14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.

Jan Šnajder, vježbe, v2.2

1 Zadatci za učenje

1. [***Svrha: Razumjeti kako podatci određuju izglednost parametara putem funkcije izglednosti.***]

   (a) Definirajte funkciju izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. Na kojoj se pretpostavci o skupu $\mathcal{D}$ temelji ta definicija?

   (b) Raspolažemo skupom (neoznačenih) primjera $\mathcal{D} = \{x^{(i)}\}_i = \{-2, -1, 1, 3, 5, 7\}$. Pretpostavljamo da se primjeri pokoravaju Gaussovoj distribuciji, $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$. Napišite funkciju izglednosti $\mathcal{L}(\mu, \sigma^2|\mathcal{D})$. Koliko iznosi izglednost parametara $\mu = 0$ i $\sigma^2 = 1$, a koliko vjerojatnost uzorka $\mathcal{D}$ uz te parametre?

   (c) Novčić bacamo N puta, pri čemu smo m puta dobili glavu, a $N - m$ puta pismo. Ishodi bacanja novčića sačinavaju naš uzorak $\mathcal{D}$. Napišite izraz za funkciju izglednosti parametra μ Bernoullijeve distribucije, parametrizirane s N i m, tj. $\mathcal{L}(\mu|N, m)$.

   (d) Skicirajte funkciju izglednosti za slučaj $N = 10$ i $m = 1$. Koja je vrijednost parametra μ najizglednija? Uz koju je vrijednost μ skup $\mathcal{D}$ najvjerojatniji?

2. [***Svrha: Osvježiti znanje matematike potrebno za izvođenje MLE-procjenitelja dviju osnovnih univarijatnih razdioba.***]

   (a) Definirajte MLE-procjenitelj $\hat{\boldsymbol{\theta}}_{\text{ML}}$.

   (b) Izvedite MLE-procjenitelj $\hat{\mu}_{\text{ML}}$ za parametar μ Bernoullijeve razdiobe $P(x|\mu)$.

   (c*) Izvedite MLE-procjenitelj $\hat{\mu}_{k,\text{MLE}}$ za parametar μ_k kategorijske ("multinulijeve") razdiobe $P(\mathbf{x}|\boldsymbol{\mu})$. Ovdje je kod optimizacije potrebno osigurati da vrijedi ograničenje $\sum_{k=1}^{K} \mu_k = 1$; za to upotrijebite metodu Lagrangeovih multiplikatora.

   (c) Izvedite MLE-procjenitelje $\hat{\mu}_{\text{ML}}$ i $\hat{\sigma}^2$ za parametre μ odnosno σ^2 univarijatne Gaussove razdiobe $p(x|\mu, \sigma^2)$.

3. [***Svrha: Isprobati izračun pristranost procjenitelja i shvatiti da MLE-procjenitelj može biti pristran, tj. da najveća izglednost ne jamči nepristranost.***]

   (a) Dokažite da je $\hat{\mu}_{\text{ML}}$ nepristran, a $\hat{\sigma}^2_{\text{ML}}$ pristran. Koliko iznosi pristranost $b(\hat{\sigma}^2)$?

   (b) Je li ta pristranost u praksi problematična? Obrazložite.

4. [***Svrha: Izvježbati izračun procjene parametara multivarijatne Gaussove razdiobe (v. primjer 3.5 u skripti). Uočiti da multikolinearnost značajki dovodi do problema.***] Raspolažemo uzorkom $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{6}$ za koji pretpostavljamo da potječe iz multivarijatne normalne razdiobe:

$$\begin{aligned} \mathbf{x}^{(1)} &= (9.5, -0.7, -2.8) & \mathbf{x}^{(4)} &= (2.3, 0.3, 1.2) \\ \mathbf{x}^{(2)} &= (8.8, -0.8, -3.2) & \mathbf{x}^{(5)} &= (2.2, \quad 0, \quad 0) \\ \mathbf{x}^{(3)} &= (6.5, -0.2, -0.8) & \mathbf{x}^{(6)} &= (3.6, 0.3, 1.2) \end{aligned}$$

   (a) Izračunajte MLE-procjenu vektora srednje vrijednosti i MLE-procjenu kovarijacijske matrice.

   (b) Izračunajte gustoću vjerojatnosti za primjer $\mathbf{x} = (-2, 1, 0)$. Je li ta gustoća dobro definirana? Zašto?

(c) Matrica kovarijacije Σ mora biti pozitivno definitna a da bi imala pozitivnu determinantu i inverz. Multikolinearnost značajki jedan je od mogućih razloga zašto matrica nije pozitivno definitna. Izračunajte Pearsonov koeficijent korelacije ρ između svih parova varijabli te izbacite varijablu koja je najviše korelirana s nekom drugom varijablom. Zatim u tako smanjenome ulaznom prostoru pokušajte ponovno izračunati funkciju gustoće za primjer $\mathbf{x}$.

5. [***Svrha: Razumjeti MAP-procjenitelj i način njegovog izračuna za Bernoullijevu distribuciju (beta-Bernoullijev model). Uočiti kako svojstvo konjugatnosti olakšava izračun aposteriorne distribucije.***]

   (a) Definirajte MAP-procjenitelj $\hat{\boldsymbol{\theta}}_{\text{MAP}}$ i objasnite zašto je on bolji od MLE-procjenitelja $\hat{\boldsymbol{\theta}}_{\text{ML}}$.

   (b) Objasnite što je to (1) konjugatna distribuicija i (2) konjugatna apriorna distribucija. Zašto nam je svojstvo konjugatnosti bitno?

   (c) Apriornu distribuciju parametra μ Bernullijeve distribucije modeliramo beta-distribucijom $p(\mu|\alpha,\beta)$. Beta-distribucija konjugatna je apriorna distribucija za Bernoullijevu funkciju izglednosti $\mathcal{L}(\mu|N,m)$. Skicirajte beta-distribuciju za (1) $\alpha=\beta=1$, (2) $\alpha=\beta=2$, (3) $\alpha=2$, $\beta=4$ i (4) $\alpha=4$, $\beta=2$.

   (d) Izvedite izraz za aposteriornu distribuciju parametra, $p(\mu|N,m,\alpha,\beta)$.

   (e) Recimo da vjerujemo da je novčić pravedan, ali da u to nismo baš u potpunosti uvjereni. To možemo modelirati beta-distribucijom $p(\mu|\alpha=2,\beta=2)$. Zatim smo u $N=10$ bacanja novčića samo $m=1$ puta dobili glavu. Skicirajte apriornu gustoću $p(\mu|\alpha=2,\beta=2)$, funkciju izglednosti $\mathcal{L}(\mu|N=10,m=1)$ te njihov umnožak. Iskoristite činjenicu da je maksimizator (mod) beta-distribucije jednak $\frac{\alpha-1}{\alpha+\beta-2}$.

   (f) Izračunajte $\hat{\mu}_{\text{MAP}}$ i $\hat{\mu}_{\text{ML}}$ te komentirajte razliku. Kako bi porast broja primjera N utjecao na ovu razliku?

   (g) Pokažite da se MAP-procjenitelj za parametar μ Bernoullijeve distribucije svodi na Laplaceov procjenitelj, ako se apriorna distribucija parametra modelira beta-distribucijom te ako se odaberu odgovarajući (koji?) parametri α i β.

6. [***Svrha: Razumjeti MAP-procjenitelj i način njegovog izračuna za kategorijsku (multinulijevu) varijablu (Dirichlet-kategorijski model).***]

   (a) Definirajte Dirichletovu distribuciju.

   (b) Definirajte Dirichlet-kategorijski model i izvedite MAP procjenitelj za $\alpha_k=2$.

7. [***Svrha: Razumjeti vezu između probabilističkih modela i poopćenih linearnih modela preko veze između MLE-procjenitelja i minimizacije empirijske pogreške. Razumjeti vezu između MAP-procjenitelja i minimizacije L2-regularizirane empirijske pogreške.***]

   (a) Pokažite da je MLE-procjena za parametre $\mathbf{w}$ kod linearne regresije (uz pretpostavku normalno distribuiranog šuma) ekvivalentna postupku najmanjih kvadrata.

   (b) Pokažite da je MLE-procjena za parametre $\mathbf{w}$ kod logističke regresije (uz pretpostavku Bernoullijeve distribucije oznaka) ekvivalentna minimizacije pogreške unakrsne entropije.

   (c*) Gornja dva zadataka demonstriraju vezu između MLE-procjenitelja i minimizacije empirijske pogreške. Postoji analogna veza između MAP-procjenitelja i minimizacije L2-regularizirane empirijske pogreške. Razmotrimo konkretno linearnu regresiju. Ako se apriorna gustoća vjerojatnosti težina $\mathbf{w}$ definira kao:

$$p(\mathbf{w}) = \mathcal{N}(0, \alpha^{-1}\mathbf{I})$$

   tj. kao multivarijatna normalna razdioba sa središtem u ishodištu prostora parametara i s izotropnom kovarijacijskom matricom pomnoženom nekim hiperparametrom α^{-1}, onda je MAP-procjenitelj ekvivalentan L2-regulariziranoj kvadratnoj pogrešci. Dokažite to. (Pomoć: slajdovi 30–31 ovdje i poglavlje 3.3.1 u PRML.)

   (d*) Je li u prethodnom zadatku bilo ključno to što je Gaussova distribucija samokonjugatna? Možemo li isti princip primijeniti i kod modela gdje izglednost nije Gaussova, npr. kod logističke regresije (i drugih poopćenih linearnih modela)? Zašto?

2 Zadatci s ispita

1. (T) Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ nije isto što i vjerojatnost. **Po čemu se izglednost razlikuje od vjerojatnosti?**

   A Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\mathcal{D}|\boldsymbol{\theta})$, samo što je izglednost funkcija parametara $\boldsymbol{\theta}$, dok je $p(\mathcal{D}|\boldsymbol{\theta})$ funkcija uzorka $\mathcal{D}$

   B Funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ jednaka je gustoći vjerojatnosti $p(\boldsymbol{\theta}|\mathcal{D})$, ali, za razliku od gustoće vjerojatnosti, nije odozgo ograničena sa 1

   C Ako su podatci diskretni (kategoričke značajke), onda je funkcija izglednosti parametara $\boldsymbol{\theta}$ isto što i zajednička vjerojatnost uzorka $\mathcal{D}$ i parametara $\boldsymbol{\theta}$

   D Za razliku od vjerojatnosti, funkcija izglednosti $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ je simetrična, u smislu da vrijedi $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\theta})$

2. (N) Raspolažemo sljedećim skupom označenih primjera:

   $$\mathcal{D} = \{x^{(i)}, y^{(i)}\} = \{(-2,1),(-2,1),(-1,0),(0,0),(1,1),(3,1)\}$$

   Na ovom skupu treniramo univarijatni Bayesov klasifikator, za što trebamo procijeniti izglednosti klasa $p(x|y)$. Te su izglednosti definirane Gaussovom gustoćom vjerojatnosti:

   $$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}$$

   Parametre μ i σ^2 gustoće vjerojatnosti $p(x|y)$ procjenjujemo MLE-om. Neka su μ_1 i σ_1^2 parametri gustoće vjerojatnosti $p(x|y=1)$ dobiveni MLE-om na podskupu primjera $\mathcal{D}_{y=1}$. **Koliko iznosi log-izglednost $\mathcal{L}(\mu_1,\sigma_1^2|\mathcal{D}_{y=1})$?**

   A -22.60 B -8.68 C -8.76 D $+0.48$

3. (P) Neka je $\mathcal{L}(\mu,\sigma^2|\mathcal{D})$ log-izglednost parametara μ i σ^2 normalne distribucije izračunata nad uzorkom $\mathcal{D}$ koji sadrži ukupno N opažanja normalne varijable x. Nadalje, neka su $(\mu_{\text{MLE}}, \sigma^2_{\text{MLE}})$ parametri distribucije procijenjeni MLE-om nad uzorkom $\mathcal{D}$, te neka je σ^2_{UB} nepristrana procjena varijance, izračunata kao $\sigma^2_{\text{UB}} = \frac{N}{N-1}\sigma^2_{\text{MLE}}$. Konačno, neka je $\mathcal{D}'$ slučajno uzorkovan podskup uzorka $\mathcal{D}$, tj. $\mathcal{D}' \subset \mathcal{D}$, pri čemu je poduzorkovanje načinjeno nakon procjene parametara. Razmotrite sljedeće četiri vrijednosti funkcije log-izglednosti $\mathcal{L}(\mu,\sigma^2|\mathcal{D})$:

   $$\begin{aligned}\mathcal{L}_0 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{MLE}}|\mathcal{D})\\ \mathcal{L}_1 &= \mathcal{L}(0,1|\mathcal{D})\\ \mathcal{L}_2 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{UB}}|\mathcal{D})\\ \mathcal{L}_3 &= \mathcal{L}(\mu_{\text{MLE}}, \sigma^2_{\text{UB}}|\mathcal{D}')\end{aligned}$$

   Što možemo zaključiti o odnosima između ovih vrijednosti funkcije log-izglednosti?

   A $\mathcal{L}_0 > \mathcal{L}_1$, $\mathcal{L}_2 \geqslant \mathcal{L}_3$

   B $\mathcal{L}_1 \geqslant \mathcal{L}_0$, $\mathcal{L}_2 \geqslant \mathcal{L}_3$

   C $\mathcal{L}_0 < \mathcal{L}_3$, $\mathcal{L}_0 \leqslant \mathcal{L}_2$

   D $\mathcal{L}_0 \geqslant \mathcal{L}_1$, $\mathcal{L}_0 > \mathcal{L}_2 > \mathcal{L}_3$

4. (T) MAP-procjenitelj definiramo kao $\hat{\boldsymbol{\theta}}_{\text{MAP}} = \text{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Pri odabiru apriorne distribucije $p(\boldsymbol{\theta})$, nastojimo da je to neka standardna teorijska distribucija i da je konjugatna distribucija

za izglednost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$. **Što to znači i zašto to želimo?**

A To znači da će umnožak izglednosti i apriorne distribucije dati distribuciju koja je iste vrste kao i apriorna distribucija, a ako je riječ o standardnoj teorijskoj distribuciji iz eksponencijalne familije, njezin mod (maksimizator) postoji u zatvorenoj formi, što nam omogućava da procjenitelj izračunamo analitički

B To znači da je apriorna distribucija ista vrsta distribucije kao i vjerojatnost podataka uz dane parametre, tj. izglednost parametara, pa će njihov umnožak biti distribucija koja je proporcionalna aposteriornoj distribuciji i čiji ćemo maksimum moći izračunati Bayesovim pravilom

C To znači da je apriorna distribucija upravljana hiperparametrima kojima možemo ugoditi distribucija parametara koji procjenjujemo, tj. parametri apriorne distribucije i parametri izglednosti su identični, što nam omogućava da te dvije distribucije pomnožimo i zatim nađemo maksimizator

D To znači da je aposteriorna distribucija parametara ista kao izglednost parametara, pa primjenom Bayesovog pravila možemo izračunati apriornu vjerojatnost parametara te, nakon zanemarivanja nazivnika koji je za fiksiran skup podataka konstantan, pronaći parametre koji maksimiziraju aposterionu vjerojatnost

5. (T) Kod MAP-procjenitelja, apriorna distribucija parametra $p(\boldsymbol{\theta})$ tipično se odabire tako da bude konjugatna za funkciju izglednosti $p(\mathcal{D}|\boldsymbol{\theta})$. Pretpostavimo da MAP-procjenitelj izračunavamo heurističkom metodom (npr., gradijentnim usponom). Što se događa ako za apriornu distribuciju parametra upotrijebimo distribuciju koja *nije* konjugatna funkciji izglednosti?

   A Zajedničku distribuciju $p(\mathcal{D}, \boldsymbol{\theta})$ ne možemo izvesti u zatvorenoj formi, pa MAP nije definiran

   B Aposteriornu distribuciju $p(\boldsymbol{\theta}|\mathcal{D})$ ne možemo izvesti u zatvorenoj formi, ali MAP možemo izračunati heurističkom optimizacijom

   C Ako je apriorna distribucija $p(\boldsymbol{\theta})$ iz eksponencijalne familije, onda je aposteriona distribucija $p(\boldsymbol{\theta}|\mathcal{D})$ u zatvorenoj formi i MAP je izračunljiv

   D Neovisno o apriornoj distribuciji parametra $p(\boldsymbol{\theta})$, MAP je izračunljiv optimizacijom drugog reda (npr., Newtonovim postupkom)

6. (N) U beta-Bernoullijevom modelu, apriornu vjerojatnost parametra μ modeliramo beta-distribucijom, čija je gustoća vjerojatnosti definirana kao:

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

   Mod (maksimizator) te distribucije jest:

$$\mu^* = \frac{\alpha - 1}{\alpha + \beta - 2}$$

   Aposteriorna distribucija parametra definirana je kao:

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

   Neka $\alpha = \beta = 2$. Računamo MAP-procjenu za parametar μ Bernoulijeve varijable. To radimo na dva uzorka, $\mathcal{D}_1 = (N_1, m_1)$ i $\mathcal{D}_2 = (N_2, m_2)$, koji nam pristižu jedan za drugim. Pritom koristimo svojstvo konjugatnosti, na način da aposteriornu gustoću vjerojatnosti izračunatu na temelju prvog uzorka koristimo kao apriornu gustoću vjerojatnosti pri procjeni na temelju drugog uzorka. U prvom uzorku, veličine $N_1 = 50$, Bernoullijeva varijabla realizirana je s vrijednošću $y = 1$ ukupno $m_1 = 42$ puta. U drugom uzorku, veličine $N_2 = 15$, Bernoullijeva varijabla realizirana je s vrijednošću $y = 1$ ukupno $m_2 = 3$ puta. Izračunajte MAP-procjene za parametar μ na temelju ova dva uzorka. **Koliko iznosi promjena u procjeni za μ između prve i druge procjene?**

   A -0.59 B $+0.45$ C -0.14 D -0.64

7. (P) Koristimo MAP-procjenitelj kako bismo procijenili parametre distribucije kategoričke (multinulijeve) varijable X. Varijabla može poprimiti tri vrijednosti, x_1, x_2 i x_3, pa dakle trebamo procijeniti vektor parametara (μ_1, μ_2, μ_3). Budući da se ovdje radi o kategoričkoj varijabli, za MAP-procjenu koristimo Dirichlet-kategorički model. Na temelju stručnog znanja o problemu koji rješavamo, u procjenu smo ugradili naše pretpostavke. To znači da smo na prikladan način definirali Dirichletovu apriornu gustoću vjerojatnosti, $p(\mu_1, \mu_2, \mu_3 | \alpha_1, \alpha_2, \alpha_3)$, gdje je $(\alpha_1, \alpha_2, \alpha_3)$ vektor hiperparametara (parametri Dirichletove distribucije). Konkretno, te smo hiperparametre definirali kao $(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 1)$. Međutim, skup podataka $\mathcal{D}$ ne odgovara našoj pretpostavci. U tom skupu, varijabla X je u pola slučajeva realizirana s vrijednošću x_2, u pola slučajeva s vrijednošću x_3, no baš niti jednom s vrijednošću x_1. **Kakva će biti naša MAP-procjena parametara** (μ_1, μ_2, μ_3)?

A $\mu_1 = 0$, $\frac{1}{2} < \mu_2 < 1$, $0 < \mu_2 < \mu_3 < 1$

B $0 < \mu_1 < \frac{1}{3}$, $\frac{1}{2} < \mu_2 < 1$, $\mu_3 = 0$

C $0 < \mu_1 < \mu_3 < 1$, $\frac{1}{3} < \mu_2 < \frac{2}{3}$

D $0 < \mu_1 < \frac{1}{3}$, $\frac{1}{3} < \mu_2 < 1$, $0 < \mu_3 < \mu_2 < 1$

8. (P) Bacanje igraće kocke modeliramo kategoričkom varijablom $\mathbf{x}$, gdje indikatorske varijable $x_1, \ldots, x_6$ odgovaraju vrijednosti koju dobivamo bacanjem kocke. Za procjenu parametara $\boldsymbol{\mu}$ kategoričke distribucije koristimo MAP-procjenitelj s Dirichletovom distribucijom za apriornu gustoću vjerojatnosti. U stvarnosti, kocka je modificirana tako da će nešto češće davati šesticu, odnosno realizaciju $x_6 = 1$, međutim mi to ne znamo. Naprotiv, na temelju manjeg broja opažanja ranijih bacanja kocke utvrdili smo da je kocka najčešće davala peticu, no svjesni smo da je naša procjena temeljena na manjem broju opažanja. **Uz koje parametre Dirichletove distribucije će naša procjena za** $\boldsymbol{\mu}$ **biti najbliža stvarnoj vrijednosti tih parametara?**

A $\boldsymbol{\alpha} = (1, 1, 1, 1, 1, 1)$

B $\boldsymbol{\alpha} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$

C $\boldsymbol{\alpha} = (2, 2, 2, 2, 2, 2)$

D $\boldsymbol{\alpha} = (1, 1, 1, 1, 3, 1)$