Strojno učenje – završni ispit

UNIZG FER, ak. god. 2013./2014.

30. siječnja 2014.

Ispit traje 180 minuta i nosi 35 bodova. Svaki zadatak rješavajte na zasebnoj stranici.

1. (*5 bodova*) Linearni diskriminativni modeli.

   (a) Definirajte poopćeni linearni model s nelinearnom granicom u ulaznome prostoru.

   (b) Skicirajte na jednome grafikonu (naznačite osi!) sljedeće funkcije gubitka: (1) gubitak 0-1, (2) kvadratni gubitak, (3) gubitak perceptrona, (4) logistički gubitak i (5) gubitak zglobnice.

   (c) Temeljem ove skice, ukratko objasnite koja je prednost odnosno koji je nedostatak logističke regresije u odnosu na linearnu regresiju i SVM.

   (d) Skup primjera je $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\} = \{((0,0),0),((1,1),1),((5,5),1)\}$. Skicirajte u $\mathbb{R}^2$ granicu između klasa koju bi pronašla tri modela: regresija, logistička regresija i SVM.

2. (*6 bodova*) Logistička regresija.

   (a) Izvedite model logističke regresije krenuvši od generativne definicije za $P(\mathcal{C}_j|\mathbf{x})$. Ne morate izvoditi jednadžbe za parametre $\tilde{\mathbf{w}}$, ali trebate navesti pretpostavke generativnog modela.

   (b) Rješavamo problem višeklasne klasifikacije primjera s $n = 50$ značajki u $K = 5$ klasa. Odredite broj parametara modela logističke regresije (shema jedan-naspram-ostali) te broj parametara odgovarajućeg generativnog modela. Koji će model bolje generalizirati?

   (c) Izvedite pogrešku unakrsne entropije krenuvši od maksimizacije log-izglednosti skupa za učenje $\mathcal{D}$.

   (d) Skicirajte na jednome grafikonu funkciju pogreške $E(\tilde{\mathbf{w}}|\mathcal{D})$ za $\lambda = 0$ i $\lambda = 100$ u ovisnosti o broju iteracija za dvije različite vrijednosti stope učenja, η_1 i η_2, gdje $\eta_1 < \eta_2$ (ukupno četiri krivulje). Pretpostavite da algoritam u svim ovim slučajevima konvergira.

3. (*6 bodova*) Stroj potpornih vektora.

   (a) Formulirajte problem maksimalne margine i pripadnu Lagrangeovu funkciju $L(\mathbf{w}, w_0, \boldsymbol{\alpha})$ (primarni problem, tvrda margina).

   (b) Veza između primarnih i dualnih parametara modela je

   $$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)} \boldsymbol{\phi}(\mathbf{x}^{(i)})$$

   Treniranje linearnog SVM-a rezultiralo je potpornim vektorima $\mathbf{x}^{(1)} = (-5, 1, 10, 0)$, $\mathbf{x}^{(2)} = (1, 2, -2, 5)$ i $\mathbf{x}^{(3)} = (0, -5, -1, 7)$. Prvi primjer je negativan, a druga dva su pozitivna. Lagrangeovi koeficijenti su $\alpha_1 = 0.01$, $\alpha_2 = 0.007$ i $\alpha_3 = 0.003$. Pomak je $w_0 = 0.45$. Napišite izraz za $h(\mathbf{x})$ u dualnoj formulaciji te odredite klasifikaciju novog primjera $\mathbf{x}^{(4)} = (5, 5, -50, 10)$. Nalazi li se taj primjer unutar margine?

   (c) Ukratko objasnite što je to *jezgreni trik* i koja je njegova prednost u odnosu na izravno preslikavanje funkcijom $\boldsymbol{\phi}$.

   (d) Koristimo polinomijalnu jezgrenu funkciju $\kappa(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z} + 1)^2$. Odredite vektor $\boldsymbol{\phi}(\mathbf{x})$ u koji će efektivno biti preslikan primjer $\mathbf{x} = (2, 1)$.

4. (*4 boda*) Neparametarski klasifikacijski modeli.

   (a) Skicirajte pogrešku učenja i pogrešku generalizacije modela k-nn kao funkciju hiperparametra k. Naznačite područja prenaučenosti i podnaučenosti.

   (b) Uporabom algoritma ID3 izgradite stablo odluke za klasifikaciju primjera prema kriteriju *"Nezaboravno ljetovanje na Jadranu"*. U prvom koraku napišite postupak izračuna; u daljnjim koracima izračun ne treba pisati. Skup primjera za učenje je sljedeći:

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

5. (*4 boda*) Vrednovanje klasifikatora.

   (a) Izračunajte mikro-F_1 i makro-F_1 temeljem sljedeće matrice zabune (retci odgovaraju predviđenoj kategoriji, a stupci stvarnoj kategoriji):

$$\begin{pmatrix} 4 & 2 & 5 \\ 3 & 20 & 2 \\ 6 & 10 & 31 \end{pmatrix}.$$

   (b) Raspolažemo sa 1000 označenih primjera. Treniramo SVM s Gaussovom jezgrenom, koja ima hiperparametre C i γ. Optimizaciju hiperparametara provodimo iscrpnim pretraživanjem po rešetci (10 vrijednosti za C i 10 vrijednosti za γ). Za procjenu pogreške modela koristimo ugniježđenu unakrsnu provjeru 5×5. Koliko ćemo ukupno puta provesti učenje modela i koji će biti minimalan broj primjera s kojim ćemo učiti model?

6. (*5 bodova*) Grupiranje.

   (a) Skicirajte krivulju pogreške J algoritma k-srednjih vrijednosti kao funkcije broja iteracija, i to za (1) broj grupa $K = N/10$ i $K = N/2$ te (2) za k-means++ i za slučajno inicijalizirana središta (ukupno četiri krivulje). Kako biste u obzir uzeli stohastičnost, nacrtajte krivulje koje biste dobili uprosječivanjem rezultata većeg broja pokretanja algoritma.

   (b) Raspolažemo skupom neoznačenih primjera:

$$\mathcal{D} = \left\{a = (0,1), b = (3,2), c = (1,4), d = (2,3), e = (5,2), f = (7,5), g = (4,9)\right\}.$$

   Primijenite hijerarhijsko aglomerativno grupiranje (HAC) s potpunim povezivanjem te skicirajte pripadni dendrogram.

   (c) Usporedite vremensku i prostornu složenost algoritma k-srednjih vrijednosti i algoritma HAC.

7. (*5 bodova*) Algoritam maksimizacije očekivanja.

   (a) Napišite izraz za mješavinu s latentnim varijablama i izvedite izraz za (potpunu) log-izglednost.

   (b) Napišite pseudokod algoritma maksimizacije očekivanja primijenjenog na model Gaussove mješavine.

   (c) Raspolažemo podatcima $\mathcal{D}$ koji potječu iz $K = 5$ klasa. Manji dio podataka imamo označen. Neke su značajke međusobno linearno zavisne. Za grupiranje koristimo model Gaussove mješavine, i to: (a) model s nedijeljenom kovarijacijskom matricom i (b) model s dijeljenom izotropnom kovarijacijskom matricom. Za oba modela skicirajte (1) log-izglednost $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ i (2) Randov indeks R kao funkcije broja grupa K (ukupno četiri krivulje).