15. Bayesov klasifikator

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

Jan Šnajder, natuknice s predavanja, v1.4

1 Pravila vjerojatnosti

- **Pravilo zbroja**:

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

$\Rightarrow$ **marginalna vjerojatnosti** 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*):

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

$\Rightarrow$ **faktorizacija** zajedničke vjerojatnosti na umnožak **faktora**

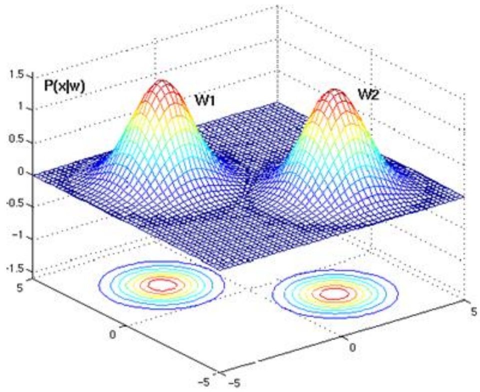
2 Bayesov klasifikator

- 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**):

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

- 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:

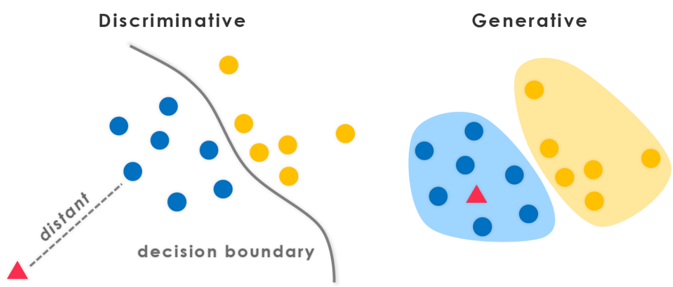
$$P(\mathbf{x}, y) = p(\mathbf{x}|y)P(y)$$

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

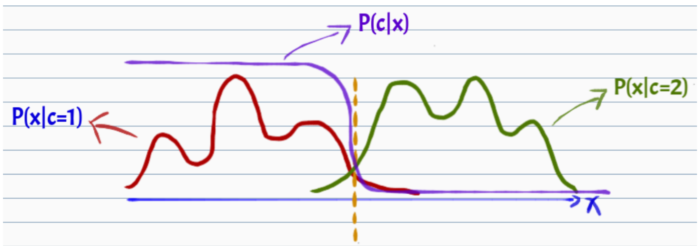
- Složeniji generativni modeli: **Bayesove mreže**, **HMM**, **GMM**, **LDA**
- Usp.: diskriminativni modeli izravno modeliraju $p(y|\mathbf{x})$; npr. logistička regresija:

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

- Diskriminativno vs. generativno:

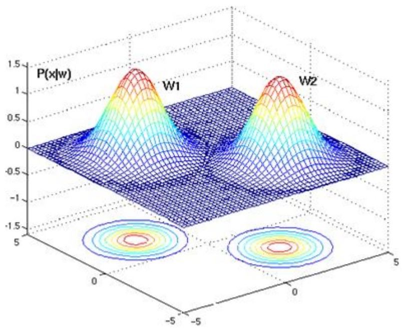


- Prednosti: laka ugradnja stručnog znanja, interpretabilnost/analiza rezultata
- Nedostatci: iziskuju mnogo primjera za učenje, nepotrebna složenost modeliranja
- Primjer: nepotrebna složenost modeliranja zajedničke vjerojatnosti:



4 Gaussov Bayesov klasifikator

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



- Jednodimenzijski 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**):

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

- Model za klasu j:

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

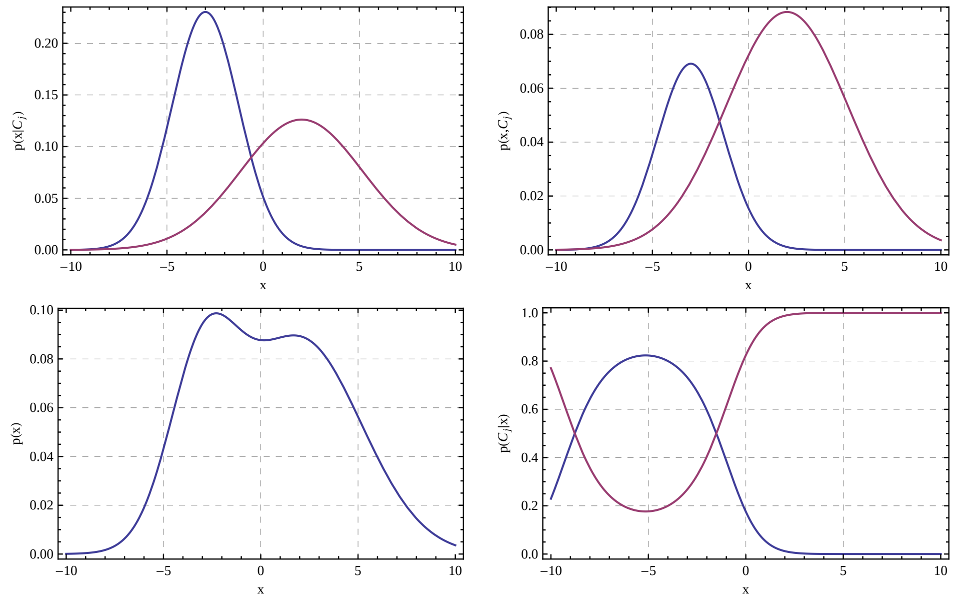
- Prelazak u logaritamsku domenu i uklanjanje konstanti:

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

- MLE procjene parametara:

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

- Primjer:
$p(x|y=1)\sim\mathcal{N}(-3,3)$, $P(y=1)=0.3$
$p(x|y=2)\sim\mathcal{N}(2,10)$, $P(y=2)=0.7$



- Više značajki ⇒ izglednosti modeliramo multivarijatnom normalnom razdiobom:

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

- Model za klasu j:

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

- MLE procjene parametara:

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{N_j}\sum_{i=1}^{N}\mathbf{1}\{y^{(i)} = j\}\mathbf{x}^{(i)}$$

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

$$\hat{\mu}_j = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N}$$

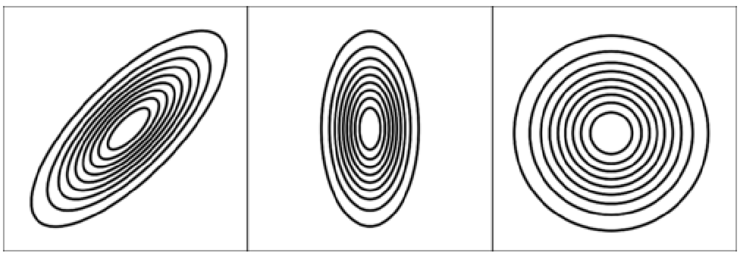
- Broj parametara: $\frac{n}{2}(n+1)K + K\cdot n + K - 1 \ \Rightarrow \mathcal{O}(n^2)$
- Granica između dviju klasa: $h_1(\mathbf{x}) - h_2(\mathbf{x}) = 0$:

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

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

5 Varijante Gaussovog Bayesovog klasifikatora

- Uvodimo pretpostavke na $\boldsymbol{\Sigma}$ koje pojednostavljuju model

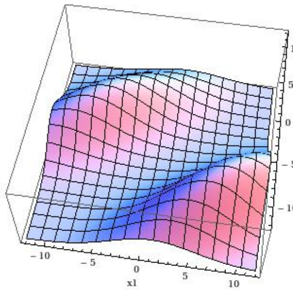


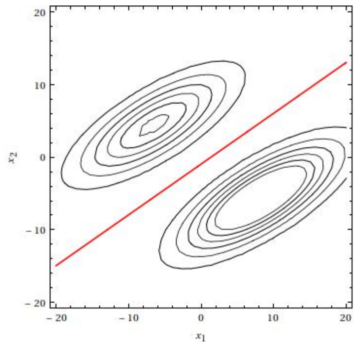
$\boldsymbol{\Sigma}$ $\quad$ $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$ $\quad$ $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

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

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

  - Model je linearan $\Rightarrow$ **linearna granica** između klasa
  - Broj parametara: $\frac{n}{2}(n+1) + nK + K - 1 \ \Rightarrow \mathcal{O}(n^2)$





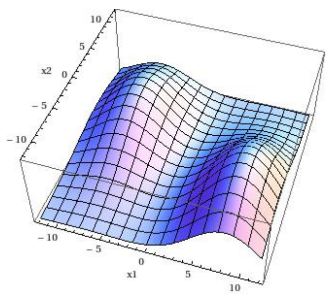
- **Dijeljena i dijagonalna kovarijacijska matrica**: $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i^2)$
  - Vrijedi $|\mathbf{\Sigma}| = \prod_i \sigma_i^2$ i $\mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_i^2)$
  - Izglednost klase:

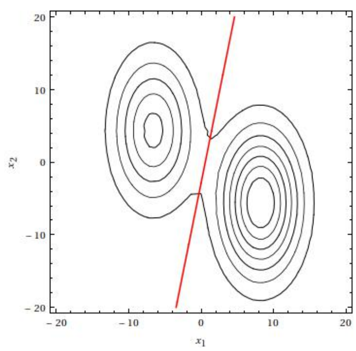
$$
\begin{aligned}
p(\mathbf{x}|y=j) &= \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\Big)\\
&= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^n \sigma_i} \exp\Big(-\frac{1}{2}\sum_{i=1}^n\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big)\\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

  $\Rightarrow$ **uvjetna nezavisnost** značajki $\Rightarrow$ **Gaussov naivan Bayesov klasifikator**
  - $x_k \perp x_j | y \;\Rightarrow\; \mathrm{Cov}(x_k|y, x_j|y) = 0 \;\Leftrightarrow\; p(\mathbf{x}|y) = \prod_k p(x_k|y)$
  - Model za klasu j:

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

  $\Rightarrow$ **normirana euklidska udaljenost** između primjera $\mathbf{x}$ i prototipa klase $\boldsymbol{\mu}_j$
  - Broj parametara: $n + n\cdot K + K - 1 \;\Rightarrow \mathcal{O}(n)$

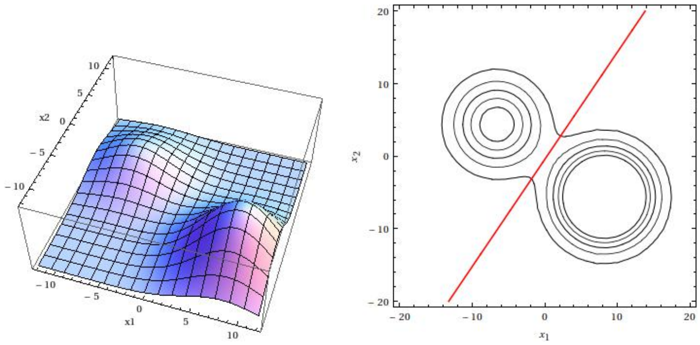




- **Izotropna kovarijacijska matrica**: $\mathbf{\Sigma} = \sigma^2\mathbf{I}$
  - Model za klasu j:

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

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



- Druge varijante:

Pretpostavka	Kov. matrica	Broj parametara
Različite, hiperelipsoidi	$\mathbf{\Sigma}_j$	$Kn(n+1)/2 + Kn$
Dijeljena, hiperelipsoidi	$\mathbf{\Sigma}$	$n(n+1)/2 + Kn$
Različite, poravnati hiperelipsoidi	$\mathbf{\Sigma}_j = \mathrm{diag}(\sigma^2_{i,j})$	$2Kn$
Dijeljena, poravnati hiperelipsoidi	$\mathbf{\Sigma} = \mathrm{diag}(\sigma^2_i)$	$n + Kn$
Različite, hipersfere	$\mathbf{\Sigma}_j = \sigma^2_j\mathbf{I}$	$K + Kn$
Dijeljena, hipersfere	$\mathbf{\Sigma} = \sigma^2\mathbf{I}$	$1 + Kn$

- Odabir modela: **unakrsnom provjerom**