UNIZG FER, **Strojno učenje**, ak. god. 2011/12.

Domaća zadaća 2

Zadano:	**21.11.2011.**
Rok predaje:	**9.12.2011.** do **17.00 sati**

Napomena: Rješenju treba priložiti izvorne kodove i datoteke sa skupovima za učenje.

1. Na raspolaganju nam je skup za učenje $\mathcal{D}$ koji se sastoji od deset primjera iz $\mathcal{X} = \mathbb{R}^4$:

i	x_1	x_2	x_3	x_4
1	−1.28	−1.51	−1.65	−2.29
2	−1.83	−0.99	−2.51	−0.72
3	3.24	0.11	2.15	−1.67
4	−2.47	−1.95	−0.65	−2.61
5	−1.07	0.15	−0.40	−2.31
6	−1.58	0.06	0.26	−0.55
7	−1.00	−0.95	−0.63	−2.12
8	−0.53	−0.67	−1.40	−1.65
9	0.50	−0.91	1.31	−1.80
10	0.70	0.11	−2.04	−1.83

   (a) Izračunajte ML-procjenu kovarijacijske matrice $\hat{\mathbf{\Sigma}}$.

   (b) Izračunajte Mahalanobisovu udaljenost između točaka $\mathbf{x}^{(1)} = (1, 1, 1, 1)$ i $\mathbf{x}^{(2)} = (1, 1, -1, -1)$. Je li ta udaljenost jednaka euklidskoj udaljenosti? Zašto? Je li to dobro ili loše?

   (c) Kako možemo utvrditi jesu li dvije varijable statistički nezavisne? Pokušajte to napraviti za varijable x_1 i x_2. Možemo li definitivno utvrditi da su varijable zavisne ili da su nezavisne? Obrazložite odgovor.

2. (a) Formalno definirajte funkciju log-izglednosti. Koje pretpostavke o skupu primjera za učenje $\mathcal{D}$ su ugrađene u tu definiciju? Kako bi izgledala definicija funkcije log-izglednosti kada te pretpostavke (sve ili neke od njih) ne bi vrijedile?

   (b) Imamo skup primjera $\mathcal{D}$ za koji pretpostavljamo da su normalno distribuirani. Izračunajte log-izglednost $\mathcal{L}(\mu = 1, \sigma^2 = 1|\mathcal{D})$ za skup primjera

$$\mathcal{D} = \{0.2, 0.5, 1, 2, 8, 10\}.$$

   (c) Dan je (neoznačen) uzorak $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$. Izvedite procjenu $\hat{a}$ za parametar geometrijske distribucije metodom najveće izglednosti. Funkcija gustoće vjerojatnosti geometrijske distribucije je $p(x|a) = a(1-a)^{x-1}$.

3. (a) Svaki algoritam strojnog učenja sastoji se se od tri osnovne komponente. Identificirajte te komponente kod naivnog Bayesovog klasifikatora.

   (b) Faktorizirajte zajedničku vjerojatnost $P(x_1, \ldots, x_5)$ na tri različita načina. Koje pretpostavke o nezavisnosti varijabli proizlaze iz tih faktorizacija? Nacrtajte odgovarajuće Bayesove mreže.

(c) Pretpostavite da se radi o binarnoj klasifikaciji i da su dvije značajke binarne, a tri su ternarne. Izračunajte broj parametara svakog od tri modela.

(d) Osmislili ste svoj potpuno nov algoritam za konstrukciju polunaivnog Bayesovog klasifikatora koji ste nazvali ASJ (*A-star Selection and Joining*). ASJ je napredna verzija algoritma FSSJ. Isto kao i algoritam FSSJ, algoritam ASJ čvorove spaja u superčvorove, ali, za razliku od algoritma FSSJ, koji koristi pohlepno pretraživanje, algoritam ASJ za pretraživanje koristi algoritam A^*. Pritom se za heuristiku koristiti empirijska pogreška, tj. u svakoj se iteraciji odabire ona operacija koja smanjuje pogrešku na skupu za učenje.

Je li algoritam ASJ, kako je ovdje opisan, dobar? Postoje li nekakvi nedostatci ili problemi koje možemo očekivati? Obrazložite odgovore. Predložite poboljšanja.

4. U ovom zadatku potrebno je napraviti programsku izvedbu naivnog Bayesovog klasifikatora i provesti eksperimente nad stvarnim podatcima.

(a) Napravite programsku izvedbu naivnog Bayesovog klasifikatora za diskretne ulaze u programskome jeziku po izboru. Klasifikator treba podržavati izračun ML-procjena te izračun zaglađenih (Laplaeovih) procjena.

(b) Skinite skup podataka *Iris* s adrese

`http://archive.ics.uci.edu/ml/datasets/iris`

U njemu se opisuju biometrijske značajke cvjetova triju vrsta perunike odnosno irisa (*virginica*, *setosa* i *versicolor*). Skup najprije razdijelite slučajnim odabirom na podskup za učenje i ispitivanje u omjeru 2:1. Diskretizirajte značajke na skupu za učenje tako da približno trećina vrijednosti u svakom stupcu bude preslikana u jednu diskretnu vrijednost čime ćemo dobiti skup s diskretnim značajkama koje imaju po tri vrijednosti (radi jednostavnosti, možemo ih označiti brojevima 0, 1 i 2).

(c) Naučite model naivnog Bayesovog klasifikatora na skupu za učenje. Ispišite empirijsku pogrešku i pogrešku generalizacije.

(d) Ponovite eksperimente koristeći Laplaceovo zaglađivanje. Komentirajte rezultate.

5. U ovom zadatku trebate izgraditi multivarijatni generativni model za dvije klase s kontinuiranim značajkama. U datoteci `su-2011-dz2-zad5-podatci.txt` nalaze se primjeri koji su nastali iz 12 višedimenzijskih Gaussovih izvora. Prva dva stupca označavaju koordinate primjera, a treći stupac označava klasu primjera. U programima Matlab/Octave podatci se mogu učitati na ovaj način:

```
podatci=dlmread('su-2011-dz2-zad5-podatci.txt');
X = podatci(:, 1:2);
y = podatci(:, 3);
```

(a) Nasumično odaberite dvije klase (od mogućih 12). Napišite koje ste klase odabrali.

(b) Za odabrane klase izračunajte $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\Sigma}}_j$ i $\hat{P}(\mathcal{C}_j)$. Podskup skupa $\mathcal{X}$ koji odgovara npr. klasi s oznakom 5 možete pronaći ovako:

```
podskupX = X(y==5, :);
```

(c) Skup podijelite na skup za učenje, skup za provjeru, i skup za ispitivanje u omjeru 3:1:1.

(d) Izgradite tri modela različite složenosti: (1) model s različitim kovarijacijskim matricama, (2) model s dijeljenom kovarijacijskom matricom i (3) model s dijeljenom izotropnom kovarijacijskom matricom.

(e) Za sva tri modela nacrtajte primjere iz prve i druge klase te decizijsku granicu između dviju klasa. Decizijska granica dobiva se rješavanjem jednadžbe $P(\mathcal{C}_1|x) = P(\mathcal{C}_2|x)$. Funkcijom `ezplot` mogu se crtati implicitne jednadžbe. Primjer crtanja u sustavu Matlab/Octave:

```
plot(X(y==5,1),X(y==5,2),'o');
hold on;
plot(X(y==6,1),X(y==6,2),'+');
ezplot('-3*x-2*y+1',[-10 10 -10 10]);
hold off;
xlabel('x');
ylabel('y');
print -deps slika.eps
```

Nakon izvođenja ovog odsječka slika će se spremiti u datoteku `slika.eps` u `eps` formatu.

(f) Izračunajte pogrešku generalizacije za svaki od tri modela na skupu za provjeru. Odaberite optimalan model. Obrazložite svoju odluku.

(g) Naučite odabrani model na uniji skupa za učenje i skupa za provjeru (4/5 ukupnog skupa). Izračunajte pogrešku generalizacije takvog modela na skupu za ispitivanje (1/5 ukupnog skupa). Je li ona manja ili veća od one koju ste dobili na skupu za provjeru u zadatku 5f? Komentirajte zašto je to tako.