

# Vytěžování dat

## Úloha 2: Bayesovské rozhodování

Michael Anděl



Evropský sociální fond  
Praha & EU: Investujeme do vaší budoucnosti

*Fakulta elektrotechnická, ČVUT*

## (Binary) Bayesian Classification – Revision

- **Given** the data  $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots, (\mathbf{x}_n, y_n)\}$  with  $y_i \in \{0, 1\}$ ,  $\mathbf{x}_i \in \mathbb{X}$ ; where  $\mathbb{X}$  is a feature space
- **Find** a rule  $f : \mathbb{X} \rightarrow \{0, 1\}$

# (Binary) Bayesian Classification – Revision

- **Given** the data  $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \dots, (\mathbf{x}_n, y_n)\}$  with  $y_i \in \{0, 1\}$ ,  $\mathbf{x}_i \in \mathbb{X}$  where  $\mathbb{X}$  is a feature space
- **Find** a rule  $f : \mathbb{X} \rightarrow \{0, 1\}$

## What about aposteriori probability?

- $f = \mathbf{x} \mapsto 1$  IF  $p(y = 1|\mathbf{x}) > p(y = 0|\mathbf{x})$  ELSE 0

## But what about $p(y|\mathbf{x})$ ?

- $p(y|\mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})}$
- $p(y = 1|\mathbf{x}) > p(y = 0|\mathbf{x}) \iff p(\mathbf{x}, y = 1) > p(\mathbf{x}, y = 0)$

## But what about $p(\mathbf{x}, y)$ ?

- $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$
- $p(\mathbf{x}|y = c) \approx \mathcal{N}(\mu_c, \sigma_c^2)$
- $\mathcal{N}(\mu, \sigma^2) \approx \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$
- $\mu_c \approx \frac{1}{|J_c|} \sum_{j \in J_c} x_j$
- $\sigma_c^2 \approx \sum_{j \in J_c} (x_j - \mu_c)^2$
- $J_c = \{i : 0 < i < n \wedge y_i = c\}$

# Assignment

1. Generate 2 populations  $\mathbf{X}_1, \mathbf{X}_2$  from  $\mathcal{N}(160, 10^2)$  and  $\mathcal{N}(185, 10^2)$  respectively, each of 100 examples. Use `randn`.
2. Plot the frequencies (histograms) of these two populations into a **one** figure. Use `hist`, `bar(..., '<color>')`,
3. Concatenate  $\mathbf{X}_1, \mathbf{X}_2$  into one data sample (vector)  $\mathbf{X}$ .
4. Create a vector  $\mathbf{y}$  assigning respective class to the elements of  $\mathbf{X}$ .
5. Make a classification rule, based on the **aposteriori probability**, which a *decision vector* saying for each element of  $\mathbf{X}$  whether the element has been generated from  $\mathcal{N}(160, 10^2)$ , or  $\mathcal{N}(185, 10^2)$ , respectively. You can employ `nodrmpdf`, `arrayfun`.
6. Compare the decision vector with the true classes in  $\mathbf{y}$  (enumerate the classification accuracy).

Just upload the functional m-file. No protocol needed for now.