

13. Procjena parametara

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.3

1 Motivacija

- **Probabilistički modeli** modeliraju vjerojatnosnu razdiobu primjera \mathbf{x} i/ili oznaka y
- **Prednosti:** (1) temeljeni na teoriji vjerojatnosti, (2) vjerojatnosna predikcija, (3) ugradnja apriornog znanja, (4) prikladni za male skupove podataka
- Npr. **Bayesov klasifikator** – $P(y|\mathbf{x}) \propto P(\mathbf{x}|y)P(y)$
 - odabrati prikladne razdiobe za $P(\mathbf{x}|y)$ i $P(y)$
 - **procijeniti parametre** razdioba na temelju podataka \Leftrightarrow učenje modela

2 Slučajne varijable

- X – slučajna varijabla (s.v.) sa skupom mogućih vrijednosti $\{x_i\}$
- **Diskretna s.v.:**
 - $P(X = x)$, kraće $P(x)$ – **vjerojatnost** da diskretna s.v. poprimi vrijednost x
 - $P(x_i) \geq 0, \sum_i P(x_i) = 1 \Rightarrow$ **diskretna razdioba (distribucija) vjerojatnosti**
- **Kontinuirana s.v.:**
 - $p(x)$ – **funkcija gustoće vjerojatnosti (PDF)**
 - $p(x) \geq 0, \int_{-\infty}^{\infty} p(x) dx = 1 \Rightarrow$ **kontinuirana razdioba (distribucija) vjerojatnosti**
- **Očekivanje** – prosječna vrijednost: $\mathbb{E}[X] = \sum_x xP(x)$ odnosno $\mathbb{E}[X] = \int_{-\infty}^{\infty} x p(x) dx$
- **Varijanca** – očekivano odstupanje od očekivanja: $\text{Var}(X) = \sigma_X^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- **Kovarijanca** – zajednička varijabilnost dviju varijabli:

$$\text{Cov}(X, Y) = \sigma_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$$\Rightarrow \text{Cov}(X, Y) = \text{Cov}(Y, X), \text{Cov}(X, X) = \text{Var}(X)$$

- **Pearsonov koeficijent korelacijske**: $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \in [-1, +1]$

- $\rho_{X,Y} = +1 \Leftrightarrow$ pozitivna **linearna zavisnost**
- $\rho_{X,Y} = 0 \Leftrightarrow$ **linearna nezavisnost**
- $\rho_{X,Y} = -1 \Leftrightarrow$ negativna **linearna zavisnost**

\Rightarrow ne mjeri **nelinearnu zavisnost!**

- **Matrica kovarijacije** – kovarijacija svih parova varijabli **slučajnog vektora** (X_1, \dots, X_n) :

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \dots & \text{Var}(X_n) \end{pmatrix} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]$$

- simetrična i pozitivno semidefinitna
- singularna (tj. nema inverz) ako postoje linearno zavisni retci ili ako $\sigma_i^2 = 0$
- $\text{Cov}(X_i, X_j) = 0 \Rightarrow$ **dijagonalna** kovarijacijska matrica, $\Sigma = \text{diag}(\sigma_i^2)$
- $\sigma_i^2 = \sigma \Rightarrow$ **izotropna** kovarijacijska matrica, $\Sigma = \sigma^2 \mathbf{I}$

- **Nezavisne varijable** $\Leftrightarrow P(X, Y) = P(X)P(Y)$

- nezavisne varijable su nekorelirane, $\text{Cov}(X, Y) = \rho_{X,Y} = 0$, ali obrat ne vrijedi!

3 Osnovne vjerojatnosne distribucije

- Diskretna varijabla:
 - Jednodimenzija binarna: **Bernoullijeva razdioba**
 - Jednodimenzija viševrijednosna: **kategorička (multinulijeva) razdioba**
 - Višedimenzija: Konkatenirani vektor binarnih/viševrijednosnih varijabli
- Kontinuirana varijable:
 - Jednodimenzija: **univariatna normalna (Gaussova) razdioba**
 - Višedimenzija: **multivariatna normalna (Gaussova) razdioba**
- **Bernoullijeva razdioba** – binarna s.v.:

$$P(x|\mu) = \mu^x(1-\mu)^{1-x}$$

$$\Rightarrow \mathbb{E}[X] = \mu, \text{Var}(X) = \mu(1-\mu)$$

- **Kategorička (multinulijeva) razdioba** – viševrijednosna diskretna s.v.:

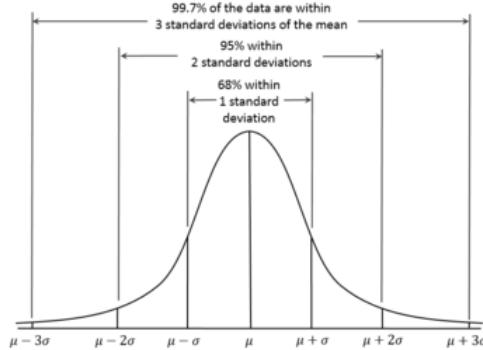
$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

- $\mathbf{x} = (x_1, x_2, \dots, x_K)^T$ – vektor indikatorskih varijabli (**one-hot encoding**)
- $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ – vjerojatnosti pojedinih vrijednosti, $\sum_k \mu_k = 1, \mu_k \geq 0$

- **Normalna (Gaussova) razdioba** – kontinuirana vrijednost uz prisustvo **šuma**:

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

$$\Rightarrow \mathbb{E}[X] = \mu, \text{Var}(X) = \sigma^2$$

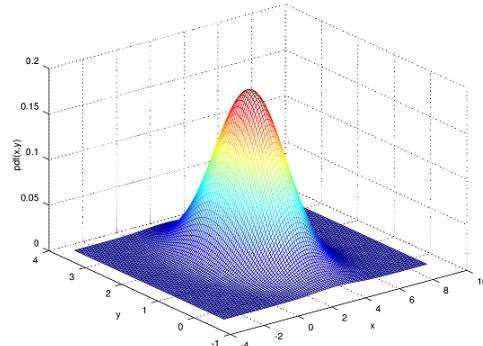


- **Multivariatna (višedimenzijska) normalna (Gaussova) razdioba**:

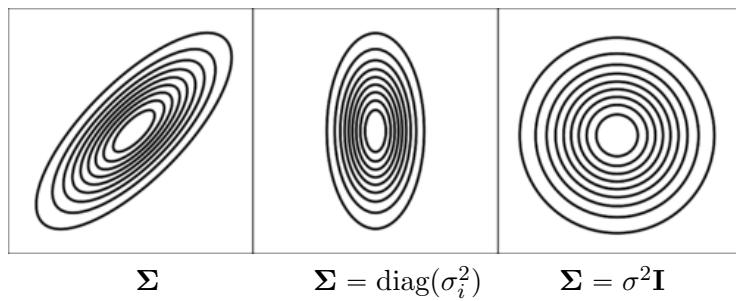
$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

$$\Rightarrow \mathbb{E}[X] = \boldsymbol{\mu}, \text{Cov}(X_i, X_j) = \boldsymbol{\Sigma}_{ij}$$

- značajke su savršeno **multikolinearne** $\Leftrightarrow \boldsymbol{\Sigma}$ je singularna $\Leftrightarrow p(\mathbf{x})$ je nedefinirana
- $\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ – **kvadratna forma**
- Δ – **Mahalanobisova udaljenost** između \mathbf{x} i $\boldsymbol{\mu}$

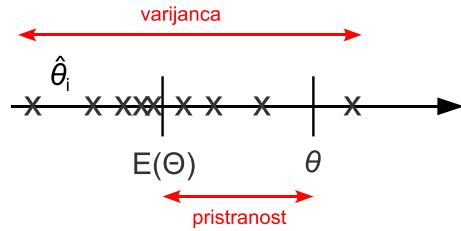


- Kovariacijska matrica određuje izgled Gaussove razdiobe:



4 Procjena parametara

- Raspolažemo konačnim i slučajnim (= reprezentativnim) uzorkom iz **populacije**
- Na temelju uzorka **procjenjujemo parametre** modela koji opisuje populaciju
- **Uzorak** – niz s.v. (X_1, X_2, \dots, X_N) koje su **iid** (identično i nezavisno distribuirane)
- **Statistika** – funkcija slučajnog uzorka, $\Theta = g(X_1, X_2, \dots, X_N)$
- **Procjenitelj (estimator)** – statistika koja odgovara parametru populacije θ
- **Procjena (estimacija)** – vrijednost procjenitelja za dani uzorak, $\hat{\theta} = g(x_1, x_2, \dots, x_N)$
- Procjenitelj je s.v., pa ima svoje očekivanje i varijancu



- **Pristranost (bias)** – razlika između očekivanja procjenitelja i parametra populacije:

$$b(\Theta) = \mathbb{E}[\Theta] - \theta$$

- **Nepristran procjenitelj (unbiased estimator)** $\Leftrightarrow \mathbb{E}[\Theta] = \theta \Leftrightarrow b(\Theta) = 0$

- Primjeri:

- $\hat{\mu} = \frac{1}{N} \sum_i x^{(i)}$ – nepristran procjenitelj srednje vrijednosti μ
- $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \hat{\mu})^2$ – pristran procjenitelj varijance σ^2 (podcjenjuje!)
- $\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x^{(i)} - \hat{\mu})^2$ – nepristran procjenitelj varijance σ^2

- Postupci za izvođenje procjenitelja:

- **Procjenitelj najveće izglednosti** (maximum likelihood estimator, **MLE**)
- **Procjenitelj maximum a posteriori (MAP)**
- **Bayesovski procjenitelj** (bayesian estimator)

- Radit ćemo MLE i MAP

14. Procjena parametara II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.5

1 Funkcija izglednosti

- Skup podataka $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ koji su **iid**; prepostavka: $\mathbf{x}^{(i)} \sim p(\mathbf{x}|\boldsymbol{\theta})$
- Vjerojatnost uzorka:

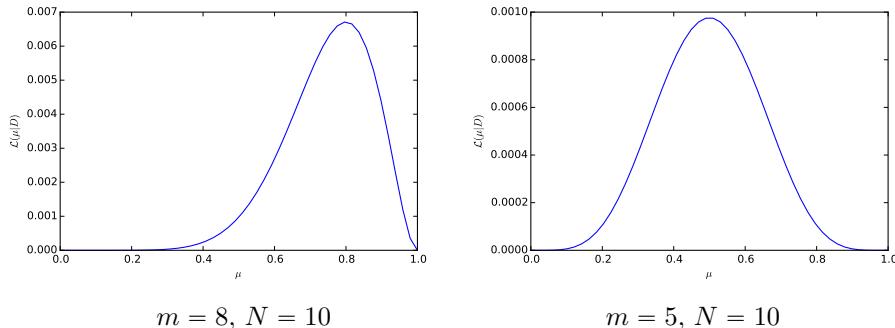
$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}|\boldsymbol{\theta}) = \prod_{i=1}^N p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \equiv \mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$$

gdje je $\mathcal{L} : \boldsymbol{\theta} \mapsto p(\mathcal{D}|\boldsymbol{\theta})$ **funkcija izglednosti**

- $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$ – vjerojatnost realizacije uzorka \mathcal{D} , ako je parametar populacije jednak $\boldsymbol{\theta}$
- Npr. funkcija izglednosti Bernoullijske varijable – m pozitivnih ishoda u N pokusa:

$$\mathcal{L}(\mu|\mathcal{D}) = P(\mathcal{D}|\mu) = P(x^{(1)}, \dots, x^{(N)}|\mu) = \prod_{i=1}^N P(x^{(i)}|\mu) = \mu^m(1-\mu)^{(N-m)}$$

gdje $m = \sum_i x^{(i)}$



2 Procjenitelj MLE

- Prepostavka: uzorak \mathcal{D} je **najvjerojatniji mogući**, inače ne bi bio izvučen
- Najbolja procjena za $\boldsymbol{\theta}$ je ona koja maksimizira izglednost $\mathcal{L}(\boldsymbol{\theta}|\mathcal{D})$

- **Procjenitelj najveće izglednosti** (*maximum likelihood estimator*) – **MLE**:

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

- Radi matematičke jednostavnosti, maksimizirat ćemo **log-izglednost**:

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

- MLE za parametar Bernoullijeve razdiobe:

$$\begin{aligned} \ln \mathcal{L}(\mu | \mathcal{D}) &= \ln \prod_{i=1}^N \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}} = \sum_{i=1}^N x^{(i)} \ln \mu + \left(N - \sum_{i=1}^N x^{(i)} \right) \ln(1-\mu) \\ \frac{\partial \ln \mathcal{L}}{\partial \mu} &= \frac{1}{\mu} \sum_{i=1}^N x^{(i)} - \frac{1}{1-\mu} \left(N - \sum_{i=1}^N x^{(i)} \right) = 0 \\ \Rightarrow \hat{\mu}_{\text{MLE}} &= \frac{1}{N} \sum_{i=1}^N x^{(i)} = \frac{m}{N} \end{aligned}$$

\Rightarrow **relativna frekvencija** (udio realizacije $x = 1$)

- MLE za parametre kategorijalne razdiobe:

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

\Rightarrow maksimizacijom po μ_k uz $\sum_{k=1}^K \mu_k = 1$ metodom **Lagrangeovih multiplikatora**:

$$\hat{\mu}_{k,\text{MLE}} = \frac{1}{N} \sum_{i=1}^N x_k^{(i)} = \frac{N_k}{N}$$

\Rightarrow relativna frekvencija k -te vrijednosti kategorijalne varijable

- MLE za parametre normalne razdiobe:

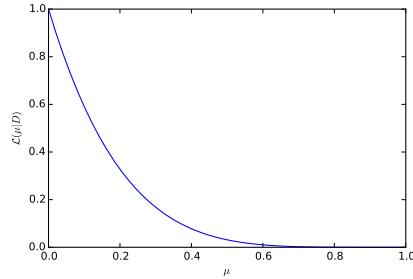
$$\begin{aligned} \ln \mathcal{L}(\mu, \sigma^2 | \mathcal{D}) &= \ln \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{\sum_i (x^{(i)} - \mu)^2}{2\sigma^2} \\ \frac{\partial \ln \mathcal{L}}{\partial \mu} &= 0 \Rightarrow \hat{\mu}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N x^{(i)} \\ \frac{\partial \ln \mathcal{L}}{\partial \sigma^2} &= 0 \Rightarrow \hat{\sigma}_{\text{MLE}}^2 = \frac{1}{N} \sum_{i=1}^N (x^{(i)} - \hat{\mu}_{\text{MLE}})^2 \end{aligned}$$

\Rightarrow procjenitelj varijance je pristran (za malen N preporuča ga se korigirati)

- MLE za parametre multivarijatne normalne razdiobe:

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathcal{D}) &= \ln \prod_{i=1}^N p(\mathbf{x}^{(i)} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= -\frac{nN}{2} \ln(2\pi) - \frac{N}{2} |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu}) \\
\nabla_{\boldsymbol{\mu}} \ln \mathcal{L} = 0 &\Rightarrow \hat{\boldsymbol{\mu}}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} \\
\nabla_{\boldsymbol{\Sigma}} \ln \mathcal{L} = 0 &\Rightarrow \hat{\boldsymbol{\Sigma}}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{MLE}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_{\text{MLE}})^T
\end{aligned}$$

- MLE smo već koristili kod izvoda funkcije gubitka za poopćene linearne modele
- Minimizacija empirijske pogreške \Leftrightarrow MLE procjena za \mathbf{w} uz odgovarajuću $p(y|\mathbf{x})$
- MLE je sklon **prenaučenosti** – osobito problematično kada je N malen
- Npr., bacanje novčića (Bernoullijeva varijabla): $m = 0, N = 5 \Rightarrow \hat{\mu}_{\text{MLE}} = 0$:



3 Procjenitelj MAP

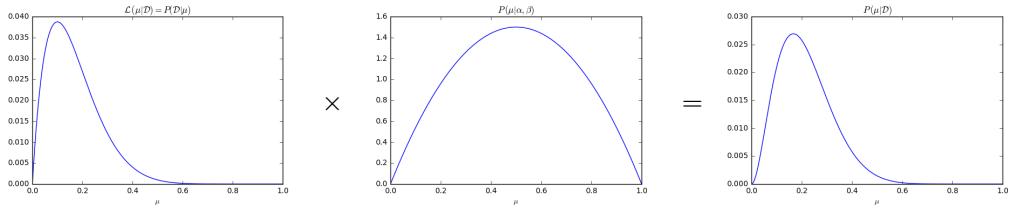
- Želimo kombinirati informacije iz podataka (izglednost $\boldsymbol{\theta}$) s **apriornim znanjem** o $\boldsymbol{\theta}$
- $p(\boldsymbol{\theta})$ – **apriorna razdioba parametra $\boldsymbol{\theta}$** (*parameter prior*)
- **Aposteriorna vjerojatnost parametra $\boldsymbol{\theta}$** (Bayesov teorem):

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\mathcal{D} | \boldsymbol{\theta}) P(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- Procjenitelj **maksimum a posteriori (MAP)**:

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

- Izglednost \times Prior \propto Posterior:



- Rješivo **analitički**, ako $p(\mathcal{D}|\theta)$ i $p(\theta)$ odaberemo tako da daju neku standardnu $p(\theta|\mathcal{D})$
- **Konjugatne distribucije** $\Leftrightarrow p(\theta|\mathcal{D})$ i $p(\theta)$ su iste vrste distribucija
- $p(\theta)$ je **konjugatna apriorna distribucija** za $p(\mathcal{D}|\theta) \Rightarrow p(\theta|\mathcal{D})$ i $p(\theta)$ su konjugatne
- Svaka $p(\mathcal{D}|\theta)$ iz **eksponečijalne familije** ima svoju konjugatnu apriornu distribuciju:
 - $p(\mathcal{D}|\theta)$ Bernoullijeva $\Rightarrow p(\theta)$ beta
 - $p(\mathcal{D}|\theta)$ kategorijkska $\Rightarrow p(\theta)$ Dirichletova
 - $p(\mathcal{D}|\theta)$ normalna $\Rightarrow p(\theta)$ normalna
 - $p(\mathcal{D}|\theta)$ multiv. normalna $\Rightarrow p(\theta)$ multiv. normalna

4 Beta-Bernoullijev model

- Konjugatna apriorna distr. za izglednost Bernoullijeve varijable je **beta-distribucija**:

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

gdje beta-funkcija B služi za normalizaciju, te $\alpha > 0$ i $\beta > 0$

 - $\alpha = \beta = 1 \Leftrightarrow$ uniformna distribucija \Rightarrow **neinformativna apriorna distribucija**
 - $\alpha > 1, \beta > 1 \Rightarrow$ veća gustoća vjerojatnosti za $\mu = 0.5$
 - $\alpha > \beta \Rightarrow$ veća gustoća vjerojatnosti za $\mu \in (0.5, 1)$
 - $\alpha < \beta \Rightarrow$ veća gustoća vjerojatnosti za $\mu \in (0, 0.5)$
- Maksimizator (mod) beta-distribucije: $\frac{\alpha-1}{\alpha+\beta-2}$ (za $\alpha, \beta > 1$)
- Aposteriorna beta-distribucija:

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

gdje $\alpha' = m + \alpha$ i $\beta' = N - m + \beta$

- MAP-procjenitelj odgovara modu aposteriorne beta-distribucije:

$$\hat{\mu}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{m + \alpha - 1}{\alpha + N + \beta - 2}$$

- Za $N \rightarrow \infty$ procjenom dominiraju podatci; za $\alpha = \beta = 1$ MAP degenerira u MLE
- MAP provodi **zaglađivanje** (*smoothing*) – preraspoređivanje mase vjerojatnosti
- **Laplaceovo zaglađivanje (Laplace smoothing)** – MAP sa $\alpha = \beta = 2$:

$$\hat{\mu}_{\text{MAP}} = \frac{m + 1}{N + 2}$$

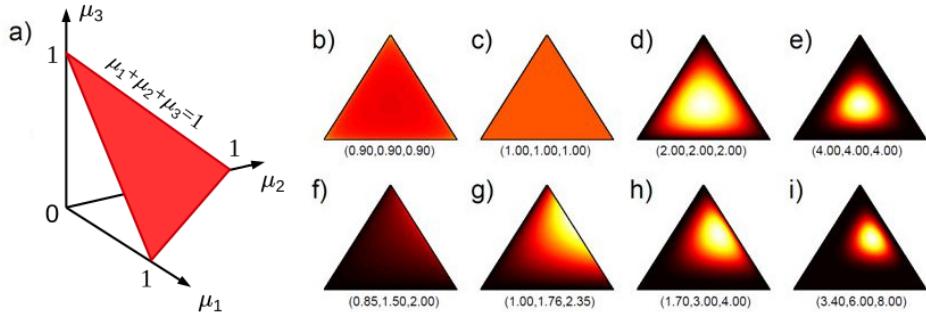
5 Dirichlet-kategorijski model

- Konjugatna apriorna distr. za multinomijalnu izglednost je **Dirichletova distribucija**

$$P(\boldsymbol{\mu}|\boldsymbol{\alpha}) = P(\mu_1, \dots, \mu_K | \alpha_1, \dots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^K \mu_k^{\alpha_k - 1}$$

gdje beta-funkcija B služi za normalizaciju, te $\alpha_k > 0$

- Dirichletova distribucija je poopćenje beta-distribucije na K varijabli
- μ_k leže na $(K - 1)$ -dimenzijskom **standardnom simpleksu**, tj. $\sum_{k=1}^K \mu_k = 1$ i $\mu_k \geq 0$
- Npr., za $K = 3$, to je trokut u trodimenzijskome prostoru:



- MAP-procjenitelj odgovara modu Dirichletove distribucije:

$$\hat{\mu}_{k,\text{MAP}} = \frac{\alpha'_k - 1}{\sum_{k=1}^K \alpha'_k - K}$$

gdje $\alpha'_k = N_k + \alpha_k$ i $N_k = \sum_i x_k^{(i)}$ (broj nastupanja k -te vrijednosti)

- Uz $\alpha_k = 2$, najvjerojatnija je uniformna distribucija po $\boldsymbol{\mu}$, a procjenitelj je:

$$\hat{\mu}_{k,\text{MAP}} = \frac{N_k + 1}{N + K}$$

15. Bayesov klasifikator

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.4

1 Pravila vjerojatnosti

- **Pravilo zbroja:**

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

⇒ **marginalna vjerojatnost** iz **zajedničke vjerojatnosti** (*joint*)

- **Pravilo umnoška:**

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

- Dva pravila izvedena iz pravila umnoška:

- **Bayesovo pravilo:**

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

- **Pravilo lanca** (*chain rule*):

$$\begin{aligned} P(x_1, \dots, x_n) &= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \cdots P(x_n|x_1, \dots, x_{n-1}) \\ &= \prod_{k=1}^n P(x_k|x_1, \dots, x_{k-1}) \end{aligned}$$

⇒ **faktorizacija** zajedničke vjerojatnosti na umnožak **faktora**

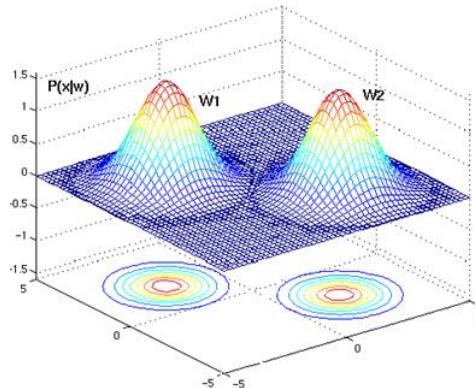
2 Bayesov klasifikator

- Model Bayesovog klasifikatora:

$$h_j(\mathbf{x}; \boldsymbol{\theta}) = P(y = j|\mathbf{x}) = \frac{p(\mathbf{x}|y = j)P(y = j)}{\sum_k p(\mathbf{x}|y = k)P(y = k)}$$

- $P(y|\mathbf{x})$ – **aposteriorna vjerojatnost** (*posterior*) klase za zadani primjer
 - $p(\mathbf{x}|y)$ – **izglednost klase** (*class likelihood*) – vjerojatnost primjera u klasi
 - $P(y)$ – **apriorna vjerojatnost klase** (*class prior*)

- Primjer: binarna klasifikacija s Gaussovim gustoćama za izglednosti klasa:



- Faktorizacija $p(\mathbf{x}, y)$ na $p(\mathbf{x}|y)P(y)$ omogućava modeliranje složenih distribucija
- Klasifikacija u najvjerojatniju klasu (**MAP-hipoteza**):

$$h(\mathbf{x}) = \operatorname{argmax}_j p(\mathbf{x}|y=j)P(y=j)$$

- Bayesov klasifikator – **parametarski** i **generativni** model

3 Generativni modeli

- Modeli modeliraju **zajedničku distribuciju** $p(\mathbf{x}, y)$
- Na temelju $p(\mathbf{x}, y)$ računamo $p(y|\mathbf{x})$ ili neku drugu distribuciju od interesa
- Modeliraju **nastajanje podataka** $\{(\mathbf{x}^{(i)}, y^{(i)})\}_i$ – tzv. **generativna priča**
- Generativna priča Bayesovog klasifikatora:

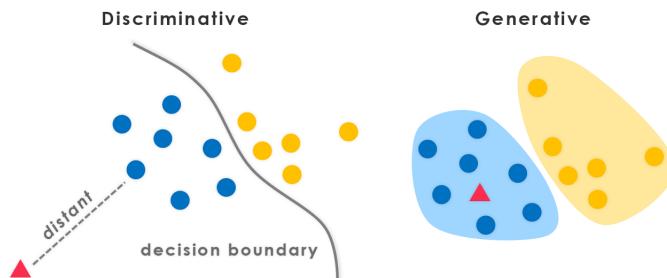
$$P(\mathbf{x}, y) = p(\mathbf{x}|y)P(y)$$

⇒ odabir oznake prema $P(y)$, zatim odabir primjera prema $P(\mathbf{x}|y)$

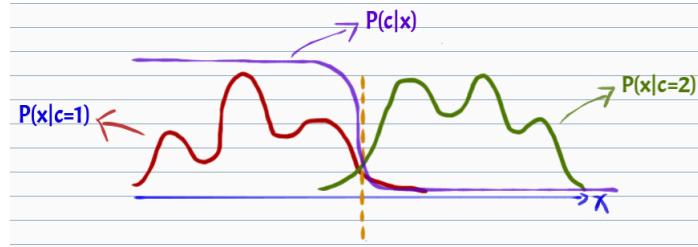
- Složeniji generativni modeli: **Bayesove mreže**, **HMM**, **GMM**, **LDA**
- Usp.: diskriminativni modeli izravno modeliraju $p(y|\mathbf{x})$; npr. logistička regresija:

$$h(\mathbf{x}; \mathbf{w}) = P(y|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

- Diskriminativno vs. generativno:

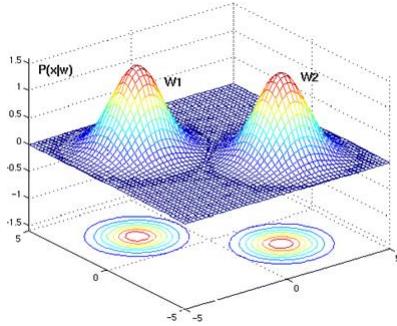


- Prednosti: laka ugradnja stručnog znanja, interpretabilnost/analiza rezultata
- Nedostatci: iziskuju mnogo primjera za učenje, nepotrebna složenost modeliranja
- Primjer: nepotrebna složenost modeliranja zajedničke vjerojatnosti:



4 Gaussov Bayesov klasifikator

- Izglednost klase modeliramo **Gaussovom (normalnom) razdiobom**: $\mathbf{x}|y \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $\boldsymbol{\mu}$ predstavlja **prototipni primjer**; primjeri odstupaju od prototipa uslijed **šuma**



- Jednodimenzionalni slučaj:

$$p(x|y = j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{(x - \mu_j)^2}{2\sigma_j^2} \right\}$$

- Model (**MAP-hipoteza**):

$$h(x) = \operatorname{argmax}_j p(x, y = j) = \operatorname{argmax}_j p(x|y = j)P(y = j)$$

- Model za klasu j :

$$h_j(x) = p(x, y = j) = p(x|y = j)P(y = j)$$

- Prelazak u logaritamsku domenu i uklanjanje konstanti:

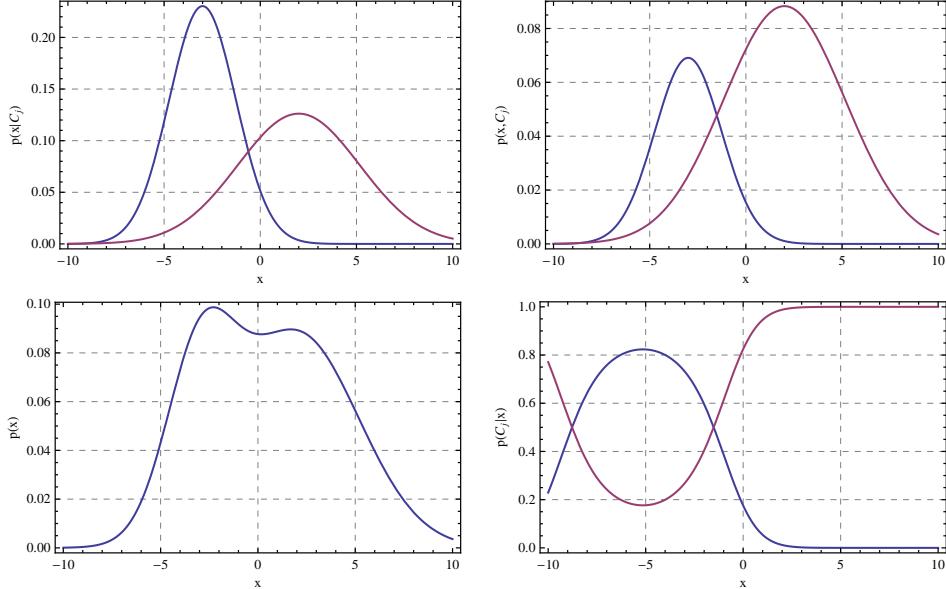
$$\begin{aligned} h_j(x) &= \ln p(x|y = j) + \ln P(y = j) \\ &= -\frac{1}{2} \ln 2\pi - \ln \sigma_j - \frac{(x - \mu_j)^2}{2\sigma_j^2} + \ln P(y = j) \end{aligned}$$

- MLE procjene parametara:

$$\begin{aligned}\hat{\mu}_j &= \frac{1}{N_j} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} x^{(i)} \\ \hat{\sigma}_j^2 &= \frac{1}{N_j} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} (x^{(i)} - \hat{\mu}_j)^2 \\ P(y = j) &= \hat{\mu}'_j = \frac{1}{N} \sum_{j=1}^N \mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N}\end{aligned}$$

- Primjer:

$$\begin{aligned}p(x|y=1) &\sim \mathcal{N}(-3, 3), P(y=1) = 0.3 \\ p(x|y=2) &\sim \mathcal{N}(2, 10), P(y=2) = 0.7\end{aligned}$$



- Više značajki \Rightarrow izglednosti modeliramo multivarijatnom normalnom razdiobom:

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

- Model za klasu j :

$$\begin{aligned}h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\ &= -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln P(y=j) \\ &\Rightarrow -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln P(y=j)\end{aligned}$$

- MLE procjene parametara:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_j &= \frac{1}{N_j} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} \mathbf{x}^{(i)} \\ \hat{\boldsymbol{\Sigma}}_j &= \frac{1}{N_j} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_j)(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}}_j)^T \\ \hat{\mu}_j &= \frac{1}{N} \sum_{j=1}^N \mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N}\end{aligned}$$

- Broj parametara: $\frac{n}{2}(n+1)K + K \cdot n + K - 1 \Rightarrow \mathcal{O}(n^2)$

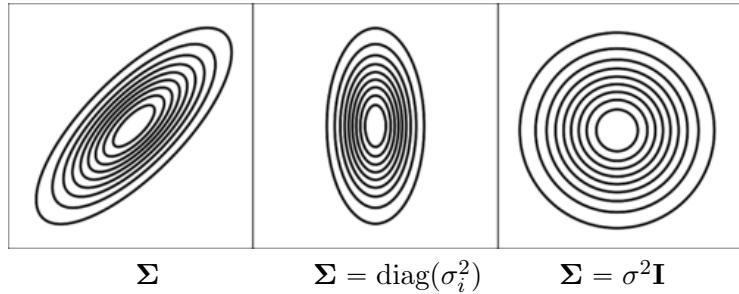
- Granica između dviju klasa: $h_1(\mathbf{x}) - h_2(\mathbf{x}) = 0$:

$$\begin{aligned}h_{12}(\mathbf{x}) &= h_1(\mathbf{x}) - h_2(\mathbf{x}) \\ &= -\frac{1}{2} \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} (\mathbf{x}^T \boldsymbol{\Sigma}_1^{-1} \mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1) + \ln P(y=1) \\ &\quad - \left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_2| - \frac{1}{2} (\mathbf{x}^T \boldsymbol{\Sigma}_2^{-1} \mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) + \ln P(y=2) \right) \\ &\quad \dots \mathbf{x}^T (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \mathbf{x} \dots\end{aligned}$$

\Rightarrow član koji kvadratno ovisi o $\mathbf{x} \Leftrightarrow$ **nelinearna granica**

5 Varijante Gaussovog Bayesovog klasifikatora

- Uvodimo prepostavke na $\boldsymbol{\Sigma}$ koje pojednostavljaju model



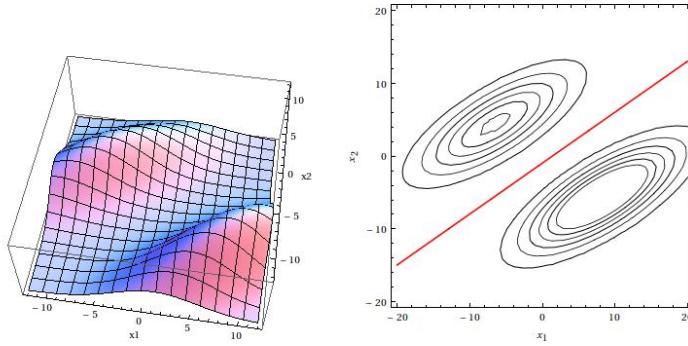
- **Dijeljena kovarijacijska matrica:** $\hat{\boldsymbol{\Sigma}} = \sum_j \hat{\mu}_j \hat{\boldsymbol{\Sigma}}_j$

- Model za klasu j :

$$\begin{aligned}h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\ &= -\frac{n}{2} \cancel{\ln 2\pi} - \frac{1}{2} \cancel{\ln |\boldsymbol{\Sigma}|} - \frac{1}{2} (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - 2\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j) + \ln P(y=j)\end{aligned}$$

- Model je linearan \Rightarrow **linearna granica** između klasa

- Broj parametara: $\frac{n}{2}(n+1) + nK + K - 1 \Rightarrow \mathcal{O}(n^2)$



- **Dijeljena i dijagonalna kovarijacijska matrica:** $\Sigma = \text{diag}(\sigma_i^2)$

- Vrijedi $|\Sigma| = \prod_i \sigma_i^2$ i $\Sigma^{-1} = \text{diag}(1/\sigma_i^2)$
- Izglednost klase:

$$\begin{aligned}
p(\mathbf{x}|y=j) &= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_j)^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right) \\
&= \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \sigma_i} \exp\left(-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2\right) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2} \left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2\right\} \\
&= \prod_{i=1}^n \mathcal{N}(\mu_{ij}, \sigma_i^2) = \prod_{i=1}^n p(x_i|y)
\end{aligned}$$

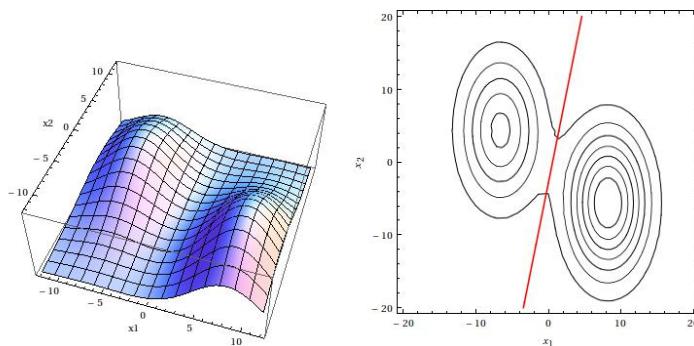
\Rightarrow **uvjetna nezavisnost** značajki \Rightarrow **Gaussov naivan Bayesov klasifikator**

- $x_k \perp x_j | y \Rightarrow \text{Cov}(x_k|y, x_j|y) = 0 \Leftrightarrow p(\mathbf{x}|y) = \prod_k p(x_k|y)$
- Model za klasu j :

$$\begin{aligned}
h_j(\mathbf{x}) &= \ln p(\mathbf{x}|y=j) + \ln P(y=j) \\
&= \sum_{i=1}^n \ln \frac{1}{\sqrt{2\pi}\sigma_i} + \sum_{i=1}^n \left(-\frac{1}{2} \left(\frac{x_i - \mu_{ij}}{\sigma_i}\right)^2\right) + \ln P(y=j)
\end{aligned}$$

\Rightarrow **normirana euklidska udaljenost** između primjera \mathbf{x} i prototipa klase $\boldsymbol{\mu}_j$

- Broj parametara: $n + n \cdot K + K - 1 \Rightarrow \mathcal{O}(n)$

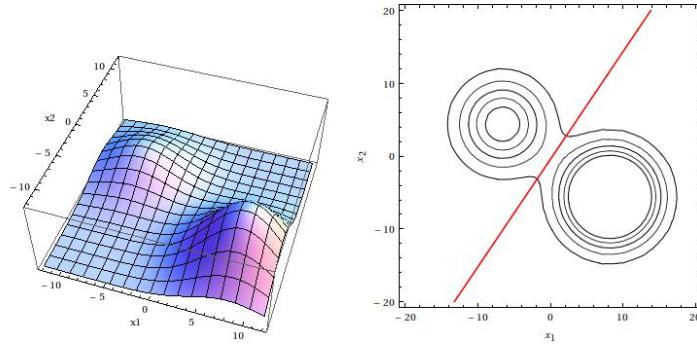


- Izotropna kovarijacijska matrica: $\Sigma = \sigma^2 \mathbf{I}$

– Model za klasu j :

$$h_j(\mathbf{x}) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_{ij})^2 + \ln P(y = j)$$

– Broj parametara: $1 + Kn + K - 1 \Rightarrow \mathcal{O}(n)$



- Druge varijante:

Pretpostavka	Kov. matrica	Broj parametara
Različite, hiperelipsoidi	Σ_j	$Kn(n+1)/2 + Kn$
Dijeljena, hiperelipsoidi	Σ	$n(n+1)/2 + Kn$
Različite, poravnati hiperelipsoidi	$\Sigma_j = \text{diag}(\sigma_{i,j}^2)$	$2Kn$
Dijeljena, poravnati hiperelipsoidi	$\Sigma = \text{diag}(\sigma_i^2)$	$n + Kn$
Različite, hipersfere	$\Sigma_j = \sigma_j^2 \mathbf{I}$	$K + Kn$
Dijeljena, hipersfere	$\Sigma = \sigma^2 \mathbf{I}$	$1 + Kn$

- Odabir modela: **unakrsnom provjerom**

16. Bayesov klasifikator II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.4

1 Bayesov klasifikator vs. logistička regresija

- Ideja: pokazati da logistička regresija i Bayesov klasifikator izračunavaju isti $P(y|\mathbf{x})$
- Model **logističke regresije**:

$$h(\mathbf{x}; \mathbf{w}) = P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

- Aposteriorna vjerojatnost za **kontinuirani Bayesov klasifikator** (za dvije klase):

$$\begin{aligned} P(y = 1|\mathbf{x}) &= \frac{p(\mathbf{x}|y = 1)P(y = 1)}{p(\mathbf{x}|y = 1)P(y = 1) + p(\mathbf{x}|y = 2)P(y = 2)} = \frac{1}{1 + \frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}} = \\ &= \frac{1}{1 + \exp\left(\ln \frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}\right)} = \frac{1}{1 + \exp(-\alpha)} = \sigma(\alpha) \end{aligned}$$

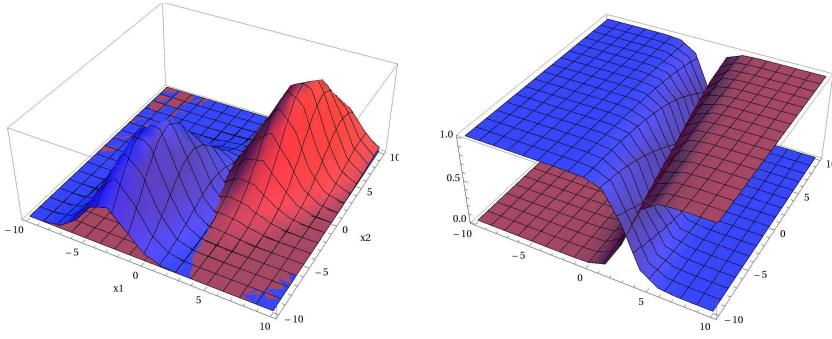
gdje

$$\alpha = \ln \frac{p(\mathbf{x}|y = 1)P(y = 1)}{p(\mathbf{x}|y = 2)P(y = 2)} = \underbrace{\ln p(\mathbf{x}|y = 1)P(y = 1)}_{h_1(\mathbf{x})} - \underbrace{\ln p(\mathbf{x}|y = 2)P(y = 2)}_{h_2(\mathbf{x})}$$

- Možemo li α prikazati kao linearu kombinaciju težina, $\alpha = \mathbf{w}^T \mathbf{x}$?
- Da, ako prepostavimo **dijeljenu kovarijacijsku matricu**:

$$\begin{aligned} \alpha &= h_1(\mathbf{x}) - h_2(\mathbf{x}) \\ &= \mathbf{x}^T \underbrace{\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}_{\mathbf{w}} - \underbrace{\frac{1}{2}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 + \ln \frac{P(y = 1)}{P(y = 2)}}_{w_0} = \mathbf{w}^T \mathbf{x} + w_0 \end{aligned}$$

\Rightarrow logistička regresija istovjetna je Bayesovom klasifikatoru s dijeljenom Σ



- Broj parametara: $\frac{n}{2}(n+1) + 2n + 1$ (Bayes) vs. $n+1$ (logistička regresija)
 \Rightarrow diskriminativan model daje istu predikciju, ali s manje parametara

2 Naivan Bayesov klasifikator

- $\mathbf{x} = (x_1, \dots, x_n)$ ima $\prod_{k=1}^n K_k$ mogućih vrijednosti
- $p(\mathbf{x}|y)$ kao kategorička razdioba od \mathbf{x} \Rightarrow **previše parametara i nema generalizacije**
- Pojednostavljenje uvođenjem **induktivnih prepostavki** u obliku uvjetnih nezavisnosti
- Prepostavka: u svakoj klasi, svaka značajka uvjetno je nezavisna od svih drugih:

$$x_k \perp (x_1, \dots, x_{k-1}) | y \Leftrightarrow P(x_k | x_1, \dots, x_{k-1}, y) = P(x_k | y)$$

- Faktorizacija uz tu prepostavku:

$$P(x_1, \dots, x_n | y) = \prod_{k=1}^n P(x_k | x_1, \dots, x_{k-1}, y) = \prod_{k=1}^n P(x_k | y)$$

- **Naivan Bayesov klasifikator** (*Naïve Bayes classifier*):

$$h(x_1, \dots, x_n) = \operatorname{argmax}_j P(y=j) \prod_{k=1}^n P(x_k | y=j)$$

- Procjena parametara – MLE za $P(y)$ i MAP za $P(\mathbf{x}|y)$:

$$\begin{aligned} P(y=j) &= \hat{\mu}_j = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N} \\ P(x_k | y=j) &= \hat{\mu}_{k,j} = \frac{\sum_{i=1}^N \mathbf{1}\{x_k^{(i)} = x_k \wedge y^{(i)} = j\} + 1}{\sum_{i=1}^N \mathbf{1}\{y^{(i)} = j\} + K_k} = \frac{N_{kj} + 1}{N_j + K_k} \end{aligned}$$

gdje je N_j broj primjera u klasi j , a N_k broj vrijednosti značajke k

- Broj parametara: $\sum_{k=1}^n (K_k - 1) \cdot K + K - 1$
- Prepostavka uvjetne nezavisnosti uglavnom ne vrijedi, no model u praksi radi dobro

3 Uvjetna nezavisnost

- **Uvjetna nezavisnost** X i Y uz dani Z – notacija: $X \perp Y | Z$:

$$\begin{aligned} P(X, Y | Z) &= P(X|Z)P(Y|Z) \\ P(X|Y, Z) &= P(X|Z) \\ P(Y|X, Z) &= P(Y|Z) \end{aligned}$$

4 Polunaivan Bayesov klasifikator

- Ideja: **združiti** (ne faktorizirati) varijable koje nisu uvjetno nezavisne
- Npr., ako $x_2 \not\perp x_3 | y$:
$$P(x_1, x_2, x_3, y) = P(x_1|y)P(x_2, x_3|y)P(y)$$

\Rightarrow slabije pretpostavke \Leftrightarrow složeniji model \Leftrightarrow više parametara
- Broj mogućih združivanja \Leftrightarrow broj particija n -članog skupa \Leftrightarrow **Bellov broj** B_n
- Previše kombinacija \Rightarrow **heurističko pretraživanje** na temelju **kriterija združivanja**
- Dva pristupa:
 - **točnost modela** (unakrsna provjera) – algoritam FSSJ
 - **procjena zavisnosti varijabli** – algoritmi TAN i k -DB

Algoritam FSSJ

1. Inicijaliziraj $X = \emptyset$. Početna faktorizacija:

$$P(x_1, \dots, x_n, y) = P(x_1) \cdots P(x_n)P(y)$$

$$P(y|x_1, \dots, x_n) = P(y)$$

Klasificiraj primjere iz skupa za provjeru: $y^* = \operatorname{argmax}_j P(y = j)$

2. Za svaku varijablu $x_i \notin X$ koja još nije uključena u model, razmotri:

- (a) Uključi x_i kao uvjetno nezavisnu u odnosu na ostale varijable za danu klasu j
- (b) Uključi x_i tako da se ona doda u zajednički faktor s nekom već uključenom varijablom

3. Izaberi x_i i opciju koja minimizira pogrešku generalizacije

4. Ponavljam od koraka (2) do konvergencije pogreške

- **Uzajamna informacija** – zavisnost varijabli kao odstupanje $P(x, y)$ od $P(x)P(y)$:

$$I(x, y) = D_{\text{KL}}(P(x, y) || P(x)P(y)) = \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{P(x)P(y)}$$

$$\Rightarrow I(x, y) = 0 \Leftrightarrow x \perp\!\!\!\perp y, \quad I(x, y) > 0 \Leftrightarrow x \not\perp\!\!\!\perp y$$

- $D_{\text{KL}}(P || Q)$ – **Kullback-Leiblerova divergencija** (odstupanje) distribucije P od Q :

$$D_{\text{KL}}(P || Q) = \sum_x P(x) \ln \frac{P(x)}{Q(x)}$$

$$\Rightarrow \text{relativna entropija } P(x) \text{ u odnosu na } Q(x)$$

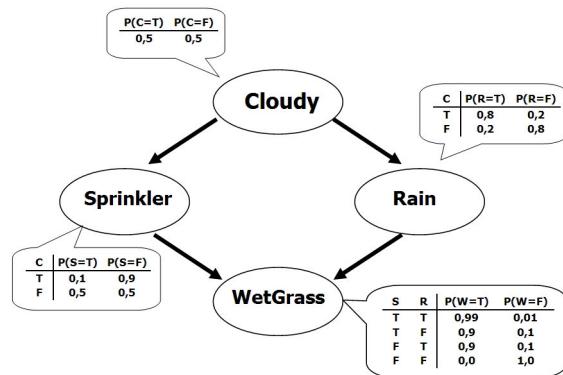
17. Probabilistički grafički modeli

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.4

1 Uvod

- **Probabilistički grafički model (PGM)** – sažet zapis zajedničke distrib. pomoću grafa
- Čvorovi grafa su varijable, bridovi su zavisnosti između varijabli

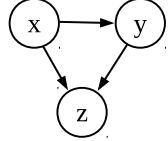


- Svrha: **probabilističko zaključivanje** (određivanje vrijednosti neopažanih varijabli)
- Tri aspekta PGM-a: (1) **reprezentacija**, (2) **zaključivanje** i (3) **učenje**
- Reprezentacija:
 - usmjereni aciklički graf \Rightarrow **Bayesove mreže**
 - neusmjereni graf \Rightarrow **Markovljeve mreže**
- Zaključivanje – određivanje vrijednosti nepažanih varijabli na temelju opažanih
- Učenje – procjena parametara ili učenje strukture mreže na temelju podatka
- Mi se fokusiramo na Bayesove mreže

2 Bayesove mreže: reprezentacija

- Usmjereni aciklički graf (*directed acyclic graph, DAG*)

- Bridovi povezuju varijablu koja uvjetuje s varijablom koja je uvjetovana
- Npr., $p(x, y, z) = p(x)p(y|x)p(z|x, y)$

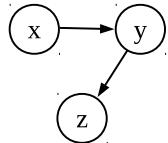


- Bez pretpostavki o uvjetnoj nezavisnosti:

$$\begin{aligned} p(x_1, \dots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \dots, x_{n-1}) \\ &= \prod_{k=1}^n p(x_k|x_1, \dots, x_{k-1}) \end{aligned}$$

\Rightarrow pravilo lanca \Leftrightarrow potpuno povezana Bayesova mreža

- Pretpostavke o uvjetnoj nezavisnosti uklanjaju bridove i pojednostavljaju mrežu
- Npr., ako $x \perp z | y$, onda $p(x, y, z) = p(x)p(y|x)p(z|y)$:



- Formalno, zajednička distribucija definirana Bayesovom mrežom je:

$$p(\mathbf{x}) = \prod_{k=1}^n p(x_k | \text{pa}(x_k))$$

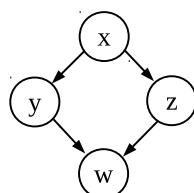
dje $\text{pa}(x_k)$ označava **čvorove roditelje** čvora x_k

- Čvorovi su poredani u **topološki uredaj** (roditelji dolaze prije djece)
- Svaki DAG ima barem jedan topološki uredaj
- **Uredajno Markovljevo svojstvo** (UMS): svaki čvor x_k ovisi samo o roditeljima:

$$x_k \perp \text{pred}(x_k) \setminus \text{pa}(x_k) \mid \text{pa}(x_k)$$

dje je $\text{pred}(x_k)$ skup prethodnika čvora x_k po topološkom uređaju

- Primjer: $p(x, y, z, w) = p(x)p(y|x)p(z|x)p(w|y, z)$



- Faktorizacija:

$$\begin{aligned}
 p(x)p(y|x)p(z|x)p(w|y,z) &= p(x,y)p(z|x)p(w|y,z) \\
 y \perp z | x &\Rightarrow p(x,y)p(z|x, \textcolor{red}{y})p(w|y,z) \\
 &= p(x,y,z)p(w|y,z) \\
 x \perp w | y, z &\Rightarrow p(x,y,z)p(w|\textcolor{red}{x},y,z) \\
 &= p(x,y,z,w)
 \end{aligned}$$

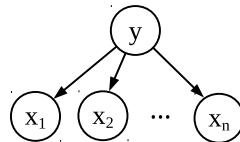
- Uvjetne nezavisnosti proizlaze iz UMS-a:

$$\begin{aligned}
 x_k \perp \text{pred}(x_k) \setminus \text{pa}(x_k) \mid \text{pa}(x_k) \\
 y \perp \{x\} \setminus \{x\} \mid \{x\} \\
 z \perp \{x, y\} \setminus \{x\} \mid \{x\} \Rightarrow y \perp z | x \\
 w \perp \{x, y, z\} \setminus \{y, z\} \mid \{x, y\} \Rightarrow x \perp w | y, z
 \end{aligned}$$

3 Primjeri Bayesovih mreža

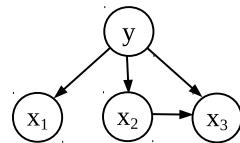
- **Naivan Bayesov klasifikator:**

$$P(\mathbf{x}, y) = P(y) \prod_{i=1}^n P(x_i|y)$$



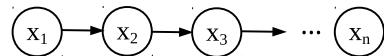
- Polunaivan Bayesov klasifikator. Npr.:

$$P(x_1, x_2, x_3, y) = P(x_1|y)P(x_2|x_3, y)P(y) = P(x_1|y)P(x_2|y)P(x_3|x_2, y)P(y)$$



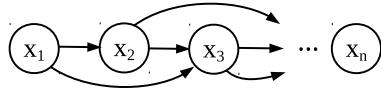
- Markovljev model prvog reda – za modeliranje slijednih podataka (npr., tekst):

$$p(\mathbf{x}) = p(x_1) \prod_{k=2}^n p(x_k|x_{k-1})$$



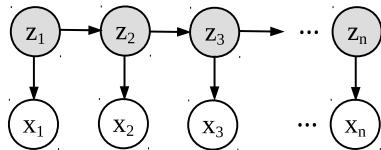
- Markovljev model drugog reda – modelira dulje zavisnosti:

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1) \prod_{k=3}^n p(x_k|x_{k-1}x_{k-2})$$



- Problem: eksplisitno modeliranje duljih zavisnosti povećava složenost modela
- **Skriveni Markovljev model** (*Hidden Markov Model, HMM*):

$$p(\mathbf{x}, \mathbf{z}) = p(z_1)p(x_1|z_1) \prod_{k=2}^n p(z_k|z_{k-1})p(x_k|z_k)$$



⇒ indirektno modelira dulje zavisnosti preko skrivenih varijabli \mathbf{z}

4 D-separacija

- Ispitivanje uvjetne nezavisnosti dviju varijabli uz zadane druge varijable
- **D-separacija**: analiziramo povezanost staze u grafu između dva čvora
- Tri pravila: račvanje, lanac, sraz
- (1) **Račvanje**: $x \leftarrow z \rightarrow y$
 - UMS: $y \perp\!\!\!\perp x | z \Leftrightarrow x \perp\!\!\!\perp y | z$
 - ⇒ ako je varijabla z opažena, onda su čvorovi odvojeni, inače su povezani
- (2) **Lanac**: $x \rightarrow z \rightarrow y$

$$p(x, y, z) = p(x)p(z|x)p(y|z)$$
 - UMS: $y \perp\!\!\!\perp x | z \Leftrightarrow x \perp\!\!\!\perp y | z$
 - ⇒ ako je varijabla z opažena, onda su čvorovi odvojeni, inače su povezani
- (3) **Sraz**: $x \rightarrow z \leftarrow y$

$$p(x, y, z) = p(x)p(y)p(z|x, y)$$
 - UMS: $y \perp\!\!\!\perp x | \emptyset$
 - ⇒ ako je varijabla z **neopažena**, onda su čvorovi odvojeni, inače su povezani
- Kod sraza varijable x i y se “natječu” za objašnjavanje (uzorkovanje) varijable z
- **Efekt objašnjavanja** (*explaining away*): opažanje x i z smanjuje vjerojatnost za y :

$$p(x|z) \neq p(x|y, z) \Leftrightarrow x \not\perp\!\!\!\perp y | z$$

- Primjer 1: Bacanje dva novčića ($x, y \in \{0, 1\}$) i opažanje njihove sume ($z = x + y$)
- Primjer 2: x – mononukleoza, y – upala grla, z – visoka temperatura

D-separacija čvorova

Raspolažemo skupom varijabli E koje su opažene.

Za **stazu** P od čvora x do čvora y kažemo da je **d-odvojena (d-separated)** akko vrijedi **barem jedno** od sljedećeg:

- P sadrži **lanac** $x \rightarrow z \rightarrow y$ ili $x \leftarrow z \leftarrow y$ i $z \in E$
- P sadrži **račvanje** $x \leftarrow z \rightarrow y$ i $z \in E$
- P sadrži **sraz** $x \rightarrow z \leftarrow y$ i varijabla z **nije** u E i nijedan sljedbenik od z nije u E

Za **par čvorova** x i y kažemo da su čvorovi x i y d-separirani za dani E ako su **sve staze** između ta dva čvora d-separirane za dani E .

Čvorovi x i y su d-separirani za dani E **akko** su uvjetno nezavisni za dani E .

- Implementacija: algoritam **Bayesove kuglice** (*Bayes-Ball*)

18. Probabilistički grafički modeli II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.4

1 Zaključivanje

- Primjer trave i prskalice (v. predavanje 17):

$$p(c, s, r, w) = p(c)p(s|c)p(r|c)p(w|s, r)$$

- Upit: Ako je trava mokra ($w = 1$), koja je vjerojatnost kiše (r) i prskalice (s)?

$$\begin{aligned} P(s = 1|w = 1) &= \frac{P(s = 1, w = 1)}{P(w = 1)} \\ &= \frac{\sum_{c,r} P(c, s = 1, r, w = 1)}{\sum_{c,r,s} P(c, s, r, w = 1)} = 0.2781/0.6471 = 0.43 \\ P(r = 1|w = 1) &= \frac{P(r = 1, w = 1)}{P(w = 1)} \\ &= \frac{\sum_{c,s} P(c, s, r = 1, w = 1)}{\sum_{c,r,s} P(c, s, r, w = 1)} = 0.4851/0.6471 = 0.708 \end{aligned}$$

gdje je $P(w = 1)$ **vjerojatnost dokaza**

- Dvije vrste upita: (1) posteriorni upiti i (2) MAP-upiti
- **Posteriorni upit** je izračun uvjetne vjerojatnosti:

$$p(\mathbf{x}_q|\mathbf{x}_o) = \frac{\sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)}{p(\mathbf{x}_o)} = \frac{\sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)}{\sum_{\mathbf{x}'_n, \mathbf{x}'_q} p(\mathbf{x}'_q, \mathbf{x}_o, \mathbf{x}'_n)}$$

gdje su \mathbf{x}_q varijable upita, \mathbf{x}_o su opažene varijable, a \mathbf{x}_n varijable smetnje (*nuisance*)

- **MAP-upiti** – najvjerojatnija vrijednost varijabli upita:

$$\mathbf{x}_q^* = \operatorname{argmax}_{\mathbf{x}_q} \sum_{\mathbf{x}_n} p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$$

2 Zaključivanje: eliminacija varijabli

- Odgovaranje upita iziskuje konstrukciju $p(\mathbf{x}_q, \mathbf{x}_o, \mathbf{x}_n)$ pa marginalizaciju/normalizaciju
- Velik $n \Rightarrow$ **kombinatorna eksplozija** \Rightarrow NP-složen problem
- Poništava prednost Bayesove mreže (sažet zapis zajedničke distribucije)
- Alternative: **egzaktno zaključivanje** i **približno zaključivanje**
- **Eliminacija varijabli** – egzaktno zaključivanje pomoću dinamičkog programiranja
- **Eliminacija varijabli zbroj-umnožak** – potiskivanje marginalizacije što dublje:

$$\begin{aligned}
p(w) &= \sum_c \sum_s \sum_r p(c, s, r, w) \\
&= \sum_c \sum_s \sum_r p(c)p(s|c)p(r|c)p(w|s, r) \\
&= \sum_s \sum_r p(w|s, r) \underbrace{\sum_c p(c)p(s|c)p(r|c)}_{t_1(s, r)} \\
&= \sum_s \underbrace{\sum_r p(w|s, r)t_1(s, r)}_{t_2(s, w)} \\
&= \sum_s t_2(s, w) \\
&= t_3(w)
\end{aligned}$$

- Varijante algoritma za skriveni Markovljev model (HMM):
 - eliminacija varijabli \Rightarrow **algoritam naprijed nazad** (*forward-backward algorithm*)
 - MAP-upiti \Rightarrow **Viterbijev algoritam**
- Za općenite Bayesove mreže eliminacija varijabli je presložena
- Alternativa: približno zaključivanje – **propagacijski algoritmi** i **metode uzorkovanja**

3 Zaključivanje: metode uzorkovanja

- Ideja: procjena distribucije na temelju uzorka
- Ako uzorkujemo uzorke $\mathbf{x} \sim P(\mathbf{x})$, očekivanje je:

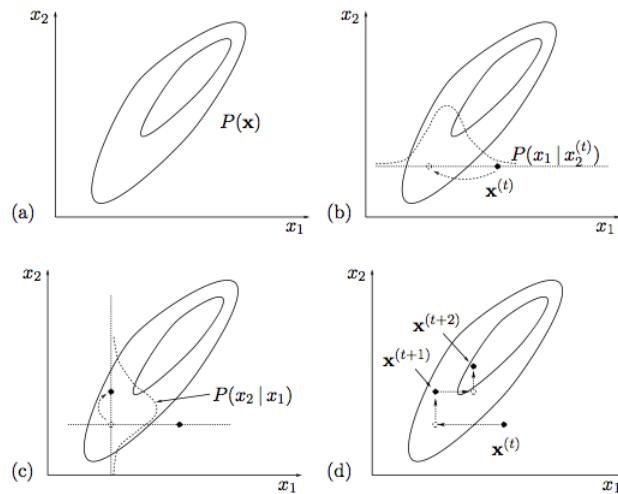
$$P(\mathbf{x} = x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{\mathbf{x} = x\}$$

- Najjjednostavnija metoda: **unaprijedno uzorkovanje** (*forward sampling*)

– Uzorkovanje varijable za varijablu, prema topološkom uređaju mreže

- Problem: želimo uzorkovati iz uvjetne vjerojatnosti $P(\mathbf{x}_q | \mathbf{x}_o)$
- **Uzorkovanje s odbijanjem** (*rejection sampling*)
 - Uzorkovanje unaprijed i odbijanje \mathbf{x} za koje \mathbf{x}_o nisu na željenim vrijednostima
 - Problem: neučinkovito, osobito ako je vjerojatnost dokaza $P(\mathbf{x}_o)$ malena
- **Uzorkovanje po važnosti** (*importance sampling*)
 - Postavljanje \mathbf{x}_o na željene vrijednosti, unaprijedno uzorkovanje i korekcija očekivanja
 - Problem: loša kvaliteta procjene, pogotovo ako su \mathbf{x}_o pri dnu Bayesove mreže
- **Gibbsovo uzokovanje** (*Gibbs sampling*)
 - Postupak iz porodice **Markov Chain Monte Carlo (MCMC)**
 - Krenuvši od slučajnog vektora \mathbf{x} , uzorkujemo ciklički varijablu po varijablu

$$\begin{aligned}
 \mathbf{x}^0 &\sim p(x_1^0, x_2^0, x_3^0) \quad \Rightarrow \text{početni vektor (npr., unaprijednim uzorkovanjem)} \\
 x_1^1 &\sim p(x_1 | x_2^0, x_3^0) \\
 x_2^1 &\sim p(x_2 | x_1^1, x_3^0) \\
 x_3^1 &\sim p(x_3 | x_1^1, x_2^1) \quad \Rightarrow \text{vektor } \mathbf{x}^1 = (x_1^1, x_2^1, x_3^1) \\
 x_1^2 &\sim p(x_1 | x_2^1, x_3^1) \\
 x_2^2 &\sim p(x_2 | x_1^2, x_3^1) \\
 x_3^2 &\sim p(x_3 | x_1^2, x_2^2) \quad \Rightarrow \text{vektor } \mathbf{x}^2 = (x_1^2, x_2^2, x_3^2) \\
 &\vdots
 \end{aligned}$$



4 Učenje

- PGM-ovi su probabilistički modeli \Rightarrow učenje se svodi na **procjenu parametara θ**
- MLE, MAP ili bayesovska procjena

- Log-izglednost za općenitu Bayesovu mrežu:

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\theta} | \mathcal{D}) &= \ln p(\mathcal{D} | \boldsymbol{\theta}) \\
&= \ln p(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)} | \boldsymbol{\theta}) \\
&= \ln \prod_{i=1}^N p(\mathbf{x}^{(i)} | \boldsymbol{\theta}) \\
&= \ln \prod_{i=1}^N \prod_{k=1}^n p(x_k^{(i)} | \text{pa}(x_k^{(i)}), \boldsymbol{\theta}_k) \\
&= \ln \prod_{k=1}^n \prod_{i=1}^N p(x_k^{(i)} | \text{pa}(x_k^{(i)}), \boldsymbol{\theta}_k) \\
&= \sum_{k=1}^n \sum_{i=1}^N \ln p(x_k^{(i)} | \text{pa}(x_k^{(i)}), \boldsymbol{\theta}_k)
\end{aligned}$$

- MLE procjena za k -ti čvor:

$$\boldsymbol{\theta}_k^* = \underset{\boldsymbol{\theta}_k}{\text{argmax}} \sum_{i=1}^N \ln p(x_k^{(i)} | \text{pa}(x_k^{(i)}), \boldsymbol{\theta}_k)$$

- MAP procjena za k -ti čvor:

$$\boldsymbol{\theta}_k^* = \underset{\boldsymbol{\theta}_k}{\text{argmax}} \left(\sum_{i=1}^N \ln p(x_k^{(i)} | \text{pa}(x_k^{(i)}), \boldsymbol{\theta}_k) + \ln p(\boldsymbol{\theta}_k) \right)$$

- MAP-procjena za kategorijsku razdiobu (Dirichlet-kategorijski model uz $\alpha = 2$):

$$\begin{aligned}
\hat{\mu}_{k,j,l} &= \frac{N_{kjl} + 1}{N_{kj} + K_k} \\
N_{kjl} &= \sum_{i=1}^N \mathbf{1}\{\mathbf{x}_{\text{pa}(x_k)}^{(i)} = j \wedge x_k^{(i)} = l\} \\
N_{kj} &= \sum_l N_{kjl}
\end{aligned}$$

gdje je K_k broj mogućih vrijednosti varijable x_k

- Primjer: MAP procjena za čvor w u mreži s travom i prskalicom (v. predavanje 17):

$$P(w|s, r) = \frac{\sum_{i=1}^N \mathbf{1}\{x_s^{(i)} = s \wedge x_r^{(i)} = r \wedge x_w^{(i)} = w\} + 1}{\sum_{i=1}^N \mathbf{1}\{x_s^{(i)} = s \wedge x_r^{(i)} = r\} + 2}$$

- Modeli sa skrivenim varijablama (npr., HMM, GMM) \Rightarrow tzv. **nepotpuni podatci**
 - Log-izglednost se ne dekomponira po strukturi grafa
 \Rightarrow MLE nema rješenje u zatvorenoj formi
 - Učenje pomoću **algoritma maksimizacije očekivanja** ili **gradijentnim usponom**
- **Učenje strukture mreže:**
 - Polunaivan Bayesov klasifikator: v. 4.3.2–4.3.4 u skripti
 - Učenje općenite strukture Bayesove mreže: **algoritam K2**

19. Grupiranje

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.1

1 Nenadzirano učenje

- Raspolažemo skupom **neoznačenih primjera** (*unlabeled data*): $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$
- Primjerni su neoznačeni jer ih (1) ne znamo označiti ili (2) označavanje je preskupo
- Osnovni zadatci nenadziranog učenja:
 - grupiranje (*clustering*)
 - procjena gustoće (*density estimation*)
 - otkrivanje novih/stršećih vrijednosti (*novelty/outlier detection*)
 - smanjenje dimenzionalnosti (*dimensionality reduction*)
- **Polunadzirano učenje:** većina primjera je neoznačena

2 Grupiranje

- Razdjeljivanje primjera u grupe (*clusters*), tako da su **slični** primjeri u istoj grupi
- Nalaženje “prirodnih” (intrinzičnih) grupa u skupu neoznačenih podataka
- Vrste grupiranja: **čvrsto/meko, partijsko/hijerarhijsko**
- Primjene: (1) istraživanje podataka, (2) kompresija, (3) polunadzirano učenje
- Grupiranje primjera / grupiranje značajki / bi-clustering

3 Algoritam K-sredina

- Particijsko grupiranje u K čvrstih grupa (K je unaprijed određen)
- **Funkcija pogreške** (kriterijska funkcija):

$$J = \sum_{k=1}^K \sum_{i=1}^N b_k^{(i)} \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\|^2$$

gdje je $\boldsymbol{\mu}_k$ centroid k -te grupe, a $b_k^{(i)}$ indikatorska varijabla pripadnosti $\mathbf{x}^{(i)}$ grupi k

- Svaki primjer $\mathbf{x}^{(i)}$ svrstavamo u grupu s njemu najbližim centroidom $\boldsymbol{\mu}_k$:

$$b_k^{(i)} = \begin{cases} 1 & \text{ako } k = \operatorname{argmin}_j \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\| \\ 0 & \text{inače} \end{cases}$$

- Tražimo grupiranje $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K$ koje minimizira pogrešku: $\operatorname{argmin}_{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K} J$
- Analitička minimizacija nije moguća jer su $b_k^{(i)}$ i $\boldsymbol{\mu}_k$ međuvisni
- Alternativa: **iterativna optimizacija**

- Fiksiramo $\boldsymbol{\mu}_k$ na neke inicijalne vrijednosti
- Pridružimo primjere grupama (izračunamo $b_k^{(i)}$ za $i = 1, \dots, N$)
- Uz fiksne $b_k^{(i)}$, minimizacija J daje formulu za ažuriranje centroida:

$$\nabla_{\boldsymbol{\mu}_k} J = \mathbf{0} \quad \Rightarrow \quad 2 \sum_{i=1}^N b_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k) = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{\mu}_k = \frac{\sum_i b_k^{(i)} \mathbf{x}^{(i)}}{\sum_i b_k^{(i)}}$$

- Ponavljamo do konvergencije $\boldsymbol{\mu}_k$ odnosno $b_k^{(i)}$

Algoritam K-sredina (k-means algorithm)

```

1:  inicijaliziraj centroide  $\boldsymbol{\mu}_k$ ,  $k = 1, \dots, K$ 
2:  ponavljam
3:    za svaki  $\mathbf{x}^{(i)} \in \mathcal{D}$ 
4:       $b_k^{(i)} \leftarrow \begin{cases} 1 & \text{ako } k = \operatorname{argmin}_j \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_j\| \\ 0 & \text{inače} \end{cases}$ 
5:    za svaki  $\boldsymbol{\mu}_k$ ,  $k = 1, \dots, K$ 
6:       $\boldsymbol{\mu}_k \leftarrow \sum_{i=1}^N b_k^{(i)} \mathbf{x}^{(i)} / \sum_{i=1}^N b_k^{(i)}$ 
7:  dok  $\boldsymbol{\mu}_k$  ne konvergiraju

```

- Vremenska složenost za T iteracija: $T(\mathcal{O}(nNK) + \mathcal{O}(nN)) = \mathcal{O}(TnNK)$

- **Konvergencija algoritma:**

- Broj konfiguracija (particija) je konačan i iznosi K^N
- J monotono pada kroz iteracije

\Rightarrow algoritam svaku konfiguraciju posjećuje najviše jednom \Rightarrow **algoritam konvergira**

- **Optimalnost algoritma:**

- Algoritam **pohlepno pretražuje** konfiguracije te nalazi **lokalni optimum** od J
- Optimalnost rješenja ovisi o odabiru početnih središta

- Pristupi za odabir početnih središta:

- Nasumičan odabir primjera kao centroida
- K slučajnih vektora prirodnih centroida cijelog skupa podataka
- K centroida iz K segmenata primjera projiciranih na prvu PCA komponentu
- **k-means++**: vjerojatnost odabira primjera $\mathbf{x}^{(i)}$ kao novog središta $\boldsymbol{\mu}_{k+1}$:

$$P(\boldsymbol{\mu}_{k+1} = \mathbf{x}^{(i)} | \mathcal{D}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k) = \frac{\min_k \|\boldsymbol{\mu}_k - \mathbf{x}^{(i)}\|^2}{\sum_j \min_k \|\boldsymbol{\mu}_k - \mathbf{x}^{(j)}\|^2}$$

\Rightarrow vjerojatnost je proporcionalna kvadratu udaljenosti od već odabranih središta

- Grupiranje treba pokrenuti više puta i uzeti rezultat s najmanjim J

4 Algoritam K-medoida

- Algoritma K-sredina: (1) primjeri moraju biti vektori, (2) udaljenost je euklidska
- **Algoritam K-medoida**: poopćenje K -sredina za općenitu mjeru sličnosti/različitosti
- Prototipi grupe nisu centroidi nego **medoidi** (odabrani primjeri u svakoj grupi)
- Funkcija pogreške:

$$J = \sum_{i=1}^N \sum_{k=1}^K b_k^{(i)} \nu(\mathbf{x}^{(i)}, \boldsymbol{\mu}_k)$$

gdje je $\nu : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ općenita **mjera različitosti** dvaju primjera

- Tipična izvedba je **algoritam PAM** (*partitioning around medoids*)

Algoritam PAM

- ```

1: inicijaliziraj medoide $\mathcal{M} = \{\boldsymbol{\mu}_k\}_{k=1}^K$ na odabrane $\mathbf{x}^{(i)}$
2: ponavljaj
3: za svaki $\mathbf{x}^{(i)} \in \mathcal{D} \setminus \mathcal{M}$
4: $b_k^{(i)} \leftarrow \begin{cases} 1 & \text{ako } k = \operatorname{argmin}_j \nu(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j) \\ 0 & \text{inače} \end{cases}$
5: za svaki $\boldsymbol{\mu}_k \in \mathcal{M}$
6: $\boldsymbol{\mu}_k \leftarrow \operatorname{argmin}_{\boldsymbol{\mu}_j \in \mathcal{D} \setminus \mathcal{M} \cup \{\boldsymbol{\mu}_k\}} \sum_i b_k^{(i)} \nu(\mathbf{x}^{(i)}, \boldsymbol{\mu}_j)$
7: dok $\boldsymbol{\mu}_k$ ne konvergiraju

```

- Složenost za  $T$  iteracija:  $T(\mathcal{O}(K(N-K)) + \mathcal{O}(K(N-K)^2)) = \mathcal{O}(TK(N-K)^2)$
- Nedostatak algoritma PAM: visoka vremenska složenost

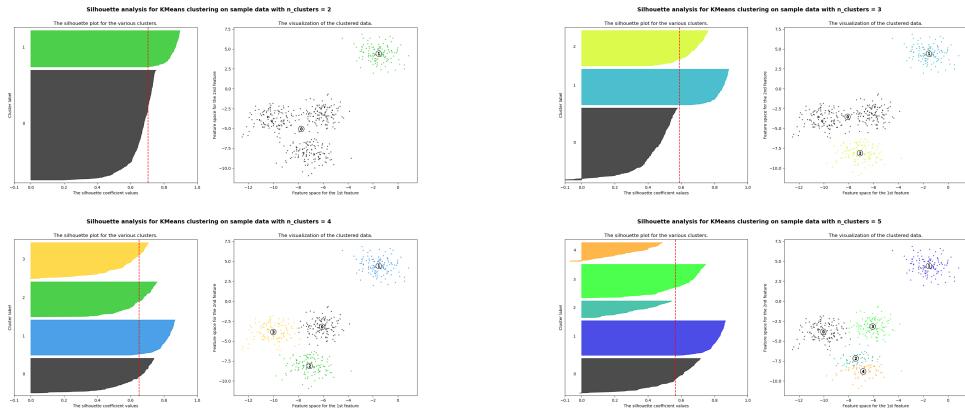
## 5 Provjera grupa

- **Broj grupa**  $K$  koji mnogih je algoritama grupiranja potrebno odrediti unaprijed
- Odabir optimalnog broja grupa dio je **provjere grupiranja** (*cluster validation*)
- $J$  ostvaruje minimum za  $K = N \Rightarrow$  nije indikativno za optimalan broj grupa
- Jednostavnije metode za odabir broja grupa:
  - Ručna provjera kvalitete grupa
  - Redukcija dimenzija u 2D-prostor (PCA, MDS, CA, t-SNE) i vizualna provjera
  - Metoda “koljena” (*elbow method*) – nalaženje platoa funkcije  $J(K)$
- **Analiza siluete** (*silhouette analysis*):
  - Silueta primjera  $\mathbf{x}^{(i)}$ :

- Silueta primjera  $\mathbf{x}^{(i)}$ :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \in [-1, +1]$$

- $a(i)$  i  $b(i)$  su prosjek udaljenost od  $\mathbf{x}^{(i)}$  do primjera iste odnosno najbliže grupe
- Računamo i grafički prikazujemo  $s(i)$  za sve primjere svake grupe
- Loše grupiranje: ispodprosječne siluete nekih grupa i/ili visoka varijanca silueta
- Primjer (scikit-learn):



- **Minimizacija regularizirane funkcije pogreške:**

- Kažnjavanje modela s velikim brojem grupa:

$$K^* = \operatorname{argmin}_K (J(K) + \lambda K)$$

- **Akaikeov kriterij (AIC)** za algoritam  $K$ -sredina:  $\lambda = 2n$

- **Točnost na podskupu primjera:**

- Raspolažemo označenim podskupom primjera ili parova primjera

- **Randov indeks** – točnost grupiranja na razini parova primjera:

$$R = \frac{a + b}{\binom{N}{2}} \in [0, 1]$$

- $a$  – broj jednakoznačenih parova u istim grupama
- $b$  – broj različito označenih parova u različitim grupama
- Optimalan  $K$  je onaj koji maksimizira  $R(K)$

## 20. Grupiranje II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.1

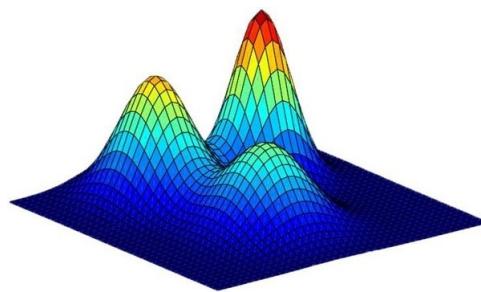
### 1 Model Gaussove mješavine

- **Model Gaussove mješavine (GMM)**  $\Rightarrow$  probabilističko meko partijsko grupiranje
- Poopćenje algoritma K-sredina: umjesto  $b_k^{(i)} \in \{0, 1\}$  imamo  $h_k^{(i)} \in [0, 1]$
- $h_k^{(i)}$  je **odgovornost** – vjerojatnost da je primjer  $\mathbf{x}^{(i)}$  generirala grupa  $k$
- GMM je poseban slučaj **modela miješane gustoće** (*mixture model*)
- Model miješane gustoće je linearna kombinacija  $K$  **komponenti**:

$$p(\mathbf{x}) = \sum_{k=1}^K p(\mathbf{x}, y=k) = \sum_{k=1}^K P(y=k)p(\mathbf{x}|y=k) = \sum_{k=1}^K \pi_k p(\mathbf{x}|\boldsymbol{\theta}_k)$$

gdje su  $\pi_k$  **koeficijenti mješavine**, a  $p(\mathbf{x}|\boldsymbol{\theta}_k)$  **gustoće komponenti**

- Primjer: model bivarijatne Gaussove mješavine s  $K = 3$  grupe:



- Odgovornost možemo izračunati Bayesovim pravilom:

$$h_k^{(i)} = P(y=k|\mathbf{x}^{(i)}) = \frac{P(y=k)p(\mathbf{x}^{(i)}|y=k)}{\sum_j P(y=j)p(\mathbf{x}^{(i)}|y=j)} = \frac{\pi_k p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)}{\sum_j \pi_j p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_j)}$$

- Parametri modela su  $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1}^K$ ,  $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- Parametre možemo (pokušati) procijeniti metodom MLE

- Log-izglednost parametara modela (tzv. **nepotpuna izglednost**):

$$\ln \mathcal{L}(\boldsymbol{\theta} | \mathcal{D}) = \ln \prod_{i=1}^N p(\mathbf{x}^{(i)}) = \ln \prod_{i=1}^N \sum_{k=1}^K \pi_k p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k) = \sum_{i=1}^N \ln \sum_{k=1}^K \pi_k p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k)$$

$\Rightarrow$  ne faktorizira se po komponentama  $\Rightarrow$  maksimizacija nema analitičko rješenje

## 2 Algoritam maksimizacije očekivanja

- Proširenje modela miješane gustoće **latentnim varijablama** (varijable koje ne opažamo)
- Latentna kategorička varijabla  $\mathbf{z}^{(i)}$  definira koja je grupa generirala primjer  $\mathbf{x}^{(i)}$ :

$$\mathbf{z}^{(i)} = (z_1^{(i)}, \dots, z_k^{(i)}, \dots, z_K^{(i)})$$

- Distribucija kategoričke varijable  $\mathbf{z}^{(i)}$ :

$$P(\mathbf{z}^{(i)} = k) = \prod_{k=1}^K \pi_k^{z_k^{(i)}}$$

- Zajednička gustoća varijabli  $\mathbf{x}$  i  $\mathbf{z}$ :

$$p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}) = P(\mathbf{z})p(\mathbf{x} | \mathbf{z}, \boldsymbol{\theta}) = \prod_{k=1}^K \pi_k^{z_k} \prod_{k=1}^K p(\mathbf{x} | \boldsymbol{\theta}_k)^{z_k} = \prod_{k=1}^K \pi_k^{z_k} p(\mathbf{x} | \boldsymbol{\theta}_k)^{z_k}$$

$\Rightarrow$  model s **latentnim varijablama**  $\mathbf{z}$

- Log-izglednost parametara modela (tzv. **potpuna izglednost**):

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta} | \mathcal{D}, \mathbf{Z}) &= \ln \prod_{i=1}^N p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)} | \boldsymbol{\theta}) = \ln \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_k^{(i)}} p(\mathbf{x} | \boldsymbol{\theta}_k)^{z_k^{(i)}} \\ &= \sum_{i=1}^N \sum_{k=1}^K z_k^{(i)} (\ln \pi_k + \ln p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k)) \end{aligned}$$

$\Rightarrow$  ako su  $\mathbf{z}^{(i)}$  poznate, maksimizacija ove log-izglednosti ima analitičko rješenje

- $\mathbf{z}^{(i)}$  su nepoznate, no možemo izračunati **očekivanje** izglednosti uz fiksirane  $\pi_k$  i  $\boldsymbol{\theta}_k$
- Može se pokazati: povećanje očekivanja od  $\mathcal{L}(\boldsymbol{\theta} | \mathcal{D}, \mathbf{Z}) \Rightarrow$  povećanje  $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$
- **Algoritam maksimizacije očekivanja (EM-algoritam)**: iterativna optimizacija  $\mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$
- Dva koraka algoritma: E-korak (*expectation*) i M-korak (*maximization*)
- **E-korak**: Izračun očekivanja potpune izglednosti uz fiksirane parametre u iteraciji  $t$ :

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\mathbf{Z} | \mathcal{D}, \boldsymbol{\theta}^{(t)}} \left[ \sum_{i=1}^N \sum_{k=1}^K z_k^{(i)} (\ln \pi_k + \ln p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k)) \right] \\ &= \sum_{i=1}^N \sum_{k=1}^K \underbrace{\mathbb{E}[z_k^{(i)} | \mathcal{D}, \boldsymbol{\theta}^{(t)}]}_{= h_k^{(i)}} (\ln \pi_k + \ln p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k)) \end{aligned}$$

- **M-korak:** Izračun parametara za iteraciju  $(t + 1)$  koji maksimiziraju očekivanje:

$$\nabla_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = 0$$

$$\nabla_{\pi_k} \left( \sum_{i=1}^N \sum_{k=1}^K h_k^{(i)} \ln \pi_k + \lambda \left( \sum_k \pi_k - 1 \right) \right) = 0 \quad \Rightarrow \quad \pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^N h_k^{(i)}$$

$$\nabla_{\boldsymbol{\theta}_k} \sum_{i=1}^N h_k^{(i)} \ln p(\mathbf{x}^{(i)} | \boldsymbol{\theta}_k) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_i h_k^{(i)} \mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}$$

$$\Rightarrow \quad \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_i h_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)}) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)})^\top}{\sum_i h_k^{(i)}}$$

### Algoritam GMM (model GMM + EM-algoritam)

inicijaliziraj parametre  $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$   
ponavlja do konvergencije log-izglednosti ili parametara

**E-korak:**

Za svaki primjer  $\mathbf{x}^{(i)} \in \mathcal{D}$  i svaku komponentu  $k = 1, \dots, K$ :

$$h_k^{(i)} \leftarrow \frac{p(\mathbf{x}^{(i)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \pi_k}{\sum_{j=1}^K p(\mathbf{x}^{(i)} | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \pi_j}$$

**M-korak:**

Za svaku komponentu  $k = 1, \dots, K$ :

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_i h_k^{(i)} \mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}, \quad \boldsymbol{\Sigma}_k \leftarrow \frac{\sum_i h_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k) (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k)^\top}{\sum_i h_k^{(i)}}, \quad \pi_k \leftarrow \frac{1}{N} \sum_{i=1}^N h_k^{(i)}$$

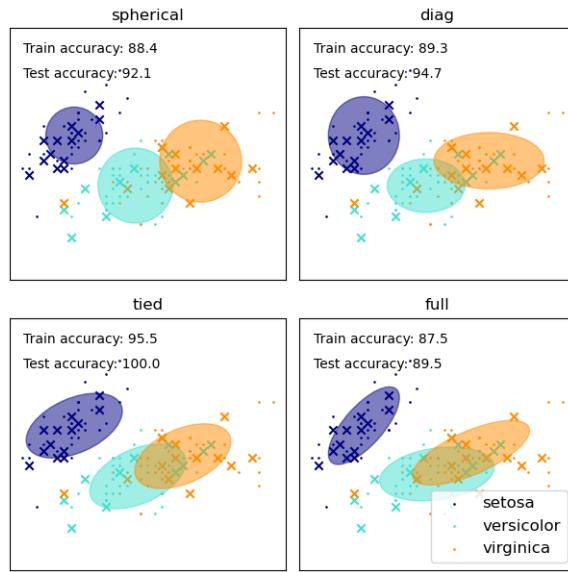
Izračunaj trenutnu vrijednost log-izglednosti

$$\ln \mathcal{L}(\boldsymbol{\theta} | \mathcal{D}) = \sum_{i=1}^N \ln \sum_{k=1}^K \pi_k p(\mathbf{x}^{(i)} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- EM-algoritam konvergira, ali ne nužno u globalni optimum log-izglednosti
- Akaikeov informacijski kriterij (AIC) za odabir optimalnog broja grupa:

$$K^* = \operatorname{argmin}_K (-2 \ln \mathcal{L}(K) + 2q(K))$$

- Moguća pojednostavljenja: dijeljena matrica, dijagonalna ili izotropna matrica  $\Sigma$



### 3 Hijerarhijsko grupiranje

- Hijerarhijsko grupiranje producira **dendrogram** – stablasti prikaz hijerarhije grupa
- Provodi se na temelju mjere udaljenosti ili mjere sličnosti/različitosti
- Može biti **aglomerativno** (bottom-up) ili **divizivno** (top-down)
- **Hijerarhijsko aglomerativno grupiranje (HAC)**: iterativno stapa najbliže parove grupa
- **Povezivanje** (*linkage*) – način izračuna udaljenosti između dvije grupe:

- **Jednostruko povezivanje** (*single linkage*)

$$d_{min}(\mathcal{G}_i, \mathcal{G}_j) = \min_{\mathbf{x} \in \mathcal{G}_i, \mathbf{x}' \in \mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

- **Potpuno povezivanje** (*complete linkage*)

$$d_{max}(\mathcal{G}_i, \mathcal{G}_j) = \max_{\mathbf{x} \in \mathcal{G}_i, \mathbf{x}' \in \mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

- **Prosječno povezivanje** (*average linkage*)

$$d_{avg}(\mathcal{G}_i, \mathcal{G}_j) = \frac{1}{N_i N_j} \sum_{\mathbf{x} \in \mathcal{G}_i} \sum_{\mathbf{x}' \in \mathcal{G}_j} d(\mathbf{x}, \mathbf{x}')$$

- **Povezivanje centroida** (*centroid linkage*)

$$d_{cent}(\mathcal{G}_i, \mathcal{G}_j) = \left\| \frac{1}{N_i} \sum_{\mathbf{x} \in \mathcal{G}_i} \mathbf{x} - \frac{1}{N_j} \sum_{\mathbf{x} \in \mathcal{G}_j} \mathbf{x} \right\|$$

#### Algoritam hijerarhijskog aglomerativnog grupiranja (HAC)

```

1: inicijaliziraj K , $k \leftarrow N$, $\mathcal{G}_i \leftarrow \{\mathbf{x}^{(i)}\}$ za $i = 1, \dots, N$
2: ponavljaj
3: $k \leftarrow k - 1$
4: $(\mathcal{G}_i, \mathcal{G}_j) \leftarrow \underset{\mathcal{G}_a, \mathcal{G}_b}{\operatorname{argmin}} d(\mathcal{G}_a, \mathcal{G}_b)$
5: $\mathcal{G}_i \leftarrow \mathcal{G}_i \cup \mathcal{G}_j$
6: dok je $k > K$

```

- Prostorna složenost: matrica udaljenosti za  $\binom{N}{2}$  parova primjera  $\Rightarrow \mathcal{O}(N^2)$
- Vremenska složenost: općenito  $\mathcal{O}(N^3)$ ,  $\mathcal{O}(N^2 \log N)$  s prioritetsnom listom

# 21. Vrednovanje modela

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.3

## 1 Osnovne mjere vrednovanja

- **Matrica zabune** (*confusion matrix*) – usporeba stvarnih oznaka i predikcija modela

|       |   | Stvarno |    |
|-------|---|---------|----|
|       |   | 1       | 0  |
| Model | 1 | TP      | FP |
|       | 0 | FN      | TN |

TP – true positives, FP – false positives, FN – false negatives, TN – true negatives

- **Točnost** (*accuracy*) je udio točno klasificiranih primjera u skupu svih primjera:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = 1 - E(h|\mathcal{D})$$

- Ako je udio klase izrazito neuravnotežen, točnost nije indikativna mjeru

- **Preciznost** (*precision*):

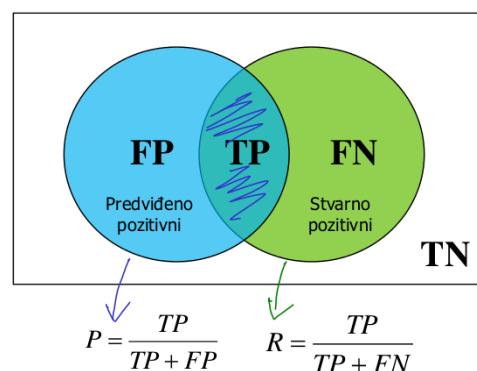
$$P = \frac{TP}{TP + FP}$$

⇒ udio pozitivno klasificiranih primjera u skupu pozitivno klasificiranih primjera

- **Odziv** (*recall, true positive rate, sensitivity*):

$$R = TPR = \frac{TP}{TP + FN}$$

⇒ udio pozitivno klasificiranih primjera u skupu svih pozitivnih primjera



- **Fall-out** (false positive rate)

$$FPR = \frac{FP}{FP + TN}$$

$\Rightarrow$  udio primjera pogrešno proglašenih pozitivnima

- **Specifičnost** (*specificity*):

$$S = \frac{TN}{TN + FP}$$

$\Rightarrow$  udio negativno klasificiranih primjera u skupu svih negativnih primjera

- **Mjera F1** – harmonijska sredina preciznosti i odziva:

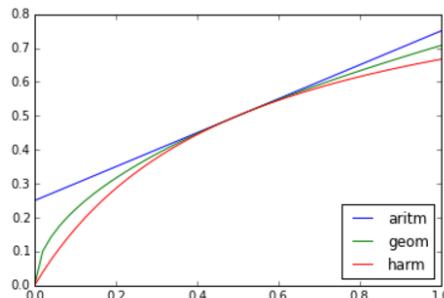
$$F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

- **Mjera F-beta** – poopćenje mjere F1 koje različito naglašava  $P$  i  $R$ :

$$F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R}$$

$\Rightarrow F_{0.5}$  dvostruko naglašava preciznost,  $F_2$  dvostruko naglašava odziv

- Harmonijska sredina je “najstroža” od triju sredina; npr. za  $R = 0.5$  i  $P \in [0, 1]$ :



- Primjer:  $N = 1000$ , od čega 100 poz. Ispravno klasificiranih 90 poz. i 650 neg.

|       |   | Stvarno |     |
|-------|---|---------|-----|
|       |   | 1       | 0   |
| Model | 1 | 90      | 250 |
|       | 0 | 10      | 650 |

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = 0.74$$

$$P = \frac{TP}{TP + FP} = \frac{90}{90 + 250} = 0.265$$

$$R = \frac{TP}{FP + FN} = \frac{90}{90 + 10} = 0.9$$

$$F_1 = \frac{2PR}{P + R} = \frac{2 \cdot 0.265 \cdot 0.9}{0.265 + 0.9} = 0.409$$

## 2 Višeklasna klasifikacija

- Iz matrice  $K \times K$  ( $K > 2$ ) izvodimo matricu  $2 \times 2$  za svaku klasu  $j$ , s elementima:
  - $\text{TP}_j$  –  $j$ -ti element dijagonale
  - $\text{FP}_j$  – zbroj nedijagonalnih elemenata  $j$ -tog retka
  - $\text{FN}_j$  – zbroj nedijagonalnih elemenata  $j$ -tog stupca
  - $\text{TN}_j = N - \text{TP}_j - \text{FP}_j - \text{FN}_j$  – zbroj po elementima izvan retka  $j$  i stupca  $j$
- **Makro-prosjek** (M): izračun mjere za svaku klasu pa uprosječivanje kroz klase

$$\text{Acc}^M = \frac{1}{K} \sum_{j=1}^K \text{Acc}_j, \quad P^M = \frac{1}{K} \sum_{j=1}^K P_j, \quad R^M = \frac{1}{K} \sum_{j=1}^K R_j, \quad F_1^M = \frac{1}{K} \sum_{j=1}^K F_{1,j}$$

$\Rightarrow$  jednak utjecaj svih klasa  $\Rightarrow$  loš rezultat na manjim klasama narušava mjeru

- **Mikro-prosjek** ( $\mu$ ): zbrajanje matrica pojedinačnih klasa pa izračun mjere

$$\text{TP} = \sum_{j=1}^K \text{TP}_j, \quad \text{FP} = \sum_{j=1}^K \text{FP}_j, \quad \text{FN} = \sum_{j=1}^K \text{FN}_j, \quad \text{TN} = \sum_{j=1}^K \text{TN}_j$$

$\Rightarrow$  vrijedi  $\text{FP} = \text{FN} \Rightarrow$  vrijedi  $P^\mu = R^\mu = F_1^\mu$

- Vrijedi  $\text{Acc}^M = \text{Acc}^\mu$
- Alternativa: neuprosječena točnost –  $\text{Acc} = \frac{1}{N} \sum_{j=1}^K \text{TP}_j = P^\mu = R^\mu = F_1^\mu$
- Primjer ( $N = 13$ ,  $K = 3$ ):

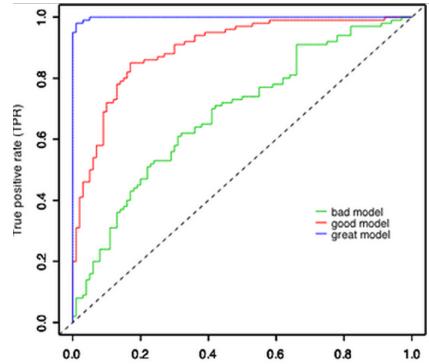
$$\begin{array}{c} \begin{array}{ccc} y = 1 & y = 2 & y = 3 \end{array} \\ \begin{array}{c} y = 1 \\ y = 2 \\ y = 3 \end{array} \left( \begin{array}{ccc} 1 & 1 & 0 \\ 2 & 2 & 3 \\ 0 & 0 & 4 \end{array} \right) \Rightarrow \begin{array}{c} \overbrace{\begin{array}{ccc} y = 1 & y = 2 & y = 3 \end{array}}^{\text{Makro}} \\ \begin{array}{ccc} (1 & 1) & (2 & 5) & (4 & 0) \\ (2 & 9) & (1 & 5) & (3 & 6) \end{array} \end{array} \Rightarrow \begin{array}{c} \overbrace{\begin{array}{cc} 7 & 6 \end{array}}^{\text{zbroj}} \\ \begin{array}{cc} 6 & 20 \end{array} \end{array} \end{array}$$

$$\begin{aligned} \text{Acc}^M &= \frac{1}{3} \left( \frac{10}{13} + \frac{7}{13} + \frac{10}{13} \right) = 0.69 & \text{Acc}^\mu &= \frac{27}{39} = 0.69 \\ P^M &= \frac{1}{3} \left( \frac{1}{2} + \frac{2}{7} + \frac{4}{4} \right) = 0.60 & P^\mu &= \frac{7}{13} = 0.54 \\ R^M &= \frac{1}{3} \left( \frac{1}{3} + \frac{2}{3} + \frac{4}{7} \right) = 0.52 & R^\mu &= \frac{7}{13} = 0.54 \\ F_1^M &= \frac{1}{3} (0.40 + 0.40 + 0.73) = 0.51 & F_1^\mu &= \frac{2P^M R^M}{P^M + R^M} = 0.54 \end{aligned}$$

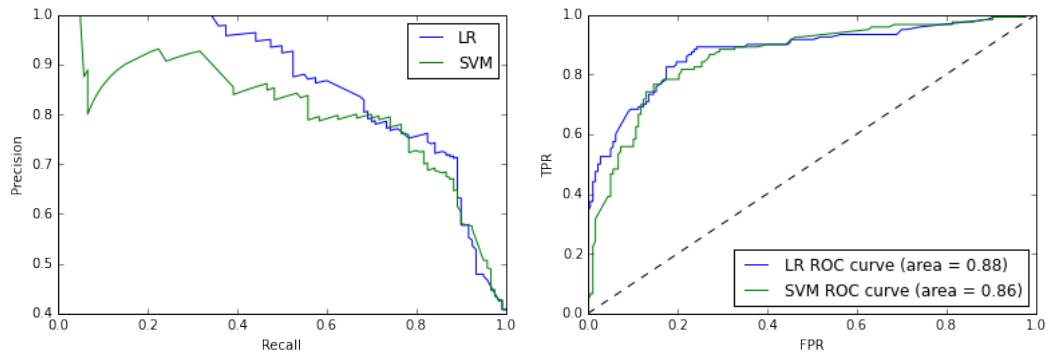
- Tipično (ali ne nužno)  $M < \mu$  jer klasifikatori rade lošije na manjim klasama

### 3 Vrednovanje klasifikatora s pragom

- Ugađanjem klasifikacijskoj praga može se ugađati  $P$  i  $R$  modela
- **Krivulja preciznost-odziv (P-R)** – preciznost kao funkcija odziva (monotonu opada)
- Agregatna mjera: **prosječna preciznost (AP)** (*average precision*)
- **Krivulja ROC** – odziv kao funkcija od FPR (fall-out)

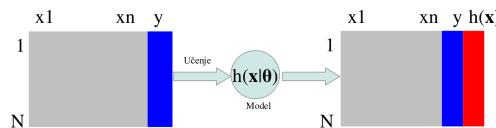


- Nasumična predikcija  $\Rightarrow TPR = FPR$ , neovisno o udjelu pozitivnih primjera
- Agregatna mjera: **površina ispod ROC krivulje (AUC)** (*area under curve*)
- Najbolji model: (1, 1) za krivulju P-R, (0, 1) za krivulju ROC



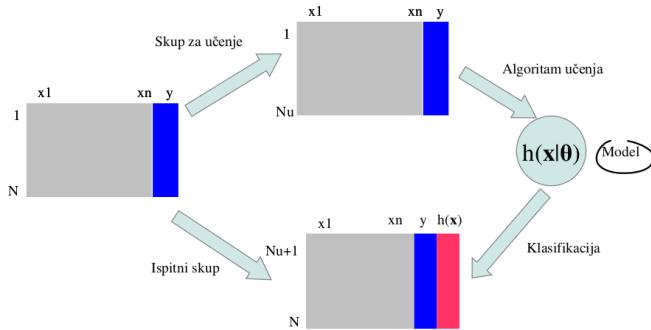
### 4 Procjena pogreške modela

- Ispitni skup je **slučajan uzorak**  $\Rightarrow$  svaka mjeri točnosti je funkcija **slučajne varijable**
- Procjena pogreške (točnosti) treba biti **dobra** (nepristrana) i **poštena** (realistična)
- Procjena na skupu za učenje  $\Rightarrow$  ne mjerimo pogrešku generalizacije  $\Rightarrow$  nepoštano



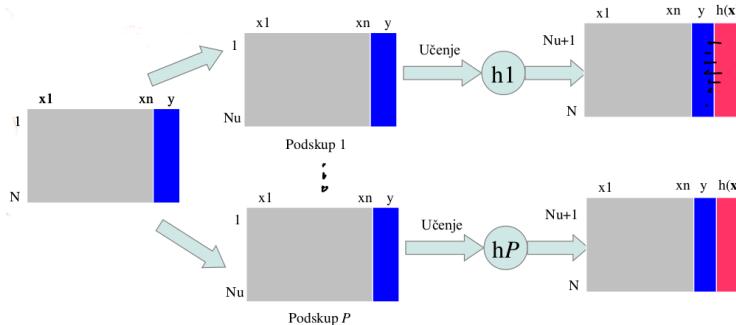
- **Metoda izdvajanja (holdout method)**

- Podjela na skup za učenje i skup za ispitivanje (npr., 70%–30%)
- Prednost: mjerimo pogrešku generalizacije
- Nedostatci: gubitak primjera za učenje, procjena na samo jednom uzorku



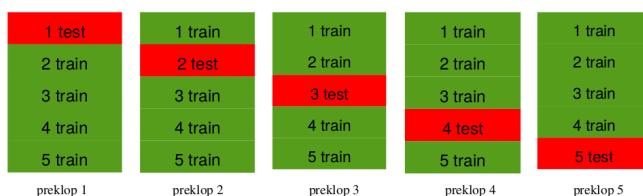
- **Ponovljeno izdvajanje (repeated holdout)**

- Višestruko uzorkovanje skupova za učenje/ispitivanje pa izračun prosjeka mjere
- Prednost: procjena pogreške generalizacije na više uzorka
- Nedostatak: ne kontroliramo koji su primjeri i koliko puta upotrijebljeni



- **$k$ -struka unakrsna provjera (CV) ( $k$ -folded cross-validation)**

- Podjela na  $k$  **preklopa (folds)** (tipično  $k = 5$  ili  $k = 10$ )
- Učenje na  $(k - 1)$  preklopa, ispitivanje na jednom preklopu, ponovljeno  $k$  puta
- Prednost: svaki je primjer iskorišten i za učenje i za ispitivanje
- Nedostatak: modeli nisu međusobno nezavisni  $\Rightarrow$  visoka varijanca procjene



- **Metoda izdvoji jednoga (LOOCV)** (*leave-one-out cross-validation*)
  - $k$ -struka unakrsna provjera uz  $k = N$
  - Prednost: gotovo svi primjeri se koriste za učenje u svakoj iteraciji
  - Nedostatci: računalno skupo, visoka varijanca procjene pogreške
- Procjena pogreške uz **odabir modela**:
  - Podjela na skup za **učenje** ( $\mathcal{D}_{\text{train}}$ ), **provjeru** ( $\mathcal{D}_{\text{validate}}$ ) i **ispitivanje** ( $\mathcal{D}_{\text{test}}$ )
  - Odabir modela: učenje na  $\mathcal{D}_{\text{train}}$  i ispitivanje na  $\mathcal{D}_{\text{validate}}$
  - Ispitivanje odabranog modela: učenje na  $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{validate}}$  i ispitivanje na  $\mathcal{D}_{\text{test}}$
- $k$ -struka CV uz odabir modela  $\Rightarrow$  **ugniježđena unakrsna provjera** (*nested CV*)

### Ugniježđena unakrsna provjera $k \times l$

```

1: podijeli \mathcal{D} na vanjske preklope \mathcal{D}_i , $i = 1, \dots, k$
2: za $i = 1, \dots, k$ radi: vanjska petlja
3: $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \setminus \mathcal{D}_i$, $\mathcal{D}_{\text{test}} \leftarrow \mathcal{D}_i$
4: za svaku odabranu vrijednost hiperparametra α radi:
5: podijeli $\mathcal{D}_{\text{train}}$ na unutarnje preklope \mathcal{D}_j , $j = 1, \dots, l$ unutarnja petlja
6: za $j = 1, \dots, l$ radi:
7: $\mathcal{D}_{\text{train}'} \leftarrow \mathcal{D}_{\text{train}} \setminus \mathcal{D}_j$, $\mathcal{D}_{\text{validate}} \leftarrow \mathcal{D}_j$
8: nauči model na $\mathcal{D}_{\text{train}'}$ i ispitaj na $\mathcal{D}_{\text{validate}}$
9: izračunaj prosjek mjere na l unutarnjih preklopa
10: odaberi hiperparametar α koji maksimizira prosjek mjere
11: nauči odabrani model na $\mathcal{D}_{\text{train}}$ i ispitaj na $\mathcal{D}_{\text{test}}$
12: izračunaj prosjek mjere na k vanjskih preklopa

```

- Odabir hiperparametara (redak 4) može biti vođen heurističkim pretraživanjem
- Kao optimalan model odabradi onaj koji je najčešće odabran u  $k$  vanjskih preklopa
- Paziti da se pri učenju modela koristi isključivo informacija iz skupa za učenje

## 22. Vrednovanje modela II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.3

### 1 Statističko zaključivanje – ideja

- Točnost modela mjerimo na slučajnome uzorku  $\Rightarrow$  točnost je **slučajna varijabla** (s.v.)
- **Statističko zaključivanje** omogućava zaključivanje na temelju slučajnog uzorka
- Osnovni pristupi: (1) **interval pouzdanosti** i (2) **statističko testiranje hipoteze**
- Osnovni pojmovi:
  - **populacija** – konačan ili beskonačan skup svih objekata od interesa
  - **uzorak** – podskup populacije veličine  $N$  dobiven (slučajnim) uzorkovanjem
  - **statistika** – procjenitelj (funkcija uzorka) koji odgovara parametru populacije
- Parametarsko statističko zaključivanje  $\Rightarrow$  temeljeno na distr. uzorkovanja statistike
- **Distribucija uzorkovanja** (*sampling distribution*) – distr. statistike na temelju sl. uzorka
- **Standardna pogreška (SE)** (*standard error*) – stand. devijacija distr. uzorkovanja
- Za neke je statistike distr. uzorkovanja u zatvorenoj formi, npr., srednja vrijednost
- **Distribucija uzorkovanja srednje vrijednosti**:

$$\text{populacija: } x \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \text{statistika: } \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$$

$$\Rightarrow \text{SE} = \sqrt{\sigma^2/N} = \sigma/\sqrt{N}$$

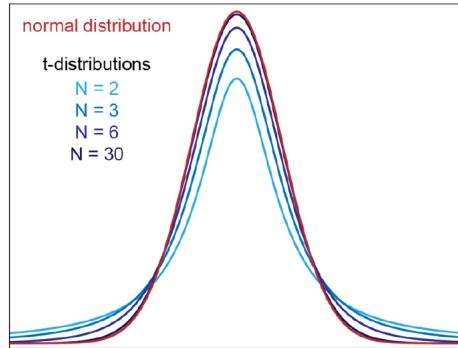
- **Središnji granični teorem:** za  $N \rightarrow \infty$  vrijedi  $\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$ , neovisno o distr. od  $x$
- U praksi, već za  $N \geq 30$  možemo pretpostaviti  $\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$
- **Standardizacijom** od  $\bar{x}$  dobivamo **z-vrijednost** s distribucijom uzorkovanja:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

- Ako je  $\sigma^2$  populacije **nepoznata**, procjenjujemo  $\hat{\sigma}^2$  iz uzorka i koristimo **t-statistiku**:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t(N - 1)$$

gdje je  $t(N - 1)$  **Studentova t-distribucija** sa  $N - 1$  stupnjeva slobode



- Za  $N \geq 30$ , t-distribucija praktički je identična normalnoj
- Sažetak pravila (za statistiku  $\bar{x}$  i varijancu  $\hat{\sigma}^2$  procijenjenu iz uzorka):
  - $N \geq 30$  ("velik uzorak")  $\Rightarrow$  koristimo z-statistiku ili t-statistiku (svejedno)
  - $N < 30$  i populacija je normalna  $\Rightarrow$  koristimo t-statistiku
  - $N < 30$  i populacija nije normalna  $\Rightarrow$  ne radimo param. stat. zaključivanje!
- Provjera normalnosti: Shapiro-Wilkov test ili **Q-Q plot za normalnu distribuciju**

## 2 Statističko zaključivanje za vrednovanje modela

- Distribucija uzorkovanja nije poznata za sve mjere vrednovanja (npr., za F1-mjeru)
- Ideja: koristiti srednju vrijednost odabrane mјere izračunatu na  $K$  preklopa:
  - **populacija** – svi mogući primjeri (moguće beskonačno)
  - **uzorak** – vrijednosti mјere na  $K$  preklopa višestruke unakrsne provjere
  - **statistika** – srednja vrijednost mјere kroz  $K$  preklopa
- NB: Veličina statističkog uzorka je  $K$  (broj preklopa), a ne  $N$  (broj primjera)!
- Tipično  $K < 30$ , pa treba provjeriti normalnost

## 3 Interval pouzdanosti

- **Interval pouzdanosti** srednje vrijednosti populacije  $\mu$  na temelju sredine uzorka  $\bar{x}$ :

$$\mu = \bar{x} \pm SE \quad (\text{sa X\% pouzdanosti})$$

- Budući da

$$z = \frac{\bar{x} - \mu}{\text{SE}} \sim \mathcal{N}(0, 1)$$

to (za  $X = 95\%$ ) vrijedi

$$P(-1.96 \leq (\bar{x} - \mu)/\text{SE} \leq 1.96) = 0.95$$

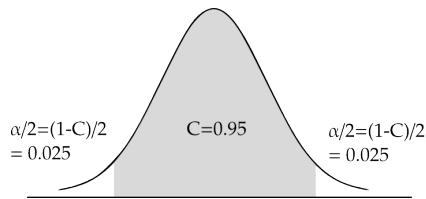
što (uz  $\text{SE} = \sigma/\sqrt{N}$ ) daje

$$P(\bar{x} - 1.96\sigma/\sqrt{N} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{N}) = 0.95$$

odnosno

$$\mu = \bar{x} \pm 1.96\sigma/\sqrt{N} \quad (\text{sa } 95\% \text{ pouzdanosti})$$

- Veza između **razine pouzdanosti**  $C \in [0, 1]$  i **razine značajnosti**  $\alpha \in [0, 1]$ :  $C = 1 - \alpha$

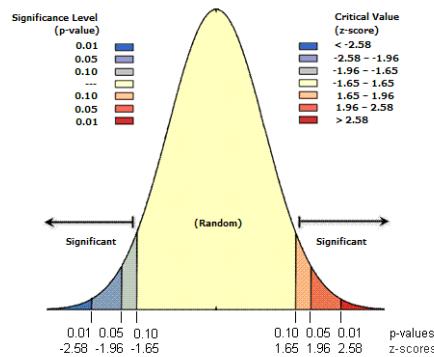


- Općenito, normalan interval pouzdanosti razine  $C = 1 - \alpha$ :

$$P(\bar{x} - z_{\alpha/2}\sigma/\sqrt{N} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{N}) = 1 - \alpha$$

gdje je  $z_{\alpha/2}$  **kritična vrijednost** distribucije  $\mathcal{N}(0, 1)$  za razinu značajnosti  $\alpha$ , tj.:

$$P(|z| \geq z_{\alpha/2}) = \alpha$$



- Ako je  $\hat{\sigma}^2$  procijenjen iz uzorka, umjesto z-statistike treba upotrijebiti t-statistiku:

$$P(\bar{x} - t_{\alpha/2}\hat{\sigma}/\sqrt{N} \leq \mu \leq \bar{x} + t_{\alpha/2}\hat{\sigma}/\sqrt{N}) = 1 - \alpha$$

gdje je  $t_{\alpha/2}$  **kritična vrijednost** distribucije  $t(N - 1)$  za razinu značajnosti  $\alpha$

- Kritične vrijednosti z-distribucije i t-distribucije očitavaju se iz [tablica](#)

## 4 Statističko testiranje hipoteze

- Pretpostavljamo da parametar populacije  $\mu$  ima neku vrijednost (**hipoteza**)
- Možemo li **odbaciti** tu hipotezu na temelju opažanja  $\bar{x}$ , koje donekle odstupa od  $\mu$ ?
- **p-vrijednost:** vjerojatnost da smo opazili  $\bar{x}$  ili ekstremnije, ako je hipoteza istinita
- Hipotezu odbacujemo ako je p-vrijednost manja od odabrane **razine značajnosti**  $\alpha$
- **t-test** za srednju vrijednost: koristi t-statistiku kao testnu statistiku

### t-test za srednju vrijednost

- Korak 1: Iskazati hipotezu  $H_0$  i njoj alternativnu hipotezu  $H_1$ :

$$H_0 : \mu = \dots$$
$$H_1 : \mu \neq \dots$$

- Korak 2: Odabrati nivo značajnosti  $\alpha$ , npr.  $\alpha = 1\%$  (ili  $5\%$ )
- Korak 3: Izračunati t-statistiku pod hipotezom  $H_0$ :  $t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}}$
- Korak 4: Na distribuciji  $t(N - 1)$  provjeriti:
  - Varijanta A (provjera kritičnog područja): je li  $|t| \geq t_{\alpha/2}$ ?
  - Varijanta B (provjera p-vrijednosti): je li  $P(|X| > t) \leq \alpha$ ?
- Korak 5:
  - Ako da: odbaciti hipotezu  $H_0$  i prihvati hipotezu  $H_1$
  - Ako ne: ne možemo odbaciti (ali niti prihvati) hipotezu  $H_0$
- Korak 6: Formulirati zaključak

- **Jednostrani test** (*one-tailed test*):

- Testiramo je li  $\bar{x}$  veće/manje od  $\mu$
- Hipoteza  $H_0$  je ista, alternativna hipoteza je  $H_1 : \mu > \dots$  ili  $H_1 : \mu < \dots$
- p-vrijednost je polovica p-vrijednosti za dvostrani test  $\Rightarrow$  lakše je odbaciti  $H_0$
- Kod vrednovanja modela u načelu treba izbjegavati jednostrani test

## 5 Usporedba modela

- Je li točnost modela A **statistički značajno** različita/bolja od točnosti modela B?

- **Upareni t-test** (*matched-pair t-test*): testiranje razlika u točnosti kroz  $K$  preklopa
- Uzorak je  $\{d_k\}_{i=1}^K$ , gdje je  $d_i = m_i^A - m_i^B$  razlika u mjeri  $m$  na preklopu  $i$
- Izračunavamo srednju vrijednost razlika,  $\bar{d} = \bar{m}^A - \bar{m}^B = \frac{1}{K} \sum_{i=1}^K d_i$
- Hipoteze:

$$H_0 : \bar{m}^A - \bar{m}^B = \bar{d} = 0 \quad \text{točnosti su iste}$$

$$H_1 : \bar{m}^A - \bar{m}^B \neq 0 \quad \text{točnosti su različite (dvostrani test)}$$

ili  $H_1 : \bar{m}^A - \bar{m}^B \leq 0 \quad \text{točnost od A je manja/veća od B (jednostrani test)}$

- t-statistika:

$$t = \frac{\bar{d} - 0}{\hat{\sigma}/\sqrt{K}}, \quad \text{gdje } \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^K (d_i - \bar{d})^2}{K - 1}}$$

- Ako je  $K < 30$ , treba provjeriti vrijedi li normalnost razlika  $d_i$

# 23. Odabir značajki

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.1

## 1 Motivacija i pristupi

- Metode za **smanjenje dimenzionalnosti** ulaznog prostora:
  - **Odabir značajki** (*feature selection*) – odabir podskupa izvornih značajki
  - **Transformacija značajki** – izvođenje novih značajki iz izvornih značajki
- Svrha:
  - Uklanjanje **irelevantnih** i **redundatnih** značajki povećava točnost modela
  - Lakše razumijevanje i objašnjavanje modela
  - Pomoć u vizualizaciji podataka
- Odabir značajki čuva izvornu semantiku značajki  $\Rightarrow$  bolja tumačivost modela

## 2 Univarijatni filter

- Procjena intrinsične vrijednosti (*merit*) svake značajke pa odabir po pragu ili rangu
- Prednosti: dobro skalira, računalno jednostavno, nezavisno od modela
- Nedostatci: nezavisno od modela, ne uzima u obzir interakciju između značajki
- Ideja: značajka  $x_k$  je **relevantna**  $\Leftrightarrow$  postoji **zavisnost** između varijabli  $x_k$  i  $y$
- **Uzajamna informacija** – zavisnost varijabli  $x$  i  $y$  kao odstupanje  $P(x, y)$  od  $P(x)P(y)$ :

$$I(x, y) = D_{\text{KL}}(P(x, y) \parallel P(x)P(y)) = \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{P(x)P(y)}$$

$\Rightarrow$  relevantnost značajke  $x_k$  za klasu  $y$  proporcionalna je sa  $I(x_k, y)$

- **t-test** (primjenjivo za  $K = 2$ )
  - Test značajnosti razlike srednje vrijednosti od  $x_k$  za klase  $y = 0$  i  $y = 1$
  - Hipoteza  $H_0$ : srednje vrijednosti su jednake

- t-statistika (pod  $H_0$  distribuirana po t-distribuciji):

$$t = \frac{\bar{x}_k^0 - \bar{x}_k^1}{\hat{\sigma}_i \sqrt{\frac{1}{N_0} + \frac{1}{N_1}}} \sim t(N_0 + N_1 - 2)$$

gdje  $N_y = \sum_{i=1}^N \mathbf{1}\{y^{(i)} = y\}$  i  $\bar{x}_k^y = \frac{1}{N_y} \sum_{i=1}^N x_k^{(i)} \mathbf{1}\{y^{(i)} = y\}$

$\Rightarrow$  relevantnost značajke  $x_k$  obrnuto je proporcionalna **p-vrijednosti**

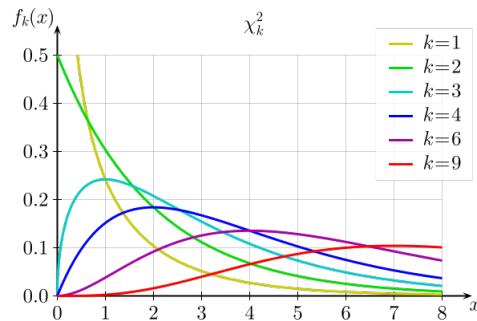
- **ANOVA** (za  $K > 2$ )

- Testiranje razlika srednjih vrijednosti značajke  $x_k$  kroz  $K$  klase
- $\chi^2$ -test (primjenjivo za kategoričke značajke)
  - Hipoteza  $H_0$ : varijable  $x_k$  i  $y$  su nezavisne ( $x_k \perp y$ )
  - $N$  – broj primjera,  $K$  – broj klase,  $K_k$  – broj vrijednosti varijable  $x_k$
  - **Tablica kontingencije** dimenzije  $K_k \times K$  sadrži opažene frekvencije  $O_{i,j}$
  - Izračun očekivanih frekvencija ( $E_{i,j}$ ) uz pretpostavku  $H_0$ :

$$\begin{aligned} P(x_k = i) &= \sum_j P(x_k = i, y = j) \\ P(y = j) &= \sum_i P(x_k = i, y = j) \\ E_{i,j} &= NP(x_k = i)P(y = j) \end{aligned}$$

- $\chi^2$ -statistika (pod  $H_0$  distribuirana po  $\chi^2$ -distribuciji):

$$\chi^2 = \sum_{i=1}^{K_k} \sum_{j=1}^K \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \sim \chi^2((K_k - 1)(K - 1))$$



$\Rightarrow$  relevantnost značajke  $x_k$  obrnuto je proporcionalna **p-vrijednosti**

- **p-mjera** – neparametarska usporedba srednjih vrijednosti  $x_k$  za klase  $y = 0$  i  $y = 1$ :

$$p(x_k, y) = \frac{\bar{x}_k^0 - \bar{x}_k^1}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

$\Rightarrow$  relevantnost značajke  $x_k$  proporcionalna je vrijednosti p-mjere

- **RELIEF** (Kira i Rendell, 1992) – neparametarska iterativna metoda (za  $K = 2$ )

- Iterativno ugadanje vektora relevantnosti svih  $n$  značajki (vektor  $\mathbf{w}$ )
- Slučajan odabir pivotnog primjera i primjera iste (*hit*) i različite klase (*miss*)
- Relevantnost  $x_k$  pada ako primjeri istih klasa imaju različite vrijednosti
- Relevantnost  $x_k$  raste ako primjeri različitih klasa imaju različite vrijednosti

### Algoritam RELIEF

```

1: postavi $w_k \leftarrow 0$ za svaku značajku $k = 1, \dots, n$
2: za $i = 1, \dots, m$ radi: -- m je broj iteracija
3: nasumično odabereti primjer $\mathbf{x} \in \mathcal{D}$
4: pronađi najbliži pogodak $\mathbf{x}^h \in \mathcal{D}$ i promašaj $\mathbf{x}^m \in \mathcal{D}$ (po L2-normi)
5: za $k = 1, \dots, n$ radi:
6: $w_k = w_k - \frac{1}{N}(x_k - x_k^h)^2 + \frac{1}{N}(x_k - x_k^m)^2$

```

## 3 Multivarijatni filter

- Univarijatne metode ocjenjuju relevantnost, neovisno o redundanciji značajki
- Multivarijatne metode ocjenjuju relevantnost i redundantnost skupa značajki
- **Uklanjanje značajki faktorom inflacije varijance (VIF)** (*variance inflation factor*)
  - Ideja:  $x_k$  je redundantna  $\Leftrightarrow$  može ju se dobro predvidjeti iz drugih varijabli
  - Model linearne regresije sa  $x_k$  kao zavisnom varijablom:

$$h_k(x_k; \mathbf{w}) = w_1 x_1 + \cdots + w_{k-1} x_{k-1} + w_{k+1} x_{k+1} + \cdots + w_n x_n$$

- VIF varijable  $x_k$ :

$$\text{VIF}_k = \frac{1}{1 - R_k^2} \in [1, \infty)$$

gdje je  $R_k^2$  **koeficijent determinacije** za  $h_k$  (v. [odjeljak 5.1.3](#) dodatka skripti)

- U praksi, značajke za koje  $\text{VIF} \geq 10$  smatraju se redundantnim
- Iterativno uklanjanje redundantnih značajki i ažuriranja VIF vrijednosti
- VIF uklanja isključivo redundante značajke (ne odabire relevantne značajke)

### Postepeno (stepwise) uklanjanje značajki VIF-om

```

1: $S \leftarrow \{1, \dots, n\}$
2: za $k \in S$ radi:
3: izračunaj VIF $_k$ sa $S \setminus \{k\}$ kao nezavisnim varijablama
4: $m \leftarrow \operatorname{argmax}_{k \in S} \text{VIF}_k$
5: dok $\text{VIF}_m \geq 10$ radi:
6: $S \leftarrow S \setminus \{m\}$
7: za $k \in S$ radi:
8: izračunaj VIF $_k$ sa $S \setminus \{k\}$ kao nezavisnim varijablama
9: $m \leftarrow \operatorname{argmax}_{k \in S} \text{VIF}_k$

```

- **Correlation feature selection (CFS)** – nalazi relevantne i neredundantne značajke

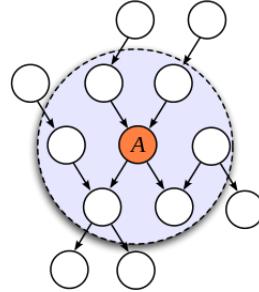
- Ocjena vrijednosti **podskupa značajki**  $S$  koji sadrži  $k$  značajki:

$$\text{Merit}_S = \frac{k \bar{r}_{x,y}}{\sqrt{k + k(k-1) \bar{r}_{x,x}}}$$

- $\bar{r}_{x,y}$  – prosječna korelacija (npr. Pearsonova) između varijabli iz  $S$  i varijable  $y$
- $\bar{r}_{x,x}$  – prosječna korelacija između svih  $k$  varijabli iz  $S$
- **Unaprijedno pretraživanje** prostora od  $2^n$  podskupova metodom **najbolji prvi**

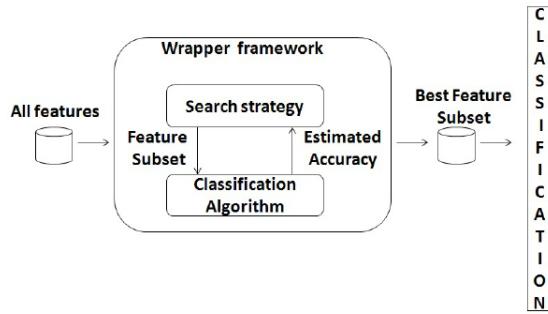
- **Markovljev pokrivač (Markov blanket)** – izravan odabir značajki za PGM-ove

- Markovljev pokrivač od  $x_k$ : roditelji od  $x_k$ , njegova djeca i roditelji djece
- Vrijednost varijable  $x_k$  u Bayesovoj mreži ovisi samo o Markovljevom pokrivaču

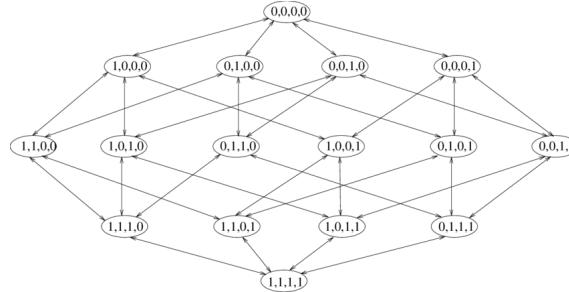


## 4 Metoda omotača

- Pretraživanje prostora od  $2^n$  podskupova značajki i provjera točnosti modela



- Kriterijska funkcija:
  - **Točnost modela** procijenjena unakrsnom provjerom
  - Mjera **prikladnosti modela** (*goodness of fit*) (npr. F-test)
- Pretraživanje:
  - **Unaprijedni odabir** – kreće od praznog skupa i dodaje značajke
  - **Unatražni odabir** – kreće od svih značajki i uklanja značajke
  - **Stepenast odabir (stepwise)** – unaprijedan odabir s unatražnim uklanjanjem



- Prednost: prilagođenost konkretnom modelu; nedostatak: računalna složenost

## 5 Ugrađene metode

- Neki algoritmi odabiru značajke pri postupku učenju: **stabla odluke, slučajne šume**
- Svaki algoritam s **L1-regularizacijom** implicitno provodi odabir značajki