

Strojno učenje – domaća zadaća 9

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

Zadano: 19. 12. 2016. Rok: 4. 1. 2017.

Napomena: Zadatke možete rješavati samostalno ili u grupi. Ako zadatke rješavate u grupi, pobrinite se da svi članovi grupe pridonose rješenju i da ga naposlijetu svi razumiju. Po potrebi konzultirajte sve dostupne izvore informacija. Rješenja zadataka ponesite na iduće auditorne vježbe. Zabilježite sve nejasnoće i nedoumice, kako bismo ih prodiskutirali.

1. [Svrha: Razumjeti model Bayesovog klasifikatora i njegove komponente. Razumjeti što su to generativni modeli, kako se razlikuju od diskriminativnih te koje su njihove prednosti i njihovi nedostatci.]
 - (a) Definirajte model Bayesovog klasifikatora i navedite sve veličine koje se pojavljaju u definiciji modela. Objasnite zašto faktoriziramo brojnik. Objasnite ulogu nazivnika i objasnite kada ga možemo zanemariti.
 - (b) Je li taj model parametarski ili neparametarski? Obrazložite odgovor.
 - (c) Objasnite zašto Bayesov model nazivamo generativnim i opišite generativnu priču Bayesovog klasifikatora.
 - (d) Objasnite razliku između generativnih i diskriminativnih modela te navedite prednosti jednih i drugih.
2. [Svrha: Isprobati izračun maksimalne aposteriorne vjerojatnosti i najvjerojatnije hipoteze uz minimizaciju rizika.] Razmotrimo problem klasifikacije neželjene el. pošte u klase *spam* ($y = 1$), *important* ($y = 2$) i *normal* ($y = 3$). Neka su apriorne vjerojatnosti tih klasa $P(y = 1) = 0.2$, $P(y = 2) = 0.05$ i $P(y = 3) = 0.75$. Za neku poruku el. pošte \mathbf{x} izglednosti iznose $p(\mathbf{x}|y = 1) = 0.8$ i $p(\mathbf{x}|y = 2) = p(\mathbf{x}|y = 3) = 0.5$. Izračunajte aposteriorne vjerojatnosti za svaku od klasa te maksimalnu aposteriornu hipotezu za primjer \mathbf{x} .
3. [Svrha: Razumjeti faktorizaciju zajedničke vjerojatnosti uz pretpostavku uvjetne nezavisnosti te povezanost toga s induktivnom pristranošću i, posljedično, brojem značajki modela.]
 - (a) Definirajte naivan Bayesov klasifikator i pretpostavku na kojoj se temelji.
 - (b) Zašto nam treba pretpostavka o uvjetnoj nezavisnosti značajki te kojoj vrsti induktivne pristranoosti ona odgovara?
 - (c) Naivan Bayesov klasifikator koristimo za klasifikaciju rukom pisanih znamenki u jednu od deset klasa. Znamenke su prikazane kao vektor binarnih značajki (crno/bijeli slikovni elementi) u matrici s razlučivošću 32×32 . Odredite ukupan broj parametara naivnog Bayesovog klasifikatora.

4. [Svrha: Isprobati na konkretnom primjeru izračun parametara naivnog Bayesovog klasifikatora.] Naivan Bayesov model želimo upotrijebiti za binarnu klasifikaciju “Skupo ljetovanje na Jadranu”. 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	Dalmacija	ne	privatni	avion	ne
5	Istra	ne	privatni	auto	da
6	Kvarner	ne	kamp	bus	ne
7	Dalmacija	da	hotel	auto	da

- (a) Izračunajte MLE procjene svih parametara modela te klasificirajte primjere (Istra, ne, kamp, bus) i (Dalmacija, da, hotel, bus).
- (b) Izračunajte Laplaceove (zaglađene) procjene za sve parametre modela te klasificajte nanovo iste primjere.
5. [Svrha: Razumjeti definiciju uzajamne informacije i način njezina izračuna.]
- (a) Krenuvši od definicija za entropiju i relativnu entropiju, izvedite mjeru uzajamne informacije $I(X, Y)$ kao Kullback-Leiblerovu divergenciju između zajedničke razdiobe, $P(X, Y)$, i zajedničke razdiobe uz pretpostavku nezavisnosti, $P(X)P(Y)$.
- (b) Izračunajte mjeru uzajamne informacije $I(X, Y)$ za varijable X i Y s razdiobom definiranom u zadatku (1) u zadaći 8. Biste li, temeljem vrijednosti uzajamne informacije, rekli da su varijable X i Y nezavisne?
- (c) Uzajamna informacija nije odozgo ograničena, ali je ograničena odozdo. Primjenom Jensenove nejednakosti, dokažite da vrijedi $I(X, Y) \geq 0$.

6. [Svrha: Razviti intuiciju za model kontinuiranog Bayesovog klasifikatora.]

Izrađujemo Bayesov model za klasifikaciju primjera iz $\mathcal{X} = \mathbb{R}$ u tri klase. Učenjem na skupu primjera dobili smo sljedeće parametre modela: $P(y=1) = 0.3$, $P(y=2) = 0.2$, $\mu_1 = -5$, $\mu_2 = 0$, $\mu_3 = 5$, $\sigma_1^2 = 5$, $\sigma_2^2 = 1$, $\sigma_3^2 = 10$. Skicirajte funkcije gustoće vjerojatnosti $p(x|y)$, $p(x,y)$, $p(x)$ i $p(y|x)$.

7. [Svrha: Razumjeti izvod modela kontinuiranog Bayesovog klasifikatora i osvježiti potrebno znanje matematike.]
- (a) Krenuvši od izraza (4.29) iz skripte, izvedite model višedimenzijskog Bayesovog klasifikatora s kontinuiranim ulazima s dijeljenom i dijagonalnom kovarijacijskom matricom.
- (b) Napišite broj parametara ovog modela.
- (c) Objasnite zašto je izglednost faktorizirana u produkt univarijatnih razdioba, što odgovara pretpostavci o uvjetnoj nezavisnosti, premda značajke mogu biti nelinearno uvjetno zavisne.

8. [Svrha: Razviti intuiciju za složenost modela kontinuiranog Bayesovog klasifikatora i shvatiti kako se problem u konačnici svodi na odabir optimalnog modela.] Želimo izgraditi klasifikator za klasifikaciju brukoša u jednu od dvije klase: $y = 1 \Rightarrow$ "Završava FER u roku" i $y = 2 \Rightarrow$ "Produljuje studij". Svaki je primjer opisan sa šest ulaznih varijabli: prosjek ocjena 1.–4. razreda (četiri varijable), bodovi državne mature iz matematike te bodovi državne mature iz fizike. Raspolažemo trima modelima: modelom \mathcal{H}_1 s dijeljenom kovarijacijskom matricom, modelom \mathcal{H}_2 s dijagonalnom (i dijeljenom) kovarijacijskom matricom i modelom \mathcal{H}_3 s izotropnom kovarijacijskom matricom.
- (a) Koliko svaki od ova tri modela ima parametara?
 - (b) Za koji od ova tri modela očekujete da će najbolje generalizirati u ovom konkretnom slučaju (uzmite u obzir prirodu problema i očekivane odnose između značajki)? Zašto?
 - (c) Nacrtajte skicu funkcije empirijske pogreške i pogreške generalizacije i naznačite na njoj točke koje označavaju navedenim trima modelima.
 - (d) Kako biste u praksi odredili koji će model upotrijebiti?
9. [Svrha: Razumjeti vezu između Bayesovog klasifikatora i logističke regresije, odnosno probabilističku interpretaciju logističke regresije. Razumjeti razliku u broju parametara između diskriminativnog i generativnog modela te utjecaj broja klasa i broja primjera na taj odnos.]
- (a) Izvedite model logističke regresije krenuvši od generativne definicije za $P(y = 1|\mathbf{x})$. Izvod napravite korak po korak te se uvjerite da možete obrazložiti svaki korak u izvodu. Napišite sve pretpostavke koje ste ugradili u izvod.
 - (b) Model logističke regresije koristimo za binarnu klasifikaciju primjera s $n = 100$ značajki. Odredite broj parametara modela logističke regresije te njemu odgovarajućeg generativnog modela.
 - (c) Izračunajte broj parametara za isti slučaj, ali sa $K = 5$ klasa.
 - (d) Prepostavite da klasificiramo u $K = 10$ klasa. Izračunajte koliko velika mora biti dimenzija prostora značajki n , a da bi se logistička regresija isplatila jer ima manje parametara od odgovarajućega generativnog modela.

8. POMAČA ZADÁČA

1.

$$\begin{array}{ll} P(1,1) = 0.2 & P(2,1) = 0.05 \\ P(1,2) = 0.05 & P(2,2) = 0.3 \\ \underline{P(1,3) = 0.3} & \underline{P(2,3) = 0.1} \end{array}$$

$P(x,y)$

$$E(x) = \sum_i x_i P(x_i)$$

$$P(x=1) = 0.2 + 0.05 + 0.3 = 0.55$$

$$P(x=2) = 0.05 + 0.3 + 0.1 = 0.45$$

$$E(x) = 0.55 + 2 \cdot 0.45$$

$$\boxed{E(x) = 1.45}$$

$$\text{Var}(x) = E[(x - E[x])^2]$$

$$\text{Var}(x) = (1.45 - 1)^2 \cdot 0.55 + (1.45 - 2)^2 \cdot 0.45$$

$$\boxed{\text{Var}(x) = 0.2475}$$

$$\text{Cov}(x,y) = E[(x - E[x])(y - E[y])]$$

$$\text{Cov}(x,y) = \sum_{x_i} \sum_{y_i} P(x_i, y_i) (x_i - E[x])(y_i - E[y])$$

$$P(y=1) = 0.2 + 0.05 = 0.25$$

$$P(y=2) = 0.35$$

$$P(y=3) = 0.4$$

$$E[y] = 0.25 + 2 \cdot 0.35 + 3 \cdot 0.4 = 2.15,,$$

$$\begin{aligned} \text{Cov}(x,y) &= 0.2(1-1.45)(1-2.15) + 0.05(2-1.45)(1-2.15) + \\ &+ 0.05(1-1.45)(2-2.15) + 0.3(2-1.45)(2-2.15) + \\ &+ 0.3(1-1.45)(3-2.15) + 0.1(2-1.45)(3-2.15) \end{aligned}$$

$$\boxed{\text{Cov}(x,y) = -0.0175}$$

$$\gamma(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{\text{Var}(x)} = 0.4975,$$

$$\sigma_y = \sqrt{\text{Var}(y)} = \sqrt{(2.15-1)^2 \cdot 0.25 + (2.15-2)^2 \cdot 0.35 + (2.15-3)^2 \cdot 0.4}$$

$$\sigma_y = 0.7921,,$$

$$\gamma(x,y) = -0.0444$$

$$\Sigma = \begin{bmatrix} \text{Var}(x) & \text{Cov}(x,y) \\ \text{Cov}(y,x) & \text{Var}(y) \end{bmatrix} \quad \text{Cov}(y,x) = \text{Cov}(x,y)$$

$$\text{Var}(y) = \sigma_y^2$$

$$\Sigma = \begin{bmatrix} 0.2475 & -0.0175 \\ -0.0175 & 0.6275 \end{bmatrix}$$

2) a) VARIJABLE x i y SU NEZAVISNE AKO:

$$P(x,y) = P(x) P(y), \quad \text{I. E.}$$

$$P(x|y) = P(x), \quad P(y|x) = P(y)$$

b) VARIJABLE IZ 1. ZADATKA NISU LINEARNO ZAVISNE JER IM JE OBZOR KOEFICIJENTA KORELACIJE BLIZU NULE. NE ZNAMO JESU LI NEZAVISNE, MOGU BITI NELINEARNO ZAVISNE (ZNAMO SAMO DA NISU LINEARNO ZAVISNE).

- c) i) ZAVISNE ; iii) ZAVISNE
 ii) LINEARNO ZAVISNE ; iv) ZAVISNE

d) $P(X, Y) = P(X)P(Y) \rightarrow \text{NEZAVISNOST}$

$$E(XY) = \sum p(x,y)xy = \sum p(x)x \cdot p(y)y = \sum x p(x) \sum y p(y)$$

$$E[XY] = E[X]E[Y],$$

$$\text{cov}(X, Y) = E[(X - E[X])(Y - E[Y])] =$$

$$= E[XY - YE[X] - XE[Y] + E[X]E[Y]] =$$

$$= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] = 0,$$

$$\text{cov}(X, Y) = 0 \rightarrow \text{Y}(X, Y) = 0 \rightarrow \text{NEKORELIRANOST},$$

NEZAVISNE VARIABLE SU LINEARNO NEKORELIRANE.

(3.)

a) $\mathcal{L}: \Theta \rightarrow P(D|\Theta)$

$$\mathcal{L}(\theta|D) = p(D|\theta) = p(x^{(1)}, \dots, x^{(n)}|\theta) = \prod_{i=1}^N p(x^{(i)}|\theta)$$

\hookrightarrow VJEROJATNOST REALIZACIJE UZORKA D AKO JE PARAMETAR POPULACIJE JEDNAK θ

TEMELJI SE NA IID PRETPOSTAVCI; PRETPOSTAVLJAMO DA SU PRIMJERI U SKUPU D NEZAVISNI I DA POTJEČU OD IDENTIČNE RAZDILOBE.

$$b) \mathcal{L}(\mu, \sigma^2 | D) = \prod_{i=1}^N p(x^{(i)} | \mu, \sigma^2) = \left\{ p(x^{(i)} | \mu, \sigma^2) \sim N(\mu, \sigma^2) \right\}$$

$$= \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right),$$

$$\mu = 0, \sigma^2 = 1$$

D - ZADAN

$$\mathcal{L}(\mu, \sigma^2 | D) = p(D | \mu, \sigma^2) = \underbrace{\{ \text{MATLAB} \}}_{1.9026 \cdot 10^{-22}},$$

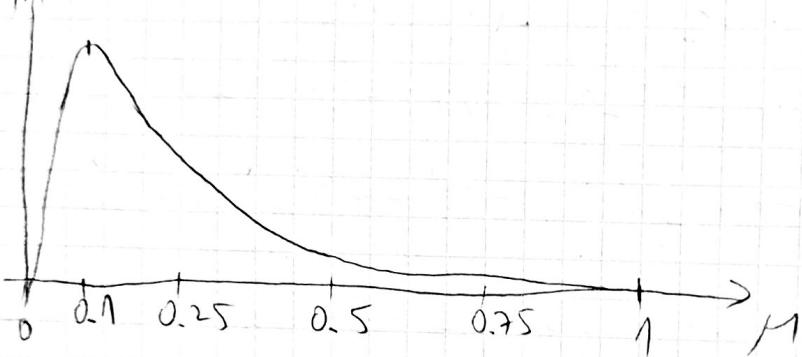
c) D SE POKORAVA BERNOULLIJEVJ VARIABLE
(REZULTAT JE ILI GLAVA ILI PISMO)

$$\mathcal{L}(\mu | N, m) = \mu^m (1-\mu)^{N-m}$$

μ — VEROJATNOST DA SRIO DOBIL GLAVU.

d) $N=10, m=1$

$$\mathcal{L}(\mu | N, m)$$



(4)

a)

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta | D)$$

b)

$$\begin{aligned} \ln \mathcal{L}(\mu | D) &= \ln \prod_{i=1}^N p(x_i | \mu) = \ln \prod_{i=1}^N \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}} = \\ &= \sum_{i=1}^N \left(x^{(i)} \ln \mu + (1-x^{(i)}) \ln (1-\mu) \right) \end{aligned}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = \frac{1}{\mu} \sum_{i=1}^N x^{(i)} - \frac{1}{1-\mu} \sum_{i=1}^N (1-x^{(i)}) = \frac{1}{\mu} \sum_{i=1}^N x^{(i)} -$$

$$- \frac{N}{1-\mu} + \frac{1}{1-\mu} \sum_{i=1}^N x^{(i)} = \frac{1}{\mu(1-\mu)} \sum_{i=1}^N x^{(i)} - \frac{N}{1-\mu} = 0$$

$$\frac{1}{\mu} \sum_{i=1}^N x^{(i)} - N = 0 \rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)}}$$

$$c) \ln L(\mu | D) = \ln \prod_{i=1}^N p(x_i | \mu) = \ln \prod_{i=1}^N \prod_{k=1}^K \mu_k^{x_k^{(i)}} =$$

$$= \sum_{i=1}^N \sum_{k=1}^K x_k^{(i)} \ln \mu_k , \text{ UZ OGRANIČENJE:}$$

$$\sum_{k=1}^K \mu_k = 1$$

LAGRANGEJAVA FUNKCIJA:

$$L = \sum_{i=1}^N \sum_{k=1}^K x_k^{(i)} \ln \mu_k + \lambda \left(\sum_{k=1}^K \mu_k - 1 \right)$$

→ LAGRANGEOV MULTIPLIKATOR

$$\frac{\partial L}{\partial \mu_k} = \frac{1}{\mu_k} \sum_{i=1}^N x_k^{(i)} + \lambda \rightarrow \mu_k = -\frac{1}{\lambda} \sum_{i=1}^N x_k^{(i)}$$

UVRŠTAVANJE U OGRANIČENJE:

$$-\sum_{k=1}^K \mu_k = -\frac{1}{\lambda} \underbrace{\sum_{k=1}^K \sum_{i=1}^N x_k^{(i)}}_N = 1 \rightarrow \boxed{\lambda = -N}$$

SLJEDI:

$$\boxed{\mu_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N x^{(i)} = \frac{N_k}{N}} \rightarrow \text{BROJ KOLIKO JE PUTA VARIJABLA } V \text{ VREDNOSTI } k$$

$$d) \ln L(\mu, \sigma | D) = \ln \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2}\right) =$$

$$= \sum_{i=1}^N \left(-\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^N (x^{(i)} - \mu)^2 \right)$$

$$= -N \ln \sqrt{2\pi} - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^N (x^{(i)} - \mu)^2 //$$

$$- \frac{N}{2} \ln \sigma^2 \rightarrow \text{ZBOG LAKŠEG RAČUNA KASNIJE}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = -\frac{1}{2\sigma^2} \cdot 2 \sum_{i=1}^N (x^{(i)} - \mu) \cdot (-1)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^N (x^{(i)} - \mu) = 0$$

$$\sum_{i=1}^N x^{(i)} = \sum_{i=1}^N \mu \rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x^{(i)}}$$

$$\frac{\partial \ln \mathcal{L}}{\partial (\sigma^2)} = -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^N (x^{(i)} - \mu)^2 = 0$$

$$N = \frac{1}{\sigma^2} \sum_{i=1}^N (x^{(i)} - \mu)^2 \rightarrow \boxed{\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \mu)^2}$$

5. a) PRISTRANOST PROCJENITELJA:

$$b_\theta(\theta) = E[\theta] - \theta$$

$$b(\hat{\mu}) = E[\hat{\mu}] - \hat{\mu}$$

$$E[\hat{\mu}] = \frac{1}{N} E\left[\sum_{i=1}^N x^{(i)}\right] = \frac{1}{N} \cdot N \mu = \hat{\mu}_{II}$$

$$b(\hat{\mu}) = 0 \rightarrow \hat{\mu}_{MLE} \text{ NEPRISTRAN!}$$

$$b(\sigma^2) = E[\hat{\sigma}^2] - \sigma^2$$

$$E[\hat{\sigma}^2] = \frac{1}{N} E\left[\sum_{i=1}^N (x^{(i)} - \hat{\mu})^2\right] = \frac{1}{N} E\left[\sum_{i=1}^N (x^{(i)2} - 2x^{(i)}\hat{\mu} + \hat{\mu}^2)\right] =$$

$$= \frac{1}{N} E\left[\sum_{i=1}^N x^{(i)2} - 2N\hat{\mu}^2 + N\hat{\mu}^2\right] = \frac{1}{N} (N E[x^2] - N E[\hat{\mu}^2]) =$$

$$= E[x^2] - E[\hat{\mu}^2]_{II}$$

$$\text{Var}(x) = E[x^2] - E[x]^2 \rightarrow E[x^2] = \text{Var}(x) + E[x]^2$$

$$E[x^2] = \sigma^2 + \mu^2$$

$$\text{Var}(\hat{\mu}) = E[\hat{\mu}^2] - E[\hat{\mu}]^2 \rightarrow E[\hat{\mu}^2] = \text{Var}(\hat{\mu}) + E[\hat{\mu}]^2$$

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{N} \sum_{i=1}^N x^{(i)}\right) = \frac{1}{N} \sum_{i=1}^N \text{Var}(x^{(i)}) = \frac{1}{N} \sigma^2$$

$$E[\hat{\mu}^2] = \frac{\sigma^2}{N} + \hat{\mu}^2$$

$$\text{er}(\hat{\mu}^2) = E[\hat{\mu}^2] - \sigma^2 = E[x^2] - E[\hat{\mu}^2] - \sigma^2 =$$

$$= \sigma^2 + \hat{\mu}^2 - \frac{\sigma^2}{N} - \hat{\mu}^2 - \sigma^2 = -\frac{\sigma^2}{N} //$$

$$\text{er}(\hat{\mu}^2) = -\frac{\sigma^2}{N}$$

b) AKO IMAMO MALI BROJ PRIMJERA, PRISTRANOST PROCJENITELJA $\text{er}(\hat{\mu}^2)$ JE RELATIVNO VELIKA PA ĆE I POGREŠKA PROCJENITELJA BITI VELIKA I TO JE PROBLEM. ZA VELIKI BROJ PRIMJERA (N), PRISTRANOST $\text{er}(\hat{\mu}^2) = -\sigma^2/N$ TEŽI K NULI PA U TOM SLUČAJU PRISTRANOST NIJE PROBLEMATIČNA.

(6) a)

JER SU $x^{(i)}$ VEKTORI $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x^{(i)}$

$$\hat{\mu}_{MLE} = \frac{1}{6} \sum_{i=1}^6 x^{(i)} = (5.483, -0.183, -0.733) //$$

$$\hat{\Sigma}_{MLE} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}^{(i)} - \hat{\mu})(\mathbf{x}^{(i)} - \hat{\mu})^T$$

↑
 $n \times n$
 $n \times 1$
 STUPAC

↑
 $1 \times n$
 REDAK

MATLAB:

$$\hat{\Sigma}_{MLE} = \begin{bmatrix} 8.7714 & -1.1981 & -4.7922 \\ -1.1981 & 0.1914 & 0.7656 \\ -4.7922 & 0.7656 & 3.0622 \end{bmatrix},$$

b) $\mathbf{x} = (-2, 1, 0)^T$

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

↓
 GUSTOĆA VJEROJATNOSTI

MATRICA KOVARIJACIJE $\hat{\Sigma}_{MLE}$ JE LOŠE KONDICIONIRANA PA JE IZRAČUN INVERZA $\hat{\Sigma}^{-1}$ NUMERIČKI NESTABILAN; NE MOŽEMO IZRAČUNATI GUSTOĆU VJEROJATNOSTI.

c)

$$\gamma_{x_1, x_2} = \frac{\text{Cov}(x_1, x_2)}{\sigma_{x_1} \sigma_{x_2}} = \frac{-1.1981}{\sqrt{8.7714 \cdot 0.1914}} = -0.9247$$

$$\gamma_{x_1, x_3} = \frac{\text{Cov}(x_1, x_3)}{\sigma_{x_1} \sigma_{x_3}} = \frac{-4.7922}{\sqrt{8.7714 \cdot 3.0622}} = -0.9247$$

$$\gamma_{x_2, x_3} = \frac{\text{Cov}(x_2, x_3)}{\sigma_{x_2} \sigma_{x_3}} = \frac{0.7656}{\sqrt{0.1914 \cdot 3.0622}} = 1$$

VALUABLE x_2 i x_3 SU LINEARNO ZAVISNE
PA IZBACUJEM x_3 :

$$x^{(1)} = (9.5, -0.7) \quad x^{(2)} = (8.8, -0.8)$$

$$x^{(3)} = (6.5, -0.2) \quad x^{(4)} = (2.3, 0.3)$$

$$x^{(5)} = (2.2, 0) \quad x^{(6)} = (3.6, 0.3)$$

GUSTOĆA VJEROJATNOSTI: (MATLAB)

$$x = (-2, 1)^T$$

$$p(x | \hat{\mu}_{MLE}, \hat{\Sigma}_{MLE}) = 0.0083 //$$

7.

a) $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} p(\theta | D) = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta | D) p(\theta),$

MAP JE BOLJI OD MLE ZATO ŠTO OMOGUĆUJE UGRADNU APRIORNOG ZNANJA U IZRAČUN PARAMETARA ČIME MOŽEMO SMANJITI PRENAUČENOST.

e) $p(\mu | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \mu^{\alpha-1} (1-\mu)^{\beta-1}$

$$\mathcal{L}(\mu | N, m) = \mu^m (1-\mu)^{N-m} = p(D | \mu)$$

$$p(\mu | D) = p(H | N, m, \alpha, \beta) = \frac{p(D | \mu) p(\mu | \alpha, \beta)}{p(D)}$$

$$p(\mu | N, m, \alpha, \beta) = \frac{1}{p(D) B(\alpha, \beta)} \mu^{\alpha+m-1} (1-\mu)^{N+\beta-m-1}$$

c) $P(\mu | \alpha=2, \beta=2), N=10, m=1$

$$\mathcal{L}(\mu | N=10, m=1)$$

U slj. je izražena aposteriorna vjerojatnost

$$P(\mu | N, m, \alpha, \beta) = \frac{1}{\rho(D)B(\alpha, \beta)} \mu^{\alpha+m-1} (1-\mu)^{N+\beta-m-1}$$

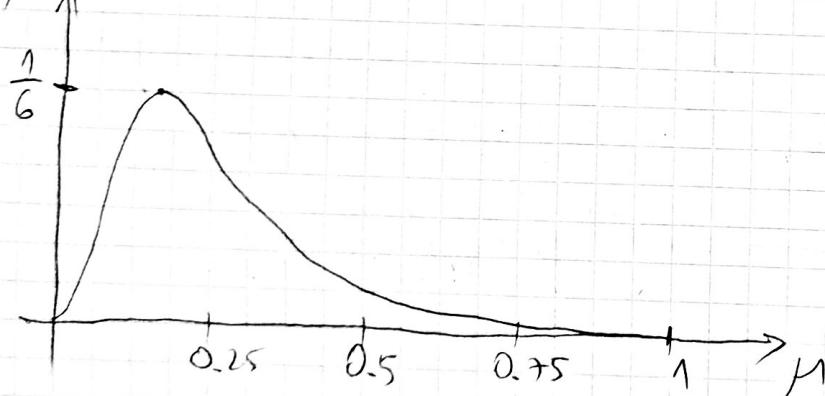
$$\underbrace{\alpha^* = \alpha + m, \beta^* = N + \beta - m}_{\text{APOSTERIORNA VJEROJATNOST JE BETA DISTRIBUCIJA S PARAMETRIMA } \alpha^*, \beta^*}$$

APOSTERIORNA VJEROJATNOST JE BETA DISTRIBUCIJA S PARAMETRIMA α^*, β^*

MAXIMUM APOSTERIORNE VJEROJATNOSTI:

$$\max P(\mu | \alpha^*, \beta^*) = \frac{\alpha^*-1}{\alpha^* + \beta^* - 2} = \frac{\alpha+m-1}{\alpha+N+\beta-2} = \frac{2}{12} = \frac{1}{6}$$

$$P(\mu, \alpha=2, \beta=2)$$



d) $\hat{\mu}_{MLE} = \frac{m}{N} = \frac{1}{10}, \quad \hat{\mu}_{MAP} = \frac{\alpha+m-1}{\alpha+N+\beta-2} = \frac{1}{6}$

MAP procjena je veća, odnosno bliža 0.5 što je pretpostavljeni μ (ako je novčić pravedan). MLE je prenaučen; previše se prilagođio podacima.

S PORASTOM BROJA PRIMJERA N , RAZLICA IZMEDU

MLE I MAP PROCJENE BI SE SMANJIVALA.

c)

APOSTERIORNA VJEROJATNOST:

$$P(\mu | N, m, \alpha, \beta) = \frac{1}{P(D)B(\alpha, \beta)} \mu^{\alpha+m-1} (1-\mu)^{N+\beta-m-1}$$

MAKSIMUM APOSTERIORNE VJEROJATNOSTI (MAP):

$$\max P(\mu | N, m, \alpha, \beta) = \frac{\alpha+m-1}{N+\alpha+\beta-2}$$

Uz $\alpha=2$, $\beta=2$:

$$\max P(\mu | N, m, \alpha, \beta) = \frac{m+1}{N+2} \Rightarrow \text{LAPLAČEOV PROCJENITELJ}$$

(8.)

a) $y^{(i)} = f(x^{(i)}) + \varepsilon \rightarrow \varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$p(y|x) = \mathcal{N}(f(x), \sigma^2) \quad h(x; w) = f(x) = w^T x$$

$$\ln \mathcal{L}(w|D) = \ln p(D|w) = \ln \prod_{i=1}^N p(y^{(i)} | x^{(i)}) =$$

$$= \sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y^{(i)} - h(x^{(i)}; w))^2}{2\sigma^2} \right) \right)$$

$$= \underbrace{-N \ln(\sqrt{2\pi}\sigma)}_{\text{konst.}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - h(x^{(i)}; w))^2$$



MAKSIMIZACIJA LOG-IZGLEDOVSTI

EKVIVALENTNA JE MINIMIZACIJA KVADRATNE
POGREŠKE

$$b) \quad H = \ln(x^{(i)}, w)$$

$$p(y|w) = \begin{cases} 1, & \text{ako } y=1 \\ 1-w, & \text{inace} \end{cases} = \underbrace{w^y (1-w)^{1-y}}_{\text{BERNOULLI}}$$

$$\begin{aligned} \ln \mathcal{L}(w|D) &= \ln p(D|w) = \ln \prod_{i=1}^N p(y^{(i)}|x^{(i)}) = \\ &= \ln \prod_{i=1}^N \ln(x^{(i)}, w)^{y^{(i)}} (1-\ln(x^{(i)}, w))^{1-y^{(i)}} \\ &= \sum_{i=1}^N \left(y^{(i)} \ln \ln(x^{(i)}, w) + (1-y^{(i)}) \ln (1-\ln(x^{(i)}, w)) \right), \end{aligned}$$

MAXIMIZACIJA LOG PREGLEDNOSTI EKUIVALENNTA JE MINIMIZACIJI POGRESKE UNAKRSNE ENTROPIJE.

$$c) \quad p(w) = N(0, \lambda^{-1} I)$$

MAP:

$$\begin{aligned} \max(p(D|w) \cdot p(w)) &= \max(\mathcal{L}(w|D) - p(w)) = \\ &= \max \ln (\mathcal{L}(w|D) p(w)), \end{aligned}$$

$$\begin{aligned} \ln(\mathcal{L}(w|D) p(w)) &= \ln \prod_{i=1}^N p(y^{(i)}|x^{(i)}) \cdot N(0, \lambda^{-1} I) = \\ &= \ln \left[\prod_{i=1}^N \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(y^{(i)} - \ln(x^{(i)}, w))^2}{2\sigma^2} \right) \right) \frac{1}{\sqrt{2\pi\lambda^{-1}}} \exp \left(-\frac{w^T w}{2\lambda} \right) \right] \\ &= \underbrace{-N \ln \sqrt{2\pi} \sigma}_{\text{konst.}} - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - \ln(x^{(i)}, w))^2 - \underbrace{\ln \sqrt{2\pi\lambda^{-1}} I}_{\text{konst.}} - \\ &\quad - \frac{w^T w}{2\lambda} \end{aligned}$$

$$\begin{aligned} \hat{\epsilon} &= -\frac{1}{\sigma^2} \sum_{i=1}^N (y^{(i)} - h(x^{(i)}, w))^2 + \lambda w^T w \\ &\equiv -\sum_{i=1}^N (y^{(i)} - h(x^{(i)}, w))^2 + \lambda w^T w // \quad \lambda = \lambda \sigma^2 \end{aligned}$$

IZRAT ZA L2 REGULARIZACIJU

MAP PROCJENITELJ VE APRIORNU VJEZOJAMOST RAZDIOBE TEŽINA w : $p(w) = N(0, \lambda^{-1} I)$
EKVIVALENTAN JE MINIMIZACIJI L2 REGULARIZIRANE POGREŠKE.

d) DA, ZATO ŠTO JE GAUSSOVA DISTRIBUCIJA SAMOKONJUGATNA, LOGARITM/RANJE JAKO OLAKŠAVA PROBLEM PA MOŽEMO ANALITIČKI PRORAČUNATI MAP PROCJENU (DOBUEMO PROBLEM EKVIVALENTAN L2 REGULARIZACIJI), T.J. IMAMO RJEŠENJE U ZATVORENOJ FORMI.

NE, JER BUDUĆI DA GAUSSOVA DISTRIBUCIJA NIJE KONJUGATNA S BERNOULLIJEVOM, RJEŠENJE MAP PROCJENE NEĆE BITI JEDNOSTAVNO PRORAČUNATI (NEĆE IMATI ZATVORENU FORMU).