Bilješka 2
Teorija vjerojatnosti

Statističko strojno učenje zasniva se na teoriji vjerojatnosti i statistici. Ovdje razmatramo koncepte koji su ključni za razumijevanje algoritama strojnog učenja.

1 Osnove teorije vjerojatnosti

Diskretno statističko obilježje interpretira se kao **diskretna slučajna varijabla** X sa zadanim skupom vrijednosti $\{x_i\}$. Vrijednost $P(X = x)$ jest vjerojatnost da slučajna varijabla X poprimi vrijednost x, tj. vjerojatnost da se slučajna varijabla X realizira kao x. U nastavku ćemo, osim ako postoji mogućnost zabune, $P(X = x)$ pisati kraće kao $P(x)$. Vrijedi $P(x_i) \geqslant 0$ i $\sum_i P(x_i) = 1$, čime je definirana **diskretna razdioba (distribucija) vjerojatnosti**.

Vjerojatnost da slučajna varijabla X poprimi vrijednost x i da slučajna varijabla Y poprimi vrijednost y pišemo kao $P(X = x, Y = y)$, odnosno kraće $P(x, y)$, i nazivamo **zajednička vjerojatnost** (engl. *joint probability*). Korištenjem **pravila zbroja**

$$P(x) = \sum_y P(x, y) \tag{1}$$

varijablu Y možemo **marginalizirati**, odnosno možemo izračunati **marginalnu vjerojatnost** $P(x)$ varijable X. Naravno, isto tako možemo marginalizirati varijablu X, odnosno izračunati marginalnu vjerojatnost $P(y)$ varijable Y.

Uvjetna vjerojatnost $P(y|x)$, odnosno vjerojatnost da varijabla Y poprimi vrijednost y, pod uvjetom da je varijabla X poprimila vjerojatnost x, definirana je kao

$$P(y|x) = \frac{P(x, y)}{P(x)} \tag{2}$$

što se često iskazuje kao **pravilo umnoška**

$$P(x, y) = P(y|x)P(x)\,. \tag{3}$$

Budući da vrijedi simetričnost $P(x, y) = P(y, x)$, primjenom pravila umnoška na lijevu i desnu stranu jednakosti dobivamo

$$P(x|y)P(y) = P(y|x)P(x)$$

iz čega slijedi poznato **Bayesovo pravilo**

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{4}$$

Ovdje treba primijetiti da se sve vrijednosti koje se pojavljuju u (4) mogu dobiti iz zajedničke vjerojatnosti $P(x, y)$ i to marginalizacijom odnosno normiranjem. Primjenom pravila zbroja i zatim pravila umnoška, marginalna vjerojatnost $P(x)$ može se izraziti kao

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

što je vrlo pogodno jer su onda izrazi u nazivniku isti kao i oni u brojniku. (Primijetite da se marginalna vjerojatnost $P(x)$ mogla izraziti i kao $\sum_y P(y|x)P(x)$, ali time ne bismo ništa dobili.) Bayesovo pravilo možemo dakle pisati kao

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_y P(x|y)P(y)}\,. \tag{5}$$

U gornjim razmatranjima ograničili smo se na slučaj diskretne slučajne varijable. Za kontinuiranu (neprekidnu) slučajnu varijablu definira se **funkcija gustoće vjerojatnosti** (engl. *probability density function*, PDF) $p(x)$, za koju vrijedi[1]

$$p(x) \geqslant 0 \tag{6}$$

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1 \tag{7}$$

Vjerojatnost da kontinuirana slučajna varijabla X poprimi vrijednost iz intervala $[a, b]$ ($a, b \in \mathbb{R}$, $a \leqslant b$) dana je s

$$P(a \leqslant X \leqslant b) = \int_a^b p(x)\,\mathrm{d}x\,.$$

Primijetite da za kontinuiranu varijablu X vrijedi $P(X = a) = 0$, tj. vjerojatnost da kontinuirana varijabla poprimi bilo koju pojedinačnu vrijednost jednaka je nuli. Kažemo da funkcijom gustoće vjerojatnosti $p(x)$ definirana **kontinuirana razdioba (distribucija) vjerojatnosti**.[2]

Pravilo zbroja i pravilo umnoška vrijede i za kontinuirane varijable (a također i za mješovite razdiobe, tj. kombinacije diskretnih i kontinuiranih varijabli):

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\,\mathrm{d}y \tag{8}$$

$$p(x, y) = p(y|x)p(x) \tag{9}$$

1.1 Očekivanje i varijanca

Prosječna vrijednost diskretne slučajne varijable X čija je razdioba $P(x)$ naziva se **(matematičko) očekivanje** varijable X i definira kao

$$\mathbb{E}[X] = \sum_x xP(x) \tag{10}$$

U slučaju kontinuirane slučajne varijable X s gustoćom vjerojatnosti $p(x)$, očekivanje je

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\,p(x)\,\mathrm{d}x\,. \tag{11}$$

[1] Funkcija gustoće vjerojatnosti u matematičkoj literaturi uobičajeno se označava s $f(x)$. Mi ćemo u nastavku koristiti oznaku $p(x)$, koja je tipična u literaturi za strojno učenje.

[2] U nastavku ćemo koristiti izraz *"gustoća $p(x)$"* ili (pomalo neprecizno) *"razdioba $p(x)$"*, mislivši pritom na funkciju gustoće vjerojatnosti $p(x)$. Kada mislimo na konkretnu vrijednost te funkcije u točci x, koristit ćemo izraz *"vrijednost $p(x)$"*.

Vrijedi:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \qquad (a, b \in \mathbb{R}) \tag{12}$$
$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{13}$$

Varijanca slučajne varijable X iskazuje koliko vrijednosti varijable variraju oko očekivane vrijednosti:

$$\mathrm{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \,. \tag{14}$$

(Provjerite da druga jednakost doista vrijedi.) Iz (14) slijedi

$$\mathrm{Var}(aX) = \mathbb{E}\Big[(aX)^2\Big] - \mathbb{E}[aX]^2 = a^2\mathbb{E}[X^2] - a^2\mathbb{E}[X]^2 = a^2\mathrm{Var}(X) \,. \tag{15}$$

Kovarijanca opisuje odnos između dviju slučajnih varijabli, odnosno opisuje u kojoj mjeri slučajne varijable zajednički variraju oko svojih očekivanih vrijednosti. Kovarijanca varijabli X i Y definirana je kao

$$\mathrm{Cov}(X, Y) = \sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \,. \tag{16}$$

(Provjerite da druga jednakost doista vrijedi.) Vrijedi $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ i $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ odnosno $\sigma_{X,X} = \sigma_X^2$.

Za slučajne varijable X i Y za koje vrijedi $\mathrm{Var}(X) > 0$ i $\mathrm{Var}(Y) > 0$ definiran je **koeficijent korelacije**

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \,. \tag{17}$$

Koeficijent korelacije upućuje na mjeru linearne zavisnosti među varijablama X i Y. Za savršenu pozitivnu linearnu ovisnost vrijedi $\rho_{X,Y} = 1$, dok za savršenu negativnu linearnu ovisnost vrijedi $\rho_{X,Y} = -1$.

1.2 Nezavisnost varijabli

Dvije slučajne varijable X i Y su **(stohastički) nezavisne** akko za sve intervale A i B, $A, B \subseteq \mathbb{R}$, vrijedi

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \,.$$

Konkretno, varijable X i Y su nezavisne akko:

$$P(X, Y) = P(X)P(Y)$$

što je ekvivalentno s

$$P(X|Y) = P(X) \qquad \text{i} \qquad P(Y|X) = P(Y) \,.$$

Intuitivno, varijable X i Y su nezavisne ako znanje o ishodu varijable Y ni na koji način ne utječe na vjerojatnost ishoda varijable X (i obrnuto).

Za nezavisne varijable X i Y vrijedi

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{18}$$
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{19}$$
$$\mathrm{Cov}(X, Y) = \rho_{X,Y} = 0 \tag{20}$$

Nezavisne varijable su nekorelirane, no obrat općenito ne vrijedi: koeficijent korelacije može biti jednak nuli, a da su varijable ipak nelinearno zavisne (koeficijent korelacije mjeri isključivo linearnu zavisnost varijabli).

U strojnom učenju važan koncept jest uvjetna nezavisnost varijabli.

Definicija 1 (Uvjetna nezavisnost.) *Slučajne varijable X i Y su* **uvjetno nezavisne** *uz danu varijablu Z, što označavamo kao $X \perp Y|Z$, akko*

$$P(X|Y,Z) = P(X|Z)$$

ili, ekvivalentno

$$P(X,Y|Z) = P(X|Z)P(Y|Z)\,.$$

Intuitivno, varijable X i Y su uvjetno nezavisne ako, jednom kada je poznat ishod varijable Z, znanje o ishodu varijable Y ni na koji način ne utječe na ishod varijable X (i obrnuto). Općenito, nezavisnost varijabli X i Y ne implicira njihovu uvjetnu nezavisnost ni po kojoj varijabli, niti obrnuto.

1.3 Višedimenzijska slučajna varijabla

Istodobno opažanje više statističkih obilježja modeliramo n-dimenzijskim slučajnim vektorom $(X_1, \ldots, X_n)$. Za slučajan vektor definirana je **matrica kovarijacije**[3] $\boldsymbol{\Sigma}$ dimenzija $n \times n$, s elementima:

$$\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(X_i, X_j) = \sigma_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]\,.$$

Odnosno, u matričnom računu, kovarijacijska je matrica definirana kao:

$$\boldsymbol{\Sigma} = \mathbb{E}\Big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}}\Big]\,. \tag{21}$$

Kovarijacijska matrica je kvadratna simetrična matrica koja na dijagonali ima varijance varijabli $X_1, \ldots, X_n$, a izvan dijagonale kovarijance svih parova varijabli:

$$\begin{aligned}\boldsymbol{\Sigma} &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \ldots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \ldots & \mathrm{Var}(X_n) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \ldots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \ldots & \sigma_n^2 \end{pmatrix}\end{aligned}$$

Primjer 1 (Kovarijacijska matrica) Razmotrimo slučaj dvodimenzijske kontinuirane slučajne varijable, $\mathbf{X} = (X_1, X_2)$. Neka su varijance $\sigma_1^2 = 1$ i $\sigma_2^2 = 4$ te neka su varijable pozitivno korelirane s faktorom $\rho_{12} = \rho_{21} = 0.75$. Kovarijacijska matrica je

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix}$$

[3] Također *kovarijancna matrica* i *disperzijska matrica*.

Ako su varijable $X_1, \ldots, X_n$ međusobno nezavisne, onda $\mathrm{Cov}(X_i, X_j) = 0$ i kovarijacijska je matrica dijagonalna matrica, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$. Nadalje, ako su nezavisne varijable $X_1, \ldots, X_n$ identično distribuirane, onda $\sigma_i^2 = \sigma^2$, pa kovarijacijska matrica degenerira u $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Takav slučaj nazivamo **izotropnom kovarijancom**.

2 Teorijske razdiobe

U nastavku ćemo se prisjetiti nekoliko vjerojatnosnih razdioba koje su nam potrebne u strojnom učenju.

2.1 Bernoullijeva razdioba

Bernuoullijeva razdioba jest razdioba diskretne slučajne varijable X koja ima dva moguća ishoda: događaj je ili nastupio ili nije, $x \in \{0, 1\}$. Razdioba je dana kao

$$P(X = x|\mu) = \mu^x(1-\mu)^{1-x} = \begin{cases} \mu & \text{ako } X = 1 \\ 1-\mu & \text{inače} \end{cases} \tag{22}$$

gdje parametar μ definira vjerojatnost nastupanja događaja, tj. vjerojatnost $P(X = 1)$. Iz (10), (14) i (22) slijedi da su očekivana vrijednost i varijanca

$$\begin{aligned} \mathbb{E}[X] &= \mu \\ \mathrm{Var}(X) &= \mu(1-\mu) \end{aligned}$$

2.2 Multinomijalna razdioba

Razmotrimo poopćenje Bernoullijeve razdiobe na slučaj kada slučajna varijabla X može poprimiti jedno od K međusobno isključivih stanja, $K \geqslant 2$ (npr. jednu od K klasa). Takvu varijablu nazivamo **multinomijalna varijabla**. Multinomijalnu varijablu prikazujemo kao vektor indikatorskih (binarnih) varijabli

$$\mathbf{x} = (x_1, x_2, \ldots, x_K)^{\mathrm{T}}$$

gdje je $x_k = 1$ ako je ishod varijable k, a inače $x_k = 0$. Npr. $\mathbf{x} = (0, 0, 1, 0)$ označava da je multinomijalna varijabla poprimila treće stanje od četiri mogućih stanja. Pritom vrijedi $\sum_k x_k = 1$ (ishodi su međusobno isključivi). Označimo vjerojatnost $P(X_k = 1)$ sa μ_k. Razdioba je dana s

$$P(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k^{(i)}} \tag{23}$$

gdje je $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$, pri čemu za parametre μ_k vrijedi $\sum_k \mu_k = 1$ i $\mu_k \geqslant 0$, budući da predstavljaju vjerojatnosti.

Navedena razdioba naziva se **kategorička razdioba** i zapravo je poseban slučaj **multinomijalne razdiobe** kod koje je broj eksperimenata jednak 1.[4] Multinomijalnu varijablu ne treba miješati s višedimenzijskom varijablom odnosno slučajnim vektorom (premda ovdje multinomijalnu varijablu modeliramo pomoću slučajnog vektora).

[4] U strojnom učenju i nekim drugim područjima ova se razlika često zanemaruje.

2.3 Gaussova razdioba

Neprekidna slučajna varijabla X ima Gaussovu (normalnu) razdiobu sa srednjom vrijednošću μ i varijancom σ^2, što označavamo kao $\mathcal{N}(\mu, \sigma^2)$, ako je njezina gustoća vjerojatnosti jednaka

$$p(X = x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}. \tag{24}$$

Za normalnu razdiobu vrijedi

$$\mathbb{E}[X] = \mu \tag{25}$$
$$\mathrm{Var}(X) = \sigma^2 \tag{26}$$

2.4 Multivarijatna Gaussova razdioba

Poopćenjem Gaussove razdiobe na n dimenzija dobivamo **multivarijatnu Gaussovu razdiobu**:

$$p(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\} \tag{27}$$

gdje je $\boldsymbol{\mu}$ n-dimenzijski vektor srednje vrijednosti, a $\boldsymbol{\Sigma}$ je kovarijacijska matrica dimenzija $n \times n$. Da bi razdioba (24) bila dobro definirana, kovarijacijska matrica $\boldsymbol{\Sigma}$ mora biti pozitivno definitna,[5] jer tada ima pozitivnu determinantu i nesingularna je (ima inverz). Vrijednost gustoće funkcijski zavisi o $\mathbf{x}$ preko **kvadratne forme**[6]

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

koja se javlja u eksponentu. Vrijednost Δ nazivamo **Mahalanobisova udaljenost** između $\mathbf{x}$ i $\boldsymbol{\mu}$. Mahalanobisova udaljenost je poopćenje euklidske udaljenosti koje je neosjetljivo na razlike u varijanci između pojedinih dimenzija te na korelacije između varijabli. Mahalanobisova udaljenost svodi se na euklidsku za $\boldsymbol{\Sigma} = \mathbf{I}$.

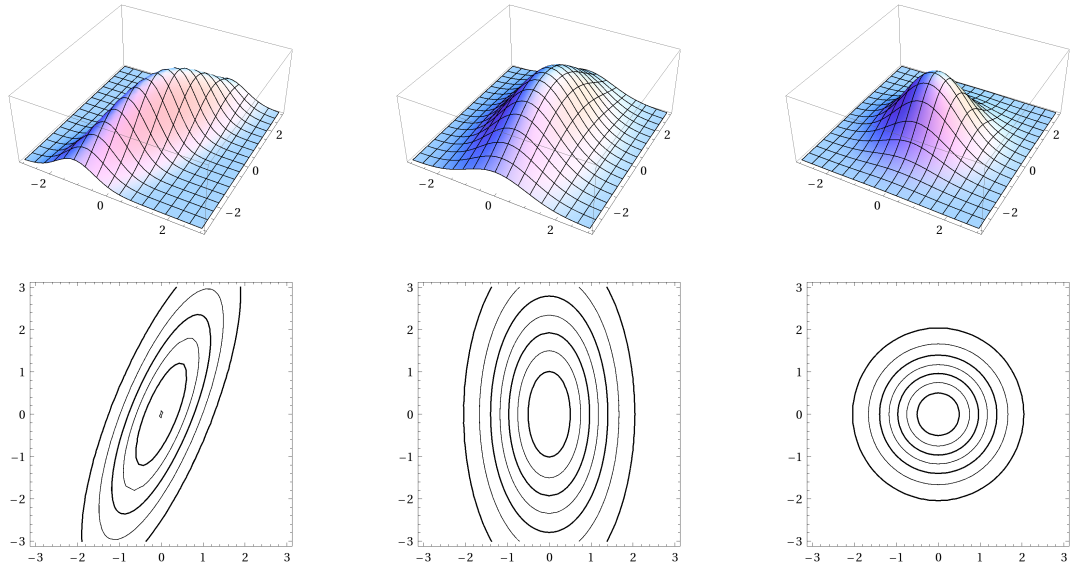
Za multivarijatnu normalnu razdiobu vrijedi

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} \tag{28}$$
$$\mathrm{Cov}(X_i, X_j) = \boldsymbol{\Sigma}_{ij} \tag{29}$$

Na slici 1 prikazane su gustoće vjerojatnosti i njima odgovarajuće konture jednake gustoće za dvodimenzijsku normalno distribuiranu varijablu sa srednjom vrijednošu $\boldsymbol{\mu} = (0, 0)^{\mathrm{T}}$. Slika 1a prikazuje najopćenitiji slučaj u kojemu su varijance pojedinih komponenti različite i korelirane. U tom slučaju konture Gaussove razdiobe su nakošene elipse (odnosno općenito hiperelipsoidi). Slika 1b prikazuje slučaj nekoreliranih varijabli s različitim varijancama (dijagonalna kovarijacijska matrica). U tom slučaju konture Gaussove razdiobe su elipse čije su osi poravnate s apscisom i ordinatom (odnosno općenito hiperelipsoidi poravnati s osima). Slika 1c prikazuje slučaj nekoreliranih i identično distribuiranih komponenti (izotropna kovarijacijska matrica). U ovom slučaju konture Gaussove razdiobe su kružnice (odnosno općenito hipersfere).

[5] Matrica $\mathbf{A}$ je *pozitivno definitna* akko $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$.

[6] Za kvadratnu matricu $\mathbf{A}$ funkcija $f : \mathbb{R}^n \to \mathbb{R}$ definirana kao $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ naziva se *kvadratna forma*. Pomoću skalarnog produkta kvadratna se forma može sažetije napisati kao $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x}|\mathbf{x}) = (\mathbf{x}|\mathbf{A}\mathbf{x})$.



(a) (b) (c)

Slika 1: Gustoća vjerojatnosti dvodimenzijske normalno distribuirane varijable sa srednjom vrijednošću $\boldsymbol{\mu} = (0, 0)^{\mathrm{T}}$: (a) slučaj za $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\rho_{12} = 0.75$, (b) za slučaj $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\rho_{12} = 0$, (c) za slučaj $\sigma_1^2 = \sigma_2^2 = 1$, $\rho_{12} = 0$.

3 Procjena parametara

Osnovna ideja procjene parametara jest na temelju uzorka najprije izračunati vrijednost određene slučajne varijable, tzv. **statistike**, koju koristimo kao **procjenitelj** (estimator) nepoznatog parametra $\boldsymbol{\theta}$ populacije (odnosno teorijske razdiobe).

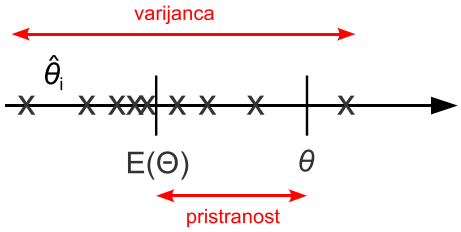
3.1 Procjenitelj

Definicija 2 (Statistika, procjenitelj i procjena) *Neka je* $(X_1, X_2, \ldots, X_n)$ *uzorak, odnosno* n*-torka slučajnih varijabli koje su iid. Slučajna varijabla* $\Theta = g(X_1, X_2, \ldots, X_n)$ *naziva se* **statistika**. *Statistika* Θ *je* **procjenitelj (estimator)** *parametra populacije* θ. *Vrijednost procjenitelja* $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ *naziva se* **procjena**.

Statistika je dakle bilo kakva funkcija koja ovisi o uzorku, dok je procjenitelj ona statistika koja se koristi za procjenu nekog parametra populacije. Procjenitelj je slučajna varijabla, što znači da ima svoje očekivanje i varijancu. Očekivanje i varijancu procjenitelja koristimo kako bismo ocijenili kvalitetu procjenitelja, odnosno kako bismo ocijenili koliko dobro procjenitelj Θ procjenjuje parametar populacije θ.

Definicija 3 (Nepristran procjenitelj) *Procjenitelj* Θ *je* **nepristran (centriran) procjenitelj** *parametra* θ *akko* $\mathbb{E}[\Theta] = \theta$.

Vrijednost procjenitelja $\hat{\theta}$ na konkretnom uzorku $\mathcal{D}$ može biti različita od prave vrijednosti parametra θ, ali ako je procjenitelj nepristran, onda će se kod ponavljanja eksperimenta prosječna vrijednost procjenitelja približiti stvarnoj vrijednosti parametra. **Pristranost** procjenitelja definirana je kao



Slika 2: Pristranost i varijanca procjenitelja: θ je parametar koji treba procijeniti, $\hat{\theta}_i$ su pojedinačne procjene na različitim uzorcima, a $\mathbb{E}[\Theta]$ je očekivana vrijednost procjenitelja.

$$b_\theta(\Theta) = \mathbb{E}[\Theta] - \theta\,.$$

Idealan procjenitelj je nepristran i ima malu varijancu. Primijetite da su ove dvije veličine nezavisne te da pristranost ovisi o vrijednosti parametra, dok varijanca ne ovisi. Pristranost i varijanca procjenitelja skicirani su na slici 2.

Dodatno poželjno svojstvo jest da s porastom veličine uzorka (odnosno broja primjera za učenje) procjena sve manje odstupa od svoje očekivane vrijednosti. Drugim riječima, želimo da s povećanjem uzorka varijanca procjenitelja teži k nuli.

Definicija 4 (Valjan procjenitelj) *Nepristan procjenitelj* Θ *je* **valjan (konzistentan)** *procjenitelj ako* $\lim_{N\to\infty} \mathrm{Var}(\Theta) = 0$.

Primjer 2 (Procjenitelj srednje vrijednosti) Neka je X slučajna varijabla s vrijednostima $x \in \mathbb{R}$. Označimo očekivanje i varijancu ove varijable s $\mathbb{E}[X] = \mu$ odnosno $\mathrm{Var}(X) = \sigma^2$. Ova se varijabla pokorava razdiobi sa srednjom vrijednošću μ i varijancom σ^2. Neka je $\{x^{(i)}\}_{i=1}^N$ uzorak ove slučajne varijable.

Prave vrijednosti parametara μ i σ^2 nisu nam poznate, ali ih možemo procijeniti na temelju uzorka. Tako za procjenu parametra μ (srednja vrijednost razdiobe) možemo koristiti **sredinu uzorka**:

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

Sredina uzorka nepristran je procjenitelj parametra razdiobe μ, budući da vrijedi

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\Big[\frac{\sum_i^N x^{(i)}}{N}\Big] = \frac{1}{N}\sum_i^N \mathbb{E}[X^{(i)}] = \frac{N\mu}{N} = \mu\,.$$

Pokažimo da je $\hat{\mu}$ ujedno i valjan procjenitelj. Vrijedi

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\Big(\frac{\sum_i^N x^{(i)}}{N}\Big) = \frac{1}{N^2}\sum_i^N \mathrm{Var}(X^{(i)}) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \tag{30}$$

pri čemu smo iskoristili jednakost (15) te jednakost (19), koja vrijedi jer su varijable $X^{(i)}$ iid. Očito

$$\lim_{N\to\infty} \frac{\sigma^2}{N} = 0$$

pa zaključujemo da je $\hat{\mu}$ valjan procjenitelj.

Primjer 3 (Procjenitelj varijance) Neka su $x^{(i)} \in \mathbb{R}$ uzorci slučajne varijable X koja se pokorava razdiobi sa srednjom vrijednošću μ i varijancom σ^2, tj. $\mathbb{E}[X] = \mu$ i $\mathrm{Var}(X) = \sigma^2$. Neka je $\hat{\mu}$ nepristran procjenitelj srednje vrijednosti (v. primjer 2). Provjerimo je li procjenitelj

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

nepristran procjenitelj parametra σ^2. Uvrštavanjem $\sum_i x^{(i)} = N\hat{\mu}$ dobivamo

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

Očekivanje je

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\Big[\frac{1}{N}\Big(\sum_{i=1}^{N}(x^{(i)})^2 - N\hat{\mu}^2\Big)\Big] = \frac{1}{N}\Big(N\mathbb{E}[(X)^2] - N\mu^2\Big) = \mathbb{E}[X^2] - \mathbb{E}[\hat{\mu}^2]\,.$$

Iz (14) slijedi $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$, pa vrijedi $\mathbb{E}[X^2] = \sigma^2 + \mu^2$. Slično, korištenjem (30), dobivamo $\mathbb{E}[\hat{\mu}^2] = \sigma^2/N + \mu^2$. Uvrštavanjem ovih jednakosti u gornju jednakost dobivamo

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 + \mu^2 - \sigma^2/N - \mu^2 = \frac{N-1}{N}\sigma^2\,.$$

Budući da $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$, zaključujemo da $\hat{\sigma}^2$ nije nepristran procjenitelj varijance σ^2. Preciznije, $\hat{\sigma}$ **podcjenjuje** pravu varijancu jer $\hat{\sigma} < \sigma$. Pristranost procjenitelja $\hat{\sigma}$ je

$$b(\hat{\sigma}^2) = \frac{N-1}{N}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{N}$$

i ona se smanjuje kako $N \to \infty$. Za manje uzorke nepristranost može predstavljati problem i tada je procjenitelj moguće korigirati (učiniti nepristranim) tako da ga se pomnoži s $N/(N-1)$, tj. da se kao procjenitelj varijance koristi:

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

3.2 Pogreška procjenitelja

Za procjenitelj Θ srednju kvadratnu pogrešku definiramo kao

$$r(\Theta, \theta) = \mathbb{E}[(\Theta - \theta)^2] \tag{31}$$

i to je u stvari funkcija rizika uz kvadratnu funkciju gubitka (prisjetite se da je funkcija rizika definirana kao očekivanje funkcije gubitka). Pokažimo kako se izraz (31) može rastaviti na pristranost i varijancu. Sasvim općenito, za slučajnu varijablu X i konstantu c

vrijedi:

$$\begin{aligned}
\mathbb{E}[(X-c)^2] &= \mathbb{E}\Big[(X-\mathbb{E}[X]+\mathbb{E}[X]-c)^2\Big] \\
&= \mathbb{E}\Big[\big((X-\mathbb{E}[X])+(\mathbb{E}[X]-c)\big)^2\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])^2+2(X-\mathbb{E}[X])(\mathbb{E}[X]-c)+(\mathbb{E}[X]-c)^2\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big]+2\underbrace{\mathbb{E}\Big[(X-\mathbb{E}[X])(\mathbb{E}[X]-c)\Big]}_{\mathbb{E}[(X-\mu)(\mu-c)]=(\mu-c)\mathbb{E}[X-\mu]=0}+\mathbb{E}\Big[(\mathbb{E}[X]-c)^2\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big]+(\mathbb{E}[X]-c)^2\,. \qquad (32)
\end{aligned}$$

Navedena jednakost vrijedi jer je $\mathbb{E}[X]=\mu$ konstanta, pa vrijedi $\mathbb{E}[X-\mu]=\mathbb{E}[X]-\mu=0$, a također $\mathbb{E}[(\mathbb{E}[X]-c)^2]=(\mathbb{E}[X]-c)^2$. Primijenimo li (32) na (31), dobivamo:

$$r(\Theta,\theta)=\underbrace{\mathbb{E}\Big[(\Theta-\mathbb{E}[\Theta])^2\Big]}_{\text{varijanca}}+\underbrace{(\mathbb{E}[\Theta]-\theta)^2}_{\text{pristranost}^2}=\mathrm{Var}(\Theta)+b_\theta(\Theta)^2\,. \qquad (33)$$

Srednja kvadratna pogreška procjenitelja može se dakle rastaviti na varijancu i kvadrat pristranosti. Pristranost nam kazuje koliko procjenitelj griješi neovisno o varijacijama u uzorku, dok varijanca kazuje koliko vrijednost procjenitelja varira oko očekivane vrijednosti kako uzorak varira. Primijetite da, ako je procjenitelj nepristran, $b_\theta(\Theta)=0$, srednja kvadratna pogreška procjenitelja jednaka je njegovoj varijanci.

4 Procjenitelj najveće izglednosti

Najjednostavniji, najvažniji i u praksi najčešće korišten procjenitelj jest **procjenitelj najveće izglednosti** (engl. *maximum likelihood estimation*, MLE).[7] Neka je dan skup neoznačenih primjera $\mathcal{D}=\{\mathbf{x}^{(i)}\}_{i=1}^N$, za koje pretpostavljamo da su nezavisni i da potječu od identične razdiobe (pretpostavka iid). Pretpostavljamo da je riječ o nekoj nama poznatoj razdiobi $p(\mathbf{x}|\theta)$, definiranoj do na parametre $\boldsymbol{\theta}$:

$$\mathbf{x}^{(i)}\sim p(\mathbf{x}|\boldsymbol{\theta})\,.$$

Naš je zadatak odrediti *najizglednije* parametre $\boldsymbol{\theta}$, odnosno takve parametre da uzorkovanje primjera $\mathbf{x}^{(i)}\in\mathcal{D}$ bude što je vjerojatnije moguće. Naime, u nedostatku dodatnih spoznaja, jedino što opravdano možemo pretpostaviti jest da je baš za uzorak $\mathcal{D}$ – dakle uzorak koji imamo na raspolaganju – bilo najvjerojatnije da bude izvučen iz populacije, i da je to razlog zašto je realiziran baš taj uzorak. Budući da pretpostavljamo da su primjeri $\mathbf{x}^{(i)}$ iid, to je gustoća vjerojatnosti uzorka $\mathcal{D}$ jednaka umnošku gustoća vjerojatnosti pojedinačnih primjera:

$$p(\mathcal{D}|\boldsymbol{\theta})=p(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(N)}|\boldsymbol{\theta})=\prod_{i=1}^N p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\;\equiv L(\boldsymbol{\theta}|\mathcal{D})\,. \qquad (34)$$

Ovime je definirana gustoća vjerojatnosti za uzorak $\mathcal{D}$, parametrizirana parametrima $\boldsymbol{\theta}$ (ako je varijabla $\mathbf{x}$ diskretna, umjesto gustoće vjerojatnosti $p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$ koristi se vjerojatnost

[7] Također: *metoda najveće vjerodostojnosti, metoda najveće vjerojatnosti, ML-metoda.*

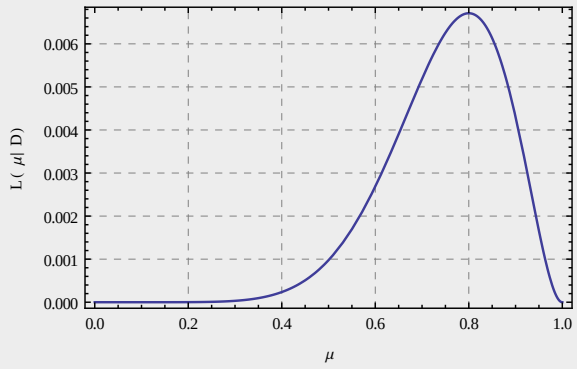
$P(\mathbf{x}^{(i)}|\boldsymbol{\theta})$). Međutim, moguće je gledati i obrnuto, pa reći da je to funkcija od $\boldsymbol{\theta}$, uz zadani parametar $\mathcal{D}$. Tada tu funkciju nazivamo **funkcija izglednosti** (engl. *likelihood function*) i označavamo s $L(\boldsymbol{\theta}|\mathcal{D})$. Ovo su zapravo dva različita pogleda na istu stvar: funkcija gustoće $p(\mathcal{D}|\boldsymbol{\theta})$ preslikava $\mathcal{D} \mapsto p(\mathcal{D}|\boldsymbol{\theta})$, dok funkcija izglednosti preslikava $\boldsymbol{\theta} \mapsto p(\mathcal{D}|\boldsymbol{\theta})$. Treba naglasiti da funkcija izglednosti $L(\boldsymbol{\theta}|\mathcal{D})$ nije funkcija gustoće vjerojatnosti od $\boldsymbol{\theta}$ te da ona ne mora biti normirana, tj. ne mora vrijediti $\int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathcal{D})\,\mathrm{d}\boldsymbol{\theta} = 1$. Očito je da $L(\boldsymbol{\theta}|\mathcal{D})$ nije normirana budući da općenito ne vrijedi $\int_y p(x|y)\mathrm{d}y = 1$.

Primjer 4 (Funkcija izglednosti) U 10 bacanja novčića ($N = 10$), glavu (H) dobivamo 8 puta, a pismo (T) 2 puta. Ishodi bacanja novčića čine naš uzorak $\mathcal{D}$.

Neka je parametar μ jednak vjerojatnosti da u bacanju novčića dobijemo glavu. Vrijedi dakle $P(X = H|\mu) = \mu$ i $P(X = T|\mu) = 1 - \mu$. Primijetite da je ovdje riječ o diskretnoj varijabli koja se pokorava Bernoullijevoj razdiobi parametriziranoj parametrom μ prema (22). Funkcija izglednosti za uzorak $\mathcal{D}$ je

$$L(\mu|\mathcal{D}) = P(\mathcal{D}|\mu) = \prod_{i=1}^{10} P(X|\mu) = \mu^8 \cdot (1-\mu)^2$$

te izgleda ovako



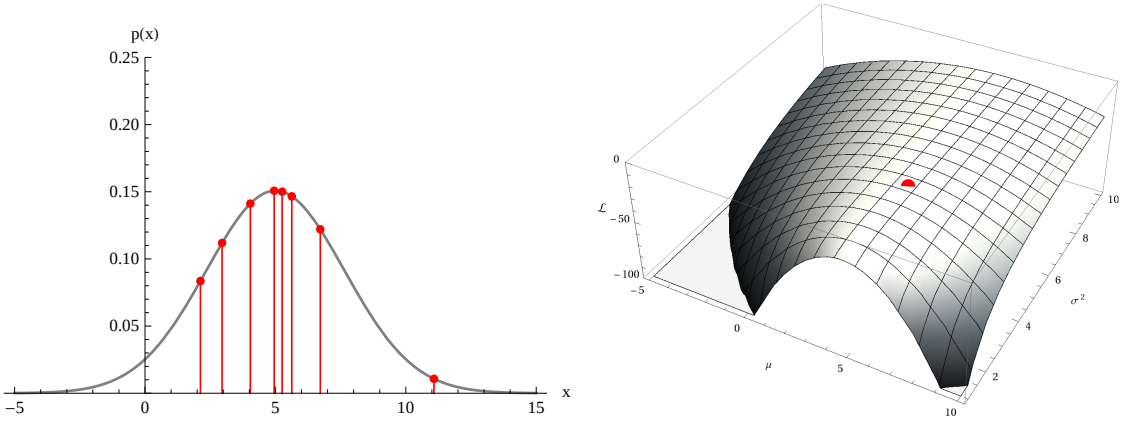
Možemo se, na primjer, pitati kolika je izglednost da je vrijednost parametra μ jednaka 0.5. Vidimo da $L(\mu = 0.5|\mathcal{D}) \approx 0.001$. Važno je naglasiti da to ne znači da je vjerojatnost da $\mu = 0.5$, uz zadani $\mathcal{D}$, jednaka 0.001 (funkcija izglednosti nije funkcija gustoće vjerojatnosti). Međutim, vrijedi obrnuto: 0.001 jest vjerojatnost uzorka $\mathcal{D}$, ako $\mu = 0.5$, tj. $P(\mathcal{D}|\mu = 0.5) = 0.001$.

Iz slike vidimo da je za uzorak najizglednije da $\mu = 0.8$, i to je vrijednost parametra kojom se maksimizira realizacija uzorka.

Kod procjene metodom najveće izglednosti naš je cilj pronaći $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ koji maksimizira funkciju izglednosti, budući da time maksimiziramo vjerojatnost pojavljivanja uzorka $\mathcal{D}$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}|\mathcal{D})\,.$$

Na slici 3 ta je ideja ilustrirana za Gaussovu razdiobu. Parametri Gaussove razdiobe su $\boldsymbol{\theta} = (\mu, \sigma^2)$, gdje je μ srednja vrijednost a σ^2 je varijanca. Gaussova razdioba koja maksimizira vjerojatnost danog uzorka je upravo ona koja je prikazana na slici 3a. Parametri te razdiobe su $\mu = 5$ i $\sigma^2 = 7$, i to su najizgledniji parametri za dani uzorak. Bilo



(a) (b)

Slika 3: Maksimizacija izglednosti parametara μ i σ^2 Gaussove razdiobe: (a) Gaussova razdioba $\mathcal{N}(\mu, \sigma^2)$ koja maksimizira vjerojatnost uzorka, (b) funkcija log-izglednosti $\mathcal{L}(\mu, \sigma^2|\mathcal{D})$ (njezin maksimum označen je crvenom točkom).

koji drugi parametri promijenili bi izgled Gaussove krivulje i to bi dovelo do smanjenja umnoška $\prod p(x|\mu, \sigma^2)$, dakle do smanjenja vjerojatnosti uzorka $\mathcal{D}$.

Umjesto maksimiziranja funkcije izglednosti, često je mnogo praktičnije maksimizirati (prirodan) logaritam te funkcije, odnosno funkciju **log-izglednosti** (engl. *log-likelihood*):

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) \equiv \ln L(\boldsymbol{\theta}|\mathcal{D}) = \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\,. \tag{35}$$

Maksimizacija funkcije log-izglednosti istovjetna je maksimizaciji funkcije izglednosti budući da ove dvije funkcije maksimum poprimaju u istim točkama (funkcija ln je monotona). Maksimizacija ove funkcije je praktičnija jer umjesto umnoška radimo sa zbrojem, a lakše je i baratati s argumentima eksponencijalne funkcije. Na slici 3b dan je primjer funkcije log-izglednosti $\mathcal{L}(\mu, \sigma^2|\mathcal{D})$ Gaussove razdiobe.

U slučajevima kada je to moguće i kada je to učinkovito, ovu optimizaciju provodimo analitički (nalaženjem nul-točaka prve derivacije). U mnogim slučajevima međutim analitičko rješenje ili nije moguće ili je računalno prezahtjevno, i tada pribjegavamo iterativnim optimizacijskim metodama.

U nastavku ćemo izvesti procjenitelje najveće izglednosti za nekoliko nama interesantnih teorijskih razdioba. U svim slučajevima koje ćemo razmotriti optimizaciju je moguće provesti analitički.

4.1 ML-procjenitelj za Bernoullijevu razdiobu

Bernoullijeva razdioba ima μ kao jedini parametar. Izvedimo njegov procjenitelj najveće izglednosti. Funkcija log-izglednosti je

$$\mathcal{L}(\mu|\mathcal{D}) = \ln \prod_{i=1}^{N} P(x|\mu) = \ln \prod_{i=1}^{N} \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}} = \sum_{i=1}^{N} x^{(i)} \ln \mu + \left(N - \sum_{i=1}^{N} x^{(i)}\right) \ln(1-\mu)\,.$$

Deriviranjem i izjednačavanjem s nulom dobivamo

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\mu} = \frac{1}{\mu}\sum_{i=1}^{N} x^{(i)} - \frac{1}{1-\mu}(N - \sum_{i=1}^{N} x^{(i)}) = 0$$

iz čega kao procjenitelj najveće izglednosti slijedi

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}\,. \tag{36}$$

Prema tome, najizglednija procjena za μ je **relativna frekvencija** događaja u uzorku, odnosno srednja vrijednost uzorka. Kako je pokazano u primjeru 2, očekivanje tog procjenitelja jednako je srednjoj vrijednosti razdiobe, $\mathbb{E}[x]$. Budući da za Bernoullijevu razdiobu vrijedi $\mathbb{E}[x] = \mu$, to slijedi $\mathbb{E}[\hat{\mu}_{\mathrm{ML}}] = \mu$, odnosno $\hat{\mu}_{\mathrm{ML}}$ je nepristran procjenitelj.

Važno je naglasiti da, premda je procjenitelj $\hat{\mu}_{\mathrm{ML}}$ nepristran (a također i valjan), procjena ipak ovisi o uzorku i ona ne mora uvijek biti dobra. Na primjer, za uzorak za koji $x^{(i)} = 1$, najizglednija procjena je $\hat{\mu}_{\mathrm{ML}} = 1$, što dovodi do **prenaučenosti** modela.

4.2 ML-procjenitelj za multinomijalnu varijablu

Funkcija log-izglednosti za razdiobu (23) je

$$\mathcal{L}(\boldsymbol{\mu}|\mathcal{D}) = \ln\prod_{i=1}^{N} P(\mathbf{x}|\boldsymbol{\mu}) = \ln\prod_{i=1}^{N}\prod_{k=1}^{K}\mu_k^{x_k^{(i)}} = \sum_{k=1}^{K}\sum_{i=1}^{N} x_k^{(i)}\ln\mu_k\,. \tag{37}$$

Kako bismo dobili ML-procjenitelj, gornji izraz potrebno je maksimizirati s obzirom na μ_k, kao što smo radili i ranije. Međutim, u ovom slučaju u obzir moramo uzeti ograničenje $\sum_{k=1}^{K}\mu_k = 1$; ukoliko to ne učinimo, nećemo dobiti ispravno rješenje (uvjerite se u to). Optimizaciju s ograničenjem možemo provesti **metodom Lagrangeovih multiplikatora**. Uvodimo novu varijablu, tzv. **Lagrangeov multiplikator** λ, i umjesto maksimizacije izraza (37), maksimiziramo odgovarajuću Lagrangeovu funkciju:

$$\sum_{i=1}^{K}\sum_{i=1}^{N} x_k^{(i)}\ln\mu_k + \lambda\Big(\sum_{k=1}^{K}\mu_k - 1\Big)\,.$$

Deriviranjem po μ_k i izjednačavanjem s nulom dobivamo

$$\mu_k = -\frac{1}{\lambda}\sum_{i=1}^{N} x_k^{(i)}\,. \tag{38}$$

Kako bismo izračunali vrijednost multiplikatora λ, dobiveni izraz uvrštavamo u ograničenje $\sum_k \mu_k = 1$ i tako dobivamo:

$$\sum_{k=1}^{K}\mu_k = -\frac{1}{\lambda}\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N} x_k^{(i)}}_{=N} = 1\,.$$

Budući da svaka multinomijalna varijabla $\mathbf{x}$ ima jedinicu postavljenu na samo jednoj komponenti, zbroj komponenata svih varijabli jednak je broju primjera N. Vrijedi dakle $\lambda = -N$, pa uvrštavanjem u (38) za ML-procjenitelj konačno dobivamo:

$$\hat{\mu}_{k,\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_k^{(i)} = \frac{N_k}{N} \tag{39}$$

gdje smo s N_k označili broj koliko je puta varijabla u uzorku poprimila vrijednost k. Kao i kod Bernoullijeve varijable, i ovdje smo za najizgledniju procjenu dobili relativnu frekvenciju, što je očekivan rezultat. Alternativno, optimizaciju izraza (37) mogli smo provesti tako da smo multinomijalnu varijablu $\mathbf{x}$ tretirali kao K nezavisnih realizacija binarne varijable.

4.3 ML-procjenitelji za Gaussovu razdiobu

Izvedimo procjenitelje maksimalne izglednosti za parametre μ i σ^2 normalne razdiobe. Za dani uzorak $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ takav da $x^{(i)} \sim \mathcal{N}(\mu, \sigma^2)$, funkcija log-izglednosti je

$$\begin{aligned}\mathcal{L}(\mu, \sigma|\mathcal{D}) &= \ln \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\Big\} = \sum_{i=1}^N \Big(-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(x^{(i)}-\mu)^2}{2\sigma^2}\Big) \\ &= -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \frac{\sum_i (x^{(i)}-\mu)^2}{2\sigma^2}\,.\end{aligned}$$

Postavljanjem $\nabla\mathcal{L}(\mu, \sigma|\mathcal{D}) = 0$ i rješavanjem po μ odnosno σ^2 dobivamo procjenitelje najveće izglednosti

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^N x^{(i)} \tag{40}$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^N (x^{(i)} - \hat{\mu}_{\mathrm{ML}})^2\,. \tag{41}$$

(Uvjerite se u ove jednakosti.) Primijetite da je procjenitelj $\hat{\sigma}^2_{\mathrm{ML}}$ izražen pomoću procjenitelja $\hat{\mu}_{\mathrm{ML}}$, zato jer nam je prava vrijednost parametra μ nepoznata.

Za procjenitelj srednje vrijednosti $\hat{\mu}_{\mathrm{ML}}$ već smo utvrdili da je to nepristran procjenitelj za koji vrijedi $\mathbb{E}[\hat{\mu}_{\mathrm{ML}}] = \mu$ i $\mathrm{Var}(\hat{\mu}_{\mathrm{ML}}) = \sigma^2/N$ (v. primjer 2). Za razliku od procjenitelja $\hat{\mu}_{\mathrm{ML}}$, koji je nepristran, procjenitelj varijance $\hat{\sigma}^2_{\mathrm{ML}}$ nije nepristran (v. primjer 3). Razlika dolazi do izražaja kod malenih uzoraka, kada je za procjenu bolje koristiti nepristran procjenitelj $N\hat{\sigma}^2_{\mathrm{ML}}/(N-1)$.

Kao što ovaj primjer ilustrira, sasvim je moguće da postupak najveće izlgednosti rezultira procjeniteljem koji nije nepristran (*najveća izglednost* ne znači nužno *nepristranost*).

4.4 ML-procjenitelji za multivarijatnu Gaussovu razdiobu

Izvedimo ML-procjenitelje za parametre $\boldsymbol{\mu}$ i $\boldsymbol{\Sigma}$ multivarijatne Gaussove razdiobe. Log-izglednost jednaka je

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) &= \ln \prod_{i=1}^N p(\mathbf{x}^{(i)}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= -\frac{nN}{2}\ln(2\pi) - \frac{N}{2}|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^N (\mathbf{x}^{(i)}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu})\,.\end{aligned} \tag{42}$$

Derivacijom po $\boldsymbol{\mu}$ i izjednačavanjem s nulom dobivamo

$$\frac{\partial}{\partial\boldsymbol{\mu}}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^N \boldsymbol{\Sigma}^{-1}(\mathbf{x}^{(i)}-\boldsymbol{\mu}) = 0$$

iz čega kao procijenitelj najveće izglednosti slijedi

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}\,. \tag{43}$$

Maksimizacija izraza (42) po $\boldsymbol{\Sigma}$ nešto je složenija, no daje očekivan rezultat:

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\mathrm{ML}})^{\mathrm{T}}\,. \tag{44}$$

Procjenitelji $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ i $\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}$ analogni su procjeniteljima $\hat{\mu}_{\mathrm{ML}}$ odnosno $\hat{\sigma}^2_{\mathrm{ML}}$ univarijatne Gaussove razdiobe. Također vrijede ista zapažanja što se tiče nepristranosti procjenitelja: procjenitelj $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ je nepristran, dok procjenitelj $\hat{\boldsymbol{\Sigma}}_{\mathrm{ML}}$ podcjenuje i na malim uzorcima treba ga korigirati faktorom $N/(N-1)$.

Primjer 5 (Procjena parametara multivarijatne razdiobe) Raspolažemo uzorkom $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{8}$ za koji znamo da potječe iz multivarijatne normalne razdiobe:

$$\begin{array}{llll}
\mathbf{x}^{(1)} = & (9.59, -0.75, \quad 0.60) & \mathbf{x}^{(5)} = & (2.24, \quad 0.02, -4.72) \\
\mathbf{x}^{(2)} = & (2.30, \quad 0.37, -2.90) & \mathbf{x}^{(6)} = & (6.59, -0.20, -0.69) \\
\mathbf{x}^{(3)} = & (8.87, -0.84, -0.28) & \mathbf{x}^{(7)} = & (3.69, \quad 0.35, -1.84) \\
\mathbf{x}^{(4)} = & (3.99, \quad 1.92, -0.13) & \mathbf{x}^{(8)} = & (3.10, \quad 1.29, -0.59)
\end{array}$$

Prema (43), ML-procjena vektora srednje vrijednosti je

$$\hat{\boldsymbol{\mu}} = \frac{1}{8}\sum_{i=1}^{8}\mathbf{x}^{(i)} \approx (5.05, 0.27, -1.32)\,.$$

Prema (44), ML-procjena kovarijacijske matrice je

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{8}\sum_{i=1}^{8}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^{\mathrm{T}} \approx \begin{pmatrix} 7.47 & -1.63 & 3.20 \\ -1.63 & 0.79 & -0.01 \\ 3.20 & -0.01 & 2.68 \end{pmatrix}.$$

Kovarijacijska matrica nam govori da je varijanca najveća za prvu, najmanja za drugu varijablu. Prema (17), koeficijenti korelacije između varijabli su

$$\begin{aligned}
\hat{\rho}_{X_1,X_2} &= \frac{\mathrm{Cov}(X_1,X_2)}{\sigma_{X_1}\sigma_{X_2}} = \frac{-1.63}{\sqrt{7.47 \times 0.79}} \approx -0.670 \\
\hat{\rho}_{X_1,X_3} &= \frac{\mathrm{Cov}(X_1,X_3)}{\sigma_{X_1}\sigma_{X_3}} = \frac{3.20}{\sqrt{7.47 \times 2.68}} \approx 0.715 \\
\hat{\rho}_{X_2,X_3} &= \frac{\mathrm{Cov}(X_2,X_3)}{\sigma_{X_2}\sigma_{X_3}} = \frac{-0.01}{\sqrt{0.79 \times 2.68}} \approx 0.007
\end{aligned}$$

Vidimo da su varijable X_1 i X_2 negativno korelirane, varijable X_2 i X_3 su pozitivno korelirane, dok varijable X_2 i X_3 gotovo da i nisu korelirane. Budući da su ovi koeficijenti računati na temelju uzorka, oni su također procjene i mogu varirati u ovisnosti o uzorku.

5 Bayesovski procjenitelj

Procjenitelj najveće izglednost najjednostavniji je način procjene parametara razdiobe. Međutim, njegov je najveći nedostatak sklonost prenaučenosti. Na primjer, ako je varijabla X Bernoullijeva, i ako se u uzorku nikada nije realizirala, najizglednija procjena za $\hat{\mu}_{\mathrm{ML}} = P(X = 1)$ jest 0. U mnogim situacijama ne možemo biti zadovoljni s takvom procjenom jer iskustveno znamo da ona nije realna i da je problem naprosto u tome što s uzorkom "nismo imali sreće". Vidjet ćemo također da procjena koja rezultira ništicom može biti loša za rad klasifikatora (npr. naivnog Bayesovog).

Uobičajen način rješavanja ovog problema jest uporaba **zaglađenih procjena** (engl. *smoothed estimates*). Zaglađena procjena preraspoređuje (zaglađuje) ukupnu masu vjerojatnosti tako da vrijednosti parametara koje bi inače imale vjerojatnost nula dobiju neku malu vjerojatnost (i obrnuto: da vrijednosti koje bi inače imale vjerojatnost 1 dobiju nešto manju vjerojatnost). Na taj se način smanjuje prenaučenost modela.

5.1 Bayesovski i frekventistički pristup

Teorijski okvir zaglađene procjene parametara daje **bayesovska statistika** (engl. *Bayesian statistics*). Bayesovska statistika koristi teoriju vjerojatnosti kao alat za modeliranje nesigurnosti znanja. Prema bayesovskom shvaćanju teorije vjerojatnosti, jednakost $P(X = x) = 0.7$ odgovara izjavi *"vjerujem da $X = x$ s vjerojatnošću 0.7"*. Suprotno bayesovskom jest **frekventističko** shvaćanje teorije vjerojatnosti, koje vjerojatnost tumači kao vjerojatnost ishoda kod ponavljanja eksperimenata, odnosno kao relativnu frekvenciju. Prema frekventističkom shvaćanju, jednakost $P(X = x) = 0.7$ odgovara izjavi *"u 70% slučajeva vrijedi $X = x$"*.

U nekim slučajevima frekventističko tumačenje intuitivnije je od bayesovskog, dok je u nekim slučajevim obrnuto. Npr., vjerojatnost da bacajući dvije kocke dobijemo dvije šestice intuitivno je lakše tumačiti frekventistički, budući da je eksperiment moguće ponavljati. S druge strane, vjerojatnost da je neki planet naseljen inteligentnim bićima intuitivno je lakše tumačiti bayesovski, budući da tako izražavamo nesigurnost našeg znanja, a sâm eksperiment uopće nije ponovljiv.

5.2 Bayesovski procjenitelj

Ideja kod bayesovskog procjenitelja jest kombinirati apriorno znanje o mogućim vrijednostima parametra θ sa znanjem koje proizlazi iz uzorka. (Kod metode najveće izglednosti tako nešto nije moguće – ondje koristimo samo znanje koje proizlazi iz uzorka.) Ako, prije nego što uopće pogledamo uzorak, imamo neko apriorno znanje o distribuciji parametara, to je znanje vrlo korisno i treba ga iskoristiti. To je osobito slučaj ako je uzorak malen, jer tada ML-procjenitelj lako može prenaučiti model.

Naravno, točna vrijednost parametra θ nije nam poznata (inače uopće ne bismo trebali raditi procjenu na temelju uzorka), ali imamo neku predodžbu o mogućim vrijednostima tog parametra. To znanje modeliramo kao "nesigurnost", na način da parametar θ tretiramo kao slučajnu varijablu i definiramo njezinu distribuciju, $p(\theta)$. Npr., možemo reći da je vrlo vjerojatno da $\mu = 0.5$, ali da je manje vjerojatno da $\mu = 0$, što možemo modelirati primjerice kao $p(\mu = 0.5) = 0.8$ i $p(\mu = 0) = 0.1$. Primijetite da se ovdje zapravo radi o distribuciji nad varijablom koja je parametar druge distribucije, odnosno o *distribuciji distribucije*.

Sada kombiniramo apriornu razdiobu $p(\theta)$ s izglednošću $p(\mathcal{D}|\theta)$ koju smo dobili na temelju uzorka (skupa primjera) $\mathcal{D}$, koristeći Bayesovo pravilo:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta')p(\theta')\mathrm{d}\theta'} \tag{45}$$

čime smo dobili aposteriornu vjerojatnost za parametar θ za dani uzorak $\mathcal{D}$. Bayesovski procjenitelj definiran je kao očekivanje te vjerojatnosti:

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{D}] = \int \theta\, p(\theta|\mathcal{D})\mathrm{d}\theta\,. \tag{46}$$

Ovaj integral može biti teško izračunljiv, ovisno o obliku apriorne distribucije $p(\theta)$. Ovisno o izboru te distribucije, umnožak $p(\mathcal{D}|\theta)p(\theta)$ bit će u različitim oblicima. Najbolji oblik za $p(\theta)$ – za koji je integral analitički izračunljiv – jest onaj koji je istog oblika kao aposteriorna gustoća $p(\theta|\mathcal{D})$. Ako je to ispunjeno, distribucije $p(\theta)$ i $p(\theta|\mathcal{D})$ zovemo **konjugatnim distribucijama**. Nadalje, distribuciju $p(\theta)$ zovemo **konjugatnom apriornom distribucijom** (engl. *conjugate prior*) za izglednost $p(\mathcal{D}|\theta)$.

Sve razdiobe koje smo razmatrali pripadaju tzv. **eksponencijalnoj familiji distribucija** i imaju konjugatne apriorne distribucije. Konkretno, konjugatna apriorna distribucija za Gaussovu distribuciju je i sama Gaussova distribucija, dok je za multinomijalnu distribuciju konjugatna apriorna distibucija **Dirichletova distribucija**.

5.3 Laplaceovo zaglađivanje

U nastavku ćemo se usredotočiti na bayesovski procjenitelj za multinomijalnu varijablu, poznat pod nazivom **Laplaceovo zaglađivanje** (engl. *Laplace smoothing*). Parametar multinomijalne razdiobe, definirane s (23), jest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. Za apriornu razdiobu parametra $p(\boldsymbol{\mu})$ koristimo Dirichletovu razdiobu, konjugatnu apriornu distribuciju za multinomijalnu distribuciju. Dirichletova razdioba

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \mathrm{Dir}(\mu_1, \ldots, \mu_K|\alpha_1, \ldots, \alpha_k)$$

ima vektor parametara $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ i također je zvonolika (konkretan izraz za Dirichletovu distribuciju ovdje nas ne zanima). Parametri $\boldsymbol{\alpha}$ određuju oblik Dirichletove razdiobe i zapravo su hiperparametri našeg modela.

Uvrštavanje Dirichletove razdiobe u (45) i izračunavanjem očekivanja (46), za zaglađeni procjenitelj parametra μ_k naposlijetku dobivamo:

$$\hat{\mu}_{k,\text{Bayes}} = \frac{N_k + \alpha_k}{N + \sum_{i=1}^{K} \alpha_i}\,.$$

Usporedimo ovaj izraz s izrazom (39) za ML-procjenitelj. Vrijednost N_k/N je relativna frekvencija koju, kao i kod ML-procjenitelja, izračunavamo na temelju uzorka. Međutim, bayesovski procjenitelj dodatno kombinira relativnu frekvenciju s apriornim znanjem. Interpretacija bayesovskog procjenitelja je sljedeća: prije nego što smo opazili uzorak $\mathcal{D}$, "virtualno" smo opazili uzorak veličine $\sum_{i=1}^{K}$, u kojemu se slučajna varijabla realizirala s vrijednošću k ukupno α_k puta. Za virtualni uzorak veličine 0, bayesovski procjenitelj degradira na ML-procjenitelj.

Možemo pojednostaviti Dirichletovu apriornu distribuciju na način da pretpostavimo $\alpha_i = \lambda$ za svaki $i = 1, \ldots, K$, tj. da koristimo apriornu distribuciju $\mathrm{Dir}(\mu_1, \ldots, \mu_K|\lambda, \ldots, \lambda)$. Tako dobivamo **Laplaceov (Lidstonov) procjenitelj**:

$$\hat{\mu}_{k,\text{Bayes}} = \frac{N_k + \lambda}{N + K\lambda}\,.$$

Odabirom različitih vrijednosti za λ dobivamo različite procjenitelje. Najčešće se uzima:

- Laplaceovo pravilo ili *add-one-rule* ($\lambda = 1$): $\hat{\mu}_{k,\text{Bayes}} = \frac{N_k+1}{N+K}$,
- Jeffreys-Perksovo pravilo ($\lambda = 1/2$): $\hat{\mu}_{k,\text{Bayes}} = \frac{N_k+1/2}{N+K/2}$,
- Schurmann-Grassbergerovo pravilo ($\lambda = 1/K$): $\hat{\mu}_{k,\text{Bayes}} = \frac{N_k+1/K}{N+1}$.

Primjer 6 (Laplaceov procjenitelj) Neka je X multinomijalna varijabla s $K = 5$ mogućih stanja, $x \in \{0, 1, 2, 3, 4\}$. Iz ove razdiobe dobiven je uzorak veličine $N = 16$:

$$\mathbf{X} = (0, 1, 2, 0, 4, 1, 0, 0, 2, 2, 1, 4, 2, 1, 0, 2)\,.$$

Prema (39), ML-procjena parametra $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)$ je

$$\hat{\boldsymbol{\mu}}_{\text{ML}} \approx (0.3125, 0.25, 0.3125, 0.0, 0.125)\,.$$

Dobivena procjena je prenaučena: vjerojatnost ishoda $X = 3$ jednaka je nuli zbog toga što se u uzorku varijabla X nikada nije realizirala s vrijednošću d. Zaglađeni procjenitelji daju drugačije procjene:

Laplace:	$\hat{\boldsymbol{\mu}} \approx (0.286, 0.238, 0.286, 0.048, 0.142)$
Jeffreys-Perks:	$\hat{\boldsymbol{\mu}} \approx (0.298, 0.243, 0.298, 0.027, 0.135)$
Schurmann-Grassberger:	$\hat{\boldsymbol{\mu}} \approx (0.306, 0.247, 0.306, 0.012, 0.129)$

Iz primjera je vidljivo da je Schurmann-Grassbergerov procjenitelj "najkonzervativniji" u smislu da događaju koji se u uzorku nije realizirao dodjeljuje najmanje vjerojatnosne mase.