatnost da bi se Malom Ivici programski jezik \mathbf{x} svidio?**

- ☐ A 0.856 ☐ B 0.431 ☐ C 0.799 ☐ D 0.694

- 4 (P) Gradimo naivan Bayesov klasifikator za klasifikaciju poruka Twittera (tvitova) prema sentimentu. Svaki tvit želimo klasificirati u jednu od tri klase: pozitivan, negativan ili neutralan. Svaki tvit sastoji se od najviše 280 riječi i prikazujemo ga kao jedan binaran vektor značajki. Vektor se sastoji od 5000 značajki, i svaka značajka odgovara jednoj od 5000 riječi iz našeg vokabulara. Ako je značajka postavljena na 1, to znači da se dotična riječ pojavila u tvitu, inače je značajka postavljena na 0. Npr., ako je $x_{42} = 1$, onda to znači da se 42. riječ iz našeg vokabulara pojavila u tvitu. Učenje ovog modela svodi se na procjenu apriornih vjerojatnosti i izglednosti klase na temelju označenog skupa podataka. **Koliko ćemo ukupno vjerojatnosti trebati procijeniti za ovaj model?**

☐ A 1683 ☐ B 30003 ☐ C 569 ☐ D 31683

- 5 (T) Stabla odluke i Bayesov klasifikator dva su algoritma nadziranog učenja. Međutim, ti se algoritmi vrlo razlikuju. **Što je prednost, a što nedostatak stabla odluke u odnosu na Bayesov klasifikator?**

- ☐ A Prednost je što možemo bolje objasniti zašto je primjer klasificiran u neku klasu, a nedostatak što nemamo vjerojatnost klasifikacijske odluke
- ☐ B Prednost je što primjere možemo klasificirati u više od jedne klase, a nedostatak što se stablo odluke može lako prenaučiti
- ☐ C Prednost je što stablo odluke možemo podrezati kako bismo spriječili prenaučenosť, a nedostatak što može doći do podljeva pri računanju vjerojatnosti
- ☐ D Prednost je što su stabla odluke otporna na male promjene u ulaznom skupu podataka, a nedostatak što pretpostavljamo uvjetnu nezavisnost značajki

- 6 (R) Raspoložemo skupom primjera za *“Nezaboravno jadransko ljeto 2025., odmah nakon pandemije koronavirusa”*. Skup se sastoji od sljedećih primjera, svaki sa 4 značajke (Mjesto, Otok, Smještaj, Prijevoz) i ciljnom oznakom y :

i	Mjesto	Otok	Smještaj	Prijevoz	y
1	Istra	ne	privatni	auto	da
2	Istra	ne	privatni	avion	da
3	Dalmacija	da	hotel	auto	da
4	Dalmacija	da	hotel	bus	da
5	Kvarner	ne	kamp	bus	ne
6	Dalmacija	da	privatni	avion	ne
7	Istra	ne	kamp	auto	ne

Primijenite na ovaj skup primjera algoritam ID3. U slučaju da je u nekom koraku više značajki ima jednaku vrijednost informacijske dobiti, izaberite onu koja je u tablici navedena prva (ona ljevija). **Kako izgleda dobiveno stablo odluke?**

- ☐ A Korijski čvor stabla je Mjesto, a njegovo dijete je čvor Smještaj s informacijskom dobiti 0.918
- ☐ B Korijski čvor stabla je Smještaj, a njegovo dijete je čvor Mjesto s informacijskom dobiti 0.918
- ☐ C Korijski čvor stabla je Mjesto, a njegovo dijete je čvor Smještaj s informacijskom dobiti 0.251
- ☐ D Korijski čvor stabla je Smještaj, a njegovo dijete je čvor Mjesto s informacijskom dobiti 0.251
- 7 (P) Treniramo model stabla odluke na skupu podataka u kojemu, nažalost, ima i nešto šuma. Svjesni smo da prisustvo šuma može dovesti do prenaučenosti modela, pa smo odlučili primijeniti unakrsnu provjeru da bismo podrezali stablo odluke. Skup označenih primjera podijelili smo na skup za učenje D_u i skup za provjeru D_p , tako da $D_u \cap D_p = \emptyset$. Sada isprobavamo nekoliko različitih dubina stabla, od $d = 1$ do $d = 42$. Ispitivanjem pogreške tih stabala različite dubine,

zaključili smo da je optimalna dubina stabla $d = 17$. Što to konkretno znači?

- ☐ A Da je pogreška na skupu D_p za stablo sa $d = 17$ veća od pogreške za isto to stablo na skupu D_u , ali veća od pogreške za bilo koju drugu dubinu stabla d na skupu D_p
- ☐ B Da je pogreška na skupu D_p za stablo sa $d = 17$ manja od pogreške na D_p za $d < 17$ i $d > 17$, ali očekivano ta je pogreška na D_p veća od pogreške na skupu D_u
- ☐ C Da stablo za $d = 17$ ostvaruje najmanju pogrešku na skupu D_u , dok na skupu D_p stablo ostvaruje uvijek veću pogrešku, s maksimumom pogreške za $d = 1$
- ☐ D Da su pogreške za stabla sa $d = 16$ i $d = 18$ veće i na skupu D_u i na skupu D_p , s time da su na skupu D_p očekivano veće nego na skupu D_u