Parametarska estimacija gustine raspodele verovatnoće

- Uvod
- Maksimalna izglednost
  - Osnovni pojmovi
  - Primeri
- Kvalitet procene - pristrasnost i varijansa
- Bayesova estimacija

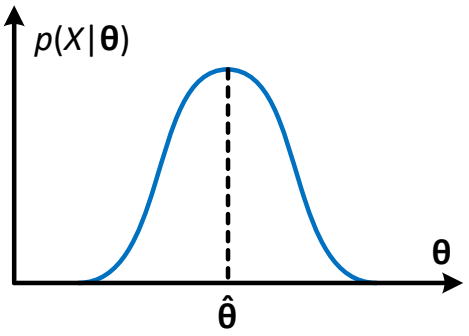
Uvod

- Prethodna predavanja su se bavila klasifikacijom (određivanjem regiona odlučivanja) uz pretpostavku da je gustina raspodele verovatnoće poznata
  - Bayesova teorija odlučivanja je formalno definisala problem
  - Kvadratni klasifikatori su rešenje za slučaj klasa sa normalnom raspodelom
- Najčešće ne poznajemo pravu gustinu raspodele verovatnoće već se ona mora proceniti na osnovu eksperimentalnih podataka
  - Parametarska estimacija
  - Neparametarska estimacija
- **Parametarska estimacija** gustine raspodele verovatnoće
  - Pretpostavlja se određeni oblik raspodele (npr. normalna raspodela), pa se problem svodi na određivanje parametara
    - Estimacija na osnovu maksimalne izglednosti
    - Bayesova estimacija
- **Neparametarska estimacija** gustine raspodele verovatnoće
  - Estimacija gustine verovatnoće pomoću kernela
  - Metoda najbližeg suseda

Maksimalna izglednost i Bayesova estimacija

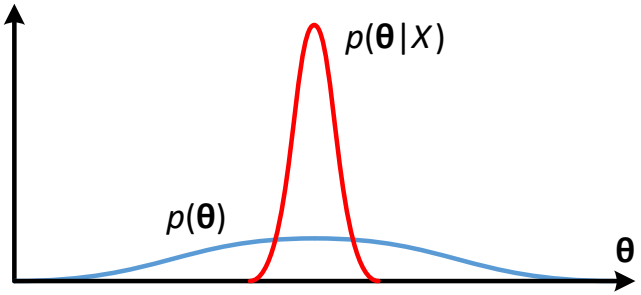
- Maksimalna izglednost (*maximum likelihood* – ML)
  - Pretpostavka je da su parametri *fiksni*, ali nepoznati
  - ML usvaja vrednosti parametara koje se najbolje „slažu“ sa skupom uzoraka X nepoznate raspodele

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})$$



- Bayesova estimacija
  - Pretpostavka je da su parametri *slučajne promenljive* sa poznatom apriornom raspodelom
  - Estimira se aposteriorna raspodela $p(\boldsymbol{\theta}|X)$ (raspodela parametara $\boldsymbol{\theta}$ ako je dat skup uzoraka X nepoznate raspodele)
  - Gustina raspodele verovatnoće u tački $\mathbf{x}$ dobija se zatim integracijom po celom prostoru parametara:

$$p(\mathbf{x}|X) = \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|X) \mathrm{d}\boldsymbol{\theta}$$



Maksimalna izglednost

- Neka se estimira gustina raspodele verovatnoće $p(\mathbf{x})$ koja zavisi od određenog broja parametara $\boldsymbol{\theta} = [\vartheta_1 \ \vartheta_2 \ \ldots \ \vartheta_p]^T$
  - Za 1-D Gaussovu raspodelu $\vartheta_1 = \mu$, $\vartheta_2 = \sigma^2$ i $p(x) = \mathcal{N}(\mu, \sigma^2)$
  - Notacija $p(\mathbf{x}|\boldsymbol{\theta})$ naglašava eksplicitnu zavisnost raspodele od parametara $\boldsymbol{\theta}$
  - Pri estimaciji $\boldsymbol{\theta}$ za više klasa pretpostavka je da su parametri različitih klasa međusobno nezavisni
- Neka se parametri $\boldsymbol{\theta}$ ocenjuju na osnovu skupa od N međusobno nezavisnih uzoraka $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ izvučenih iz raspodele $p(\mathbf{x}|\boldsymbol{\theta})$ na slučajan način:

$$p(X|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

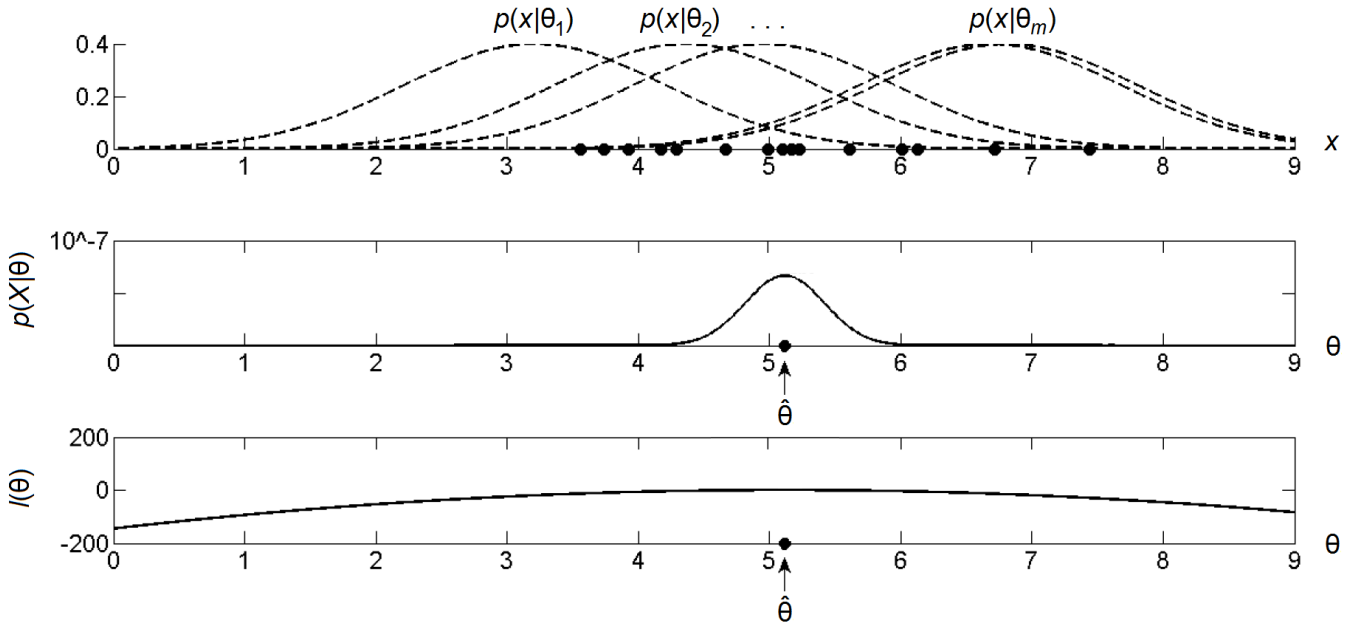
  - ML estimacija parametara $\boldsymbol{\theta}$ je ona vrednost koja maksimizuje izglednost $p(X|\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta})$$

  - Ovo odgovara intuitivno utemeljenoj ideji da se za vektor parametara $\boldsymbol{\theta}$ bira ona vrednost koja se najbolje slaže sa uzorcima dobijenim slučajnim izvlačenjem iz nepoznate raspodele

Maksimalna izglednost

- Isti uzorci za različite pretpostavljene vrednosti **θ** različito su izgledni:

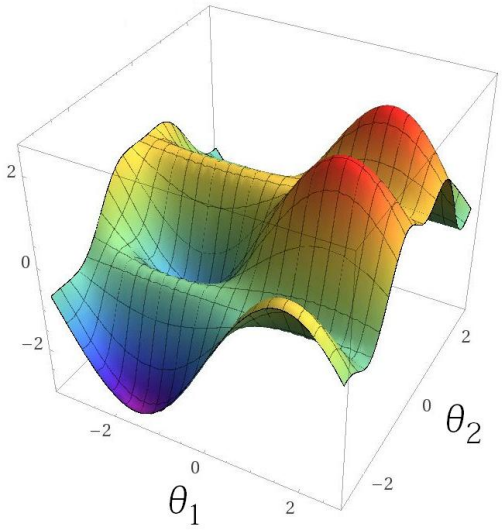


- Iz praktičnih razloga često se umesto same izglednosti maksimizuje njen logaritam $l(\boldsymbol{\theta}) = \ln p(X|\boldsymbol{\theta})$, čime se smisao maksimizacije ne menja:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(X|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \ln p(X|\boldsymbol{\theta})$$
$$= \arg\max_{\boldsymbol{\theta}} \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maksimalna izglednost

- Ako je $p(X|\boldsymbol{\theta})$ diferencijabilna, $\hat{\boldsymbol{\theta}}$ se može naći pomoću diferencijalnog računa, izjednačavanjem izvoda $l(\boldsymbol{\theta}) = \ln p(X|\boldsymbol{\theta})$ po svakoj komponenti $\boldsymbol{\theta}$ sa 0
  - Na ovaj način dobija se lokalni minimum, lokalni maksimum ili (retko) prevojna tačka $l(\boldsymbol{\theta})$
- *Gradijent* u prostoru parametara definiše se kao:



$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \partial/\partial\vartheta_1 \\ \partial/\partial\vartheta_2 \\ \vdots \\ \partial/\partial\vartheta_p \end{bmatrix}$$

tako da se maksimizacija log-izglednosti $l(\boldsymbol{\theta})$ u opštem slučaju svodi na:

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) = \mathbf{0}$$

- Maksimum funkcije više promenljivih može se naći i kroz iterativnu proceduru ekvivalentnu metodi gradijentnog silaska (samo treba ići *u pravcu* gradijenta)

Primer: 1-D Gaussova raspodela, nepoznato μ

- Za 1-D Gaussovu raspodelu sa nepoznatom srednjom vrednošću μ i poznatom varijansom σ^2, potrebno je odrediti optimalnu ML procenu μ na osnovu skupa uzoraka $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$
- Za jedan uzorak $x^{(i)}$ važi:

$$p(x^{(i)}|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}}$$

$$\ln p(x^{(i)}|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x^{(i)}-\mu)^2$$

pa se ML procena na čitavom skupu od N uzoraka dobija na osnovu:

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

$$\frac{\mathrm{d}l(\mu)}{\mathrm{d}\mu} = \frac{\mathrm{d}}{\mathrm{d}\mu}\sum_{i=1}^{N}\ln p(x^{(i)}|\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x^{(i)}-\mu) = 0$$

odakle se dobija:

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

Primer: 1-D Gaussova raspodela, nepoznato μ i σ^2

- Za 1-D Gaussovu raspodelu sa nepoznatom srednjom vrednošću μ i poznatom varijansom σ^2, potrebno je odrediti optimalnu ML procenu μ na osnovu skupa uzoraka $X = \{x^{(1)}, x^{(2)},..., x^{(N)}\}$
  - Elementi vektora parametara $\boldsymbol{\theta}$ su $\vartheta_1 = \mu$ i $\vartheta_2 = \sigma^2$
- Za jedan uzorak $x^{(i)}$ važi:

$$p(x^{(i)} \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sqrt{\vartheta_2}} e^{-\frac{(x^{(i)}-\vartheta_1)^2}{2\vartheta_2}}$$

$$\ln p(x^{(i)} \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi\vartheta_2) - \frac{(x^{(i)}-\vartheta_1)^2}{2\vartheta_2}$$

pa se ML procena na čitavom skupu od N uzoraka dobija na sledeći način:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x^{(i)} \mid \boldsymbol{\theta}) = -\frac{N}{2}\ln(2\pi\vartheta_2) - \frac{1}{2\vartheta_2}\sum_{i=1}^{N}(x^{(i)}-\vartheta_1)^2$$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0 \Rightarrow \begin{cases} \dfrac{\partial l(\boldsymbol{\theta})}{\partial \vartheta_1} = \displaystyle\sum_{i=1}^{N} \frac{1}{\vartheta_2}(x^{(i)}-\vartheta_1) = 0 \\ \dfrac{\partial l(\boldsymbol{\theta})}{\partial \vartheta_2} = -\dfrac{N}{2\vartheta_2} + \displaystyle\sum_{k=1}^{N} \frac{(x^{(i)}-\vartheta_1)^2}{2\vartheta_2^2} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x^{(i)} \\ \hat{\sigma}^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x^{(i)}-\hat{\mu})^2 \end{cases}$$

Primer: Gaussova raspodela, nepoznato $\boldsymbol{\mu}$

- Za d-dimenzionu Gaussovu raspodelu sa nepoznatom srednjom vrednošću $\boldsymbol{\mu}$ i poznatom kovarijansnom matricom $\boldsymbol{\Sigma}$, potrebno je odrediti optimalnu ML procenu $\boldsymbol{\mu}$ na osnovu skupa uzoraka $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$
- Za jedan uzorak $\mathbf{x}^{(i)}$ važi:

$$p(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}^{(i)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})}$$

$$\ln p(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^{d}|\boldsymbol{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}^{(i)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})$$

pa se ML procena na čitavom skupu od N uzoraka dobija na osnovu:

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{N} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = -\frac{N}{2}\ln\left[(2\pi)^{d}|\boldsymbol{\Sigma}|\right] - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})$$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\mu}) = \sum_{i=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu}) = 0$$

odakle, nakon množenja obe strane sa $\boldsymbol{\Sigma}$: $\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}$

- Procena na osnovu *uzoračke srednje vrednosti* u osnovi je procena na osnovu maksimizacije izglednosti

Diferenciranje matrica i vektora

$$\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} \Rightarrow \frac{d\alpha}{d\mathbf{x}} = \mathbf{y}^{\mathrm{T}}\mathbf{A}$$

$$\alpha = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \Rightarrow \frac{d\alpha}{d\mathbf{x}} = \mathbf{x}^{\mathrm{T}}(\mathbf{A}+\mathbf{A}^{\mathrm{T}})$$

Za simetrične matrice:

$$\alpha = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \Rightarrow \frac{d\alpha}{d\mathbf{x}} = 2\mathbf{x}^{\mathrm{T}}\mathbf{A}$$

(sve pod pretpostavkom da su **A** i **y** nezavisni od **x**)

Primer: Gaussova raspodela, nepoznato $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$

- Za *d*-dimenzionu Gaussovu raspodelu sa nepoznatom srednjom vrednošću $\boldsymbol{\mu}$ i nepoznatom kovarijansnom matricom $\boldsymbol{\Sigma}$, potrebno je odrediti optimalnu ML procenu $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ na osnovu skupa uzoraka $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$
  - Elementi vektora parametara $\boldsymbol{\theta}$ su elementi vektora $\boldsymbol{\mu}$ i matrice $\boldsymbol{\Sigma}$

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}) = \mathbf{0} \Rightarrow \begin{cases} \hat{\boldsymbol{\mu}} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} \mathbf{x}^{(i)} \\ \hat{\boldsymbol{\Sigma}} = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \end{cases}$$

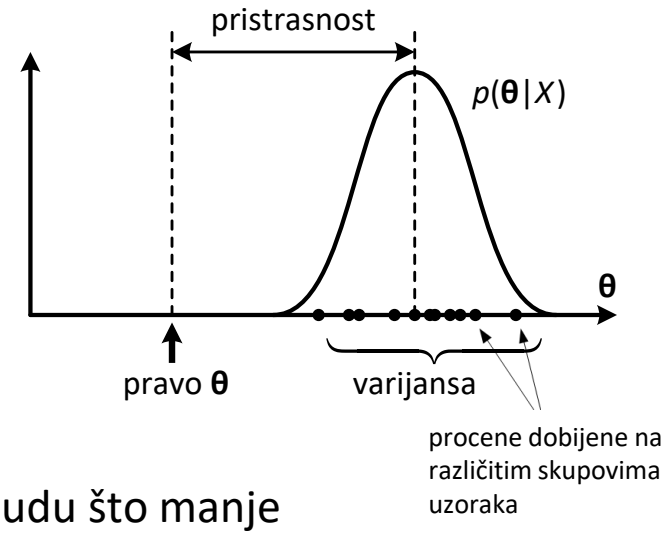
  - Procena srednje vrednosti je ponovo uzoračka srednja vrednost, a procena kovarijansne matrice je aritmetička sredina matrica oblika $(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$
    - Ovo je logično jer je prava kovarijansna matrica očekivana vrednost $(\mathbf{x} - \hat{\boldsymbol{\mu}})(\mathbf{x} - \hat{\boldsymbol{\mu}})^{\mathrm{T}}$

Pristrasnost i varijansa procene

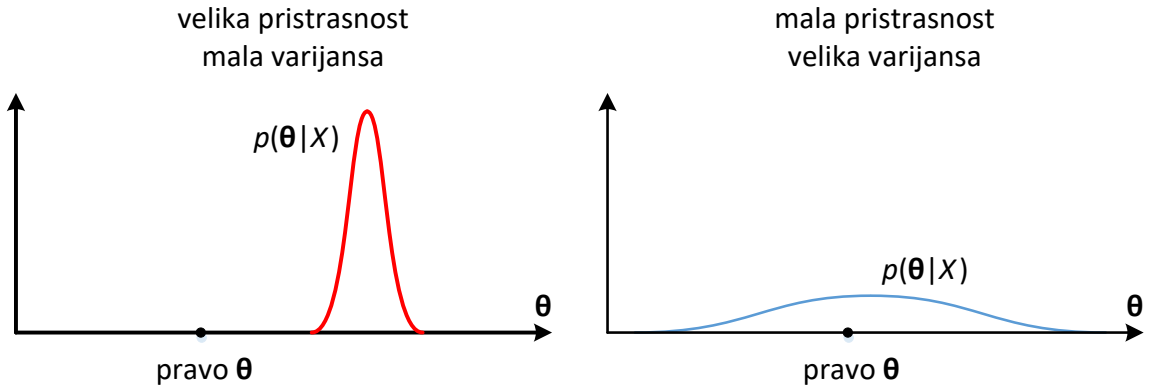
- Postavlja se pitanje kako definisati kvalitet neke procene
  - *Pristrasnost* (*necentriranost*) procene govori o tome koliko procena u proseku odstupa od tačne vrednosti na različitim skupovima uzoraka

$$\text{bias}(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta$$

  - *Varijansa* procene govori o tome koliko se procena menja na različitim skupovima uzoraka



- U praksi treba postići da pristrasnost i varijansa budu što manje
  - Obično je nemoguće ispuniti oba uslova i obično se jedna od te dve veličine minimizuje na račun druge

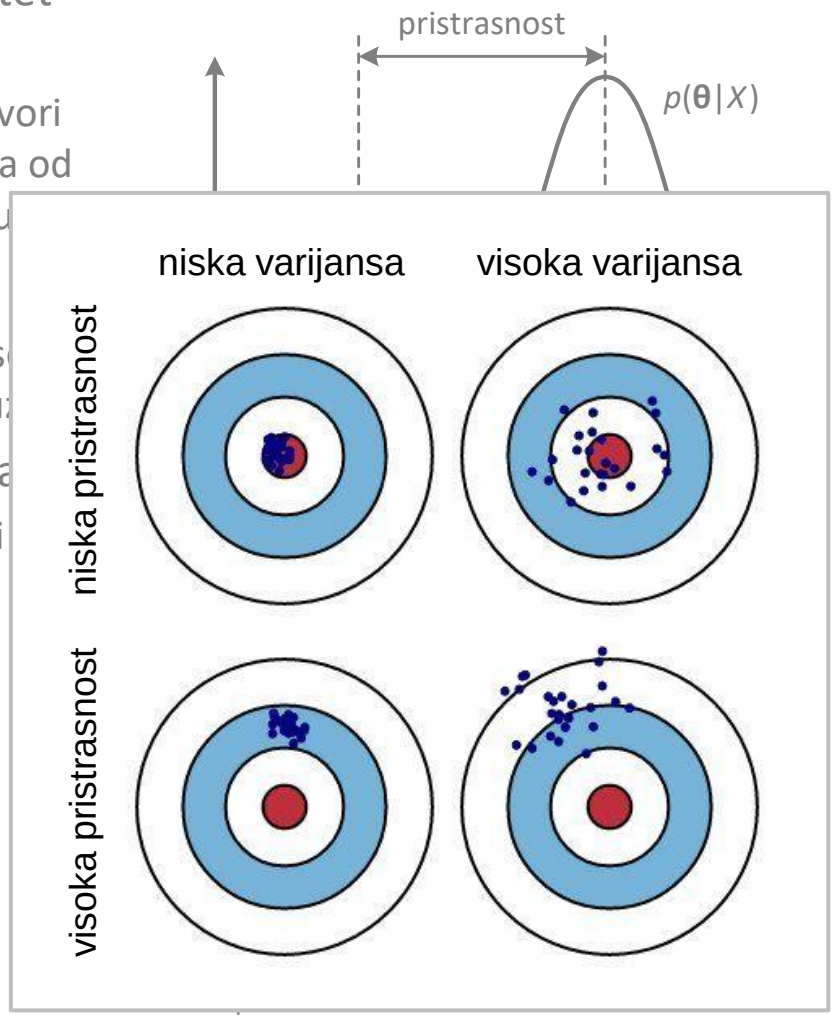


Pristrasnost i varijansa procene

- Postavlja se pitanje kako definisati kvalitet neke procene
  - *Pristrasnost* (*necentriranost*) procene govori o tome koliko procena u proseku odstupa od tačne vrednosti na različitim skupovima u

$$\text{bias}(\hat{\vartheta}) = E(\hat{\vartheta}) - \vartheta$$

  - *Varijansa* procene govori o tome koliko s procena menja na različitim skupovima u
- U praksi treba postići da pristrasnost i va
  - Obično je nemoguće ispuniti oba uslova i na račun druge



Kvalitet ML procene μ i σ^2

- ML procena srednje vrednosti μ je centrirana:

$$E\{\hat{\mu}\} = E\left\{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\right\} = \frac{1}{N}\sum_{i=1}^{N} E\{x^{(i)}\} = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

  dok ML procena varijanse σ^2 nije centrirana:

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

  iako jeste asimptotski centrirana, pošto dobijeni izraz teži σ^2 kada $N \to \infty$

  - Problem je u tome što ML estimacija varijanse koristi ML estimaciju srednje vrednosti umesto prave srednje vrednosti, i taj problem dolazi do izražaja za relativno malo N

- Postoji i centrirana procena varijanse (kao i kovarijansne matrice)

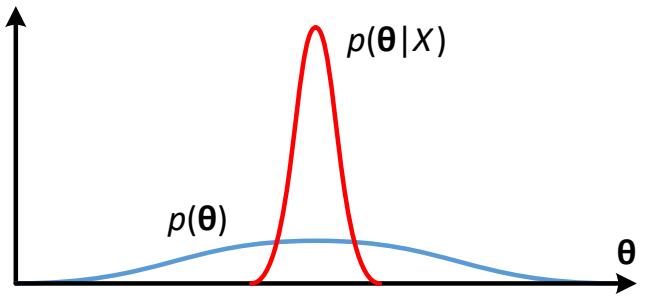
$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)} - \hat{\mu})^2$$

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

  - Nijedna od navedenih procena ne može se smatrati *ispravnom* ili *pogrešnom*, bolje pitanje je da li vodi do dobrih rezultata klasifikatora

Bayesova estimacija

- Pretpostavka je da su parametri nepoznate raspodele $\boldsymbol{\theta}$ *slučajne promenljive* sa poznatom apriornom raspodelom
  - Ova raspodela je po pravilu veoma široka, jer je apriorno znanje o $\boldsymbol{\theta}$ najčešće skromno
- Na osnovu raspoloživog uzorka X nepoznate raspodele estimira se aposteriorna raspodela $p(\boldsymbol{\theta}|X)$
  - Ova raspodela je po pravilu znatno oštrija jer očekujemo da uzorci smanje neodređenost $\boldsymbol{\theta}$



- Pri tome ne treba izgubiti iz vida da je krajnji cilj da se izračuna $p(\mathbf{x})$, ili tačnije $p(\mathbf{x}|X)$, što je gustina raspodele *ako su poznate vrednosti uzoraka*

Bayesova estimacija

- Tražena raspodela $p(\mathbf{x}|X)$ može se dobiti marginalizacijom $p(\mathbf{x},\boldsymbol{\theta}|X)$ po $\boldsymbol{\theta}$

$$p(\mathbf{x}|X) = \int p(\mathbf{x},\boldsymbol{\theta}|X)\,\mathrm{d}\boldsymbol{\theta}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta},X)\,p(\boldsymbol{\theta}|X)\,\mathrm{d}\boldsymbol{\theta} \quad \text{(na osnovu definicije uslovne verovatnoće)}$$

$$= \int p(\mathbf{x}|\boldsymbol{\theta})\,p(\boldsymbol{\theta}|X)\,\mathrm{d}\boldsymbol{\theta} \quad \text{(pošto } X \text{ postaje irelevantno ako je poznato } \boldsymbol{\theta}\text{)}$$

- Jedina nepoznata veličina ovde je $p(\boldsymbol{\theta}|X)$, a ona se može dobiti na osnovu Bayesove teoreme:

$$p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(X)} = \frac{p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(X|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}$$

gde je $p(\boldsymbol{\theta})$ apriorna gustina raspodele verovatnoće za $\boldsymbol{\theta}$

  - $p(X|\boldsymbol{\theta})$ može da se izračuna uz pretpostavku da su uzorci međusobno nezavisni

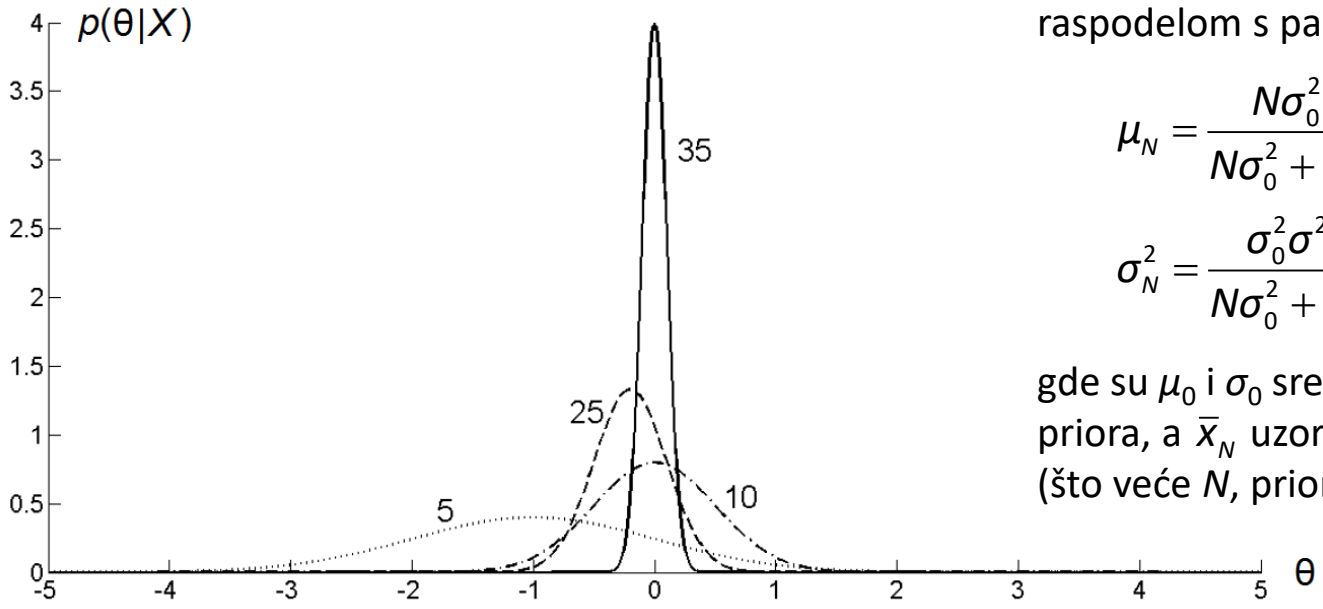
$$p(X|\boldsymbol{\theta}) = p(x^{(1)},x^{(2)},\ldots,x^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x^{(i)}|\boldsymbol{\theta})$$

Veza između ML i Bayesove procene

- Bayesov pristup zahteva računanje višedimenzionalnih integrala, što je po pravilu složenije od primene diferencijalnog računa i gradijentne pretrage koju zahteva ML pristup
- U graničnoj situaciji ova dva pristupa gotovo uvek daju isti rezultat što se tiče procenjene $p(\mathbf{x})$
- Razlike se javljaju kada je skup uzoraka relativno mali
  - Bayesova estimacija koristi potpunu estimaciju $p(\boldsymbol{\theta}|X)$, pa koristi više informacija iz skupa za obuku nego ML estimacija
  - Bayesova estimacija više odgovara načinu na koji ljudi uče, i ima intuitivnu interpretaciju kroz tzv. *inkrementalno učenje*, koje se obavlja s pristizanjem novih uzoraka
    - Počev od $p(\boldsymbol{\theta})$, rekurzivno se ocenjuju raspodele $p(\boldsymbol{\theta}|\mathbf{x}^{(1)})$, $p(\boldsymbol{\theta}|\mathbf{x}^{(1)},\mathbf{x}^{(2)})$, $p(\boldsymbol{\theta}|\mathbf{x}^{(1)},\mathbf{x}^{(2)},\mathbf{x}^{(3)})$...

Bayesova procena i inkrementalno učenje

- Kako novi uzorci pristižu, procena parametra je sve preciznija
  - Prikazani primer odnosi se na procenu nepoznate srednje vrednosti μ 1-D Gaussove raspodele poznate varijanse σ na osnovu skupa uzoraka (pri čemu je i za parametar $\vartheta = \mu$ pretpostavljeno da podleže 1-D Gaussovoj raspodeli)



Po pristizanju N-tog uzorka procena parametra μ definisana je Gaussovom raspodelom s parametrima:

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\bar{x}_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2}$$

gde su μ_0 i σ_0 srednja vrednost i varijansa priora, a $\bar{x}_N$ uzoračka srednja vrednost (što veće N, prior sve manje utiče)