Strojno učenje – završni ispit

UNIZG FER, ak. god. 2016./2017.

27. siječnja 2017.

Ispit traje 180 minuta i nosi 35 bodova. Svaki zadatak rješavajte na zasebnoj stranici. Pišite uredno i čitko. Nemojte pretpostavljati da je nešto očito; Vaše znanje može se ocijeniti samo na temelju onog što napišete. Kod skica grafikona, označite osi, budite uredni i precizni te označite ekstreme krivulja, ako postoje.

1. (*8 bodova*) Procjenitelji.

   (a) Definirajte funkciju log-izglednosti $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ i objasnite na kojoj se pretpostavci ona temelji. Zašto radimo s logaritmom izglednosti i zašto je to opravdano?

   (b) Skicirajte $\ln \mathcal{L}(\mu, \sigma^2|\mathcal{D})$ kao funkciju od μ za skup primjera $\mathcal{D} = \{0, 2, 4\}$ uz pretpostavku da se primjeri ravnaju po Gaussovoj razdiobi, $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

   (c) Definirajte ML-procjenitelj te izvedite $\hat{\mu}_{\text{ML}}$, korak po korak, za parametar μ univarijatne Gaussove razdiobe. Je li ta procjena nepristrana i što to znači?

   (d) Definirajte MAP-procjenitelj. Kada MAP-procjenitelj ima rješenje u zatvorenoj formi?

   (e) Krenuvši od MAP-procjenitelja, izvedite Laplaceov procjenitelj za parametar μ Bernoullijeve varijable. Gustoća vjerojatnosti beta-distribucije jest $p(\mu|\alpha, \beta) = \mu^{\alpha-1}(1-\mu)^{\beta-1}/B(\alpha, \beta)$, a mod je $\frac{\alpha-1}{\alpha+\beta-2}$.

   (f) Objasnite koja je veza između MLE-procjena parametara $\mathbf{w}$ kod linearne i logističke regresije i minimizacije pogrešaka tih modela.

6 bodova

2. (*7 ~~bodova~~*) Bayesov klasifikator.

   (a) Napišite model naivnog Bayesovog klasifikatora. Napišite sve pretpostavke i opišite sve induktivne pristranosti ovog modela.

   (b) Definirajte bilo kakav polunaivan diskretan Bayesov klasifikator i napišite njegovu "generativnu priču".

   (c) Izgrađujemo Bayesov model za klasifikaciju primjera iz $\mathcal{X} = \mathbb{R}$ u tri klase. Učenjem na skupu primjera dobili smo sljedeće parametre modela: $P(\mathcal{C}_1) = 0.7$, $P(\mathcal{C}_2) = 0.2$, $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 1$, $\sigma_1^2 = 2$, $\sigma_2^2 = 3$, $\sigma_3^2 = 1$. Skicirajte funkcije gustoće vjerojatnosti $p(x|\mathcal{C}_j)$, $p(x)$ i $p(\mathcal{C}_j|x)$.

   (d) Kod multivarijatnog Bayesovog klasifikatora, izglednosti klasa definirane su gustoćom:

   $$p(\mathbf{x}|y) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right\}.$$

   Izvedite općeniti izraz za model $h(\mathbf{x}|\boldsymbol{\theta})$ s dijeljenom kovarijacijskom matricom. Skicirajte u prostoru $\mathcal{X} = \mathbb{R}^2$ konture gustoća razdiobe $p(x_1, x_2, y)$ i vjerojatnosti $P(y|x_1, x_2)$.

   (e) Izvedite vezu između modela logističke regresije i modela kontinuiranog Bayesovog klasifikatora. Na tom primjeru objasnite prednosti diskriminativnih modela nad generativnim.

3. (*6 bodova*) Vrednovanje i statističko testiranje klasifikatora.

   (a) Izračunajte makro-točnosti i makro-F_1 na temelju sljedeće matrice zabune (retci odgovaraju predviđenoj, a stupci stvarnoj kategoriji):

$$\begin{pmatrix} 10 & 4 & 6 \\ 8 & 19 & 8 \\ 5 & 5 & 21 \end{pmatrix}.$$

   (b) Za vrednovanje SVM-a koristimo (naravno) ugniježđenu unakrsnu provjeru s 5 vanjskih i 5 unutarnja preklopa. Hiperparametre optimiramo pretraživanjem po rešetci. Hiperparametri su jezgra (linearna ili RBF), parametar C ($C \in \{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$) i parametar γ ($\gamma \in \{2^{-5}, 2^{-4}, \ldots, 2^{5}\}$. Koliko ćemo puta ukupno trenirati model i kako biste odredili ukupno optimalne hiperparametre?

   (c) Trenirali smo model h_2 i želimo provjeriti je li njegov F_1 statistički značajno različit od F_1 modela h_1. Oba modela vrednujemo desetorostrukom unakrsnom provjerom na ukupno N=1000 primjera te računamo točnosti oba modela na svakom od deset preklopa (lijeva tablica).

i	$F_1(h_1)$	$F_1(h_2)$	i	$F_1(h_1)$	$F_1(h_2)$
1	0.627	0.595	5	0.562	0.518
2	0.570	0.581	6	0.462	0.677
3	0.396	0.630	7	0.541	0.613
4	0.529	0.691	8	0.539	0.631

df	0.10	0.05	0.02	0.01	0.005
7	1.895	2.365	2.998	3.499	4.029
8	1.860	2.306	2.897	3.355	3.833
9	1.833	2.262	2.821	3.250	3.690
10	1.812	2.228	2.764	3.169	3.581

   Uparenim t-testom testirajte je li točnost modela h_2 statistički značajno različita od točnosti h_1 na razini značajnosti $\alpha = 1\%$ te riječima formulirajte zaključak. Kritične vrijednosti za dvostrani t-test dane su u desnoj tablici (retci: stupnjevi slobode df; stupci: α za dvostrani test).

4. (*7 bodova*) Grupiranje.

   (a) Napišite funkciju pogreške algoritma k-srednjih vrijednosti i iz nje izvedite pseudokod algoritma. Koja je vremenska a koja prostorna složenost ovog aloritma?

   (b) Sličnosti između primjera definirane su sljedećom matricom *sličnosti*:

$$S = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1.0 & 0.1 & 0.3 & 0.1 \\ 0.1 & 1.0 & 0.1 & 0.2 \\ 0.3 & 0.1 & 1.0 & 0.2 \\ 0.1 & 0.2 & 0.2 & 1.0 \end{pmatrix} \end{array}$$

   Primijenite hijerarhijsko aglomerativno grupiranje (HAC) s jednostrukim povezivanjem te skicirajte pripadni dendrogram. Na dendrogramu naznačite sličnosti na kojima se odvija spajanje grupa.

   (c) Raspolažemo skupom neoznačenih primjera i manjim podskupom od 8 primjera označenih u tri klase te želimo napraviti provjeru grupiranja. Za $K = 3$ i $K = 4$, algoritam primjere grupira u particiju $\{\{0,2,2\},\{0,0\},\{1,1,2\}\}$ odnosno $\{\{0,2,2\},\{0,0\},\{1,1\},\{2\}\}$. Izračunajte Randove indekse. Skicirajte krivulju Randovog indeksa kao funkciju broja grupa K.

   (d) Napišite izraz za mješavinski model s latentnim varijablama. Koja je značenje latentnih varijabli? Izedite izraz za (potpunu) log-izglednost i ukratko objasnite na koji način dalje provodimo optimizaciju.

   (e) Raspolažemo neoznačenim podatcima $\mathcal{D}$ koji potječu iz K=12 klasa i čije su značajke međusobno visoko linearno zavisne. Podatke grupiramo modelom Gaussovih mješavina, i to modelom s nedijeljenim kov. matricama (GMM-full) i s dijeljenom dijagonalnom kov. matricom (GMM-diag). Skicirajte očekivani izgled log-izglednosti $\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ kao funkcije broja iteracija, i to za: (1) GMM-full s nasumično odabranih K=12 središta i (2) inicijaliziran algoritmom k-srednjih vrijednosti sa K=12 središta te GMM-diag inicijaliziran algoritmom k-srednjih vrijednosti sa (3) K=12 središta algoritmom k-srednjih vrijednosti i (4) s nasumično odabranih K=100 središta (ukupno četiri krivulje).