# RPZ Test 7.2.2023

## Jméno:

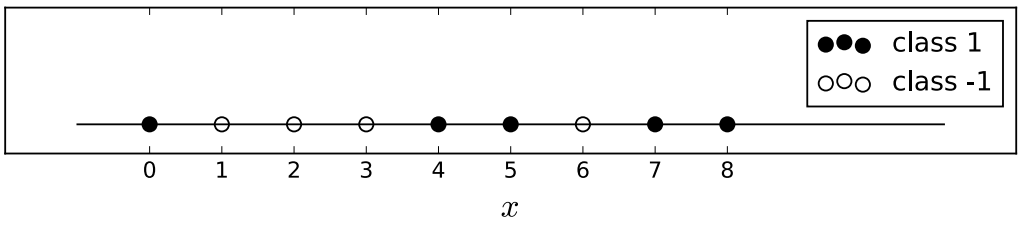
1. (2b) Bayesovská klasifikace. Uvažujte dvě třídy 1 a 2, které mají obě normální rozdělení, $N(\mu_1, \mathbf{C}_1)$ a $N(\mu_2, \mathbf{C}_2)$, a stejné apriorní pravděpodobnosti. Parametry rozdělení jsou:
   $\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $\mathbf{C}_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$, $\mathbf{C}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$.
   Najděte klasifikátor, který minimalizuje chybu klasifikace.

2. (5b) Metoda maximální věrohodnosti (Maximum Likelihood, ML).

   (a) (3b) Životnost žárovek je popsána hustotou pravděpodobnosti $p(t) = \frac{1}{\theta} e^{-t/\theta}, t \in [0, \infty]$, kde $p(t)\delta$ je pravděpodobnost, že žárovka praskne v intervalu $(t, t+\delta)$ $(\delta \to 0)$. Jaký je maximálně věrohodný odhad $\theta$, pokud máme zaznamenán čas selhání šesti žárovek: $t_0 = 0.0, t_1 = 1.0, t_2 = 2.3, t_3 = 2.7, t_4 = 6.0, t_5 = 12.0$?

   (b) (2b) Zařízení zaznamenávající časy selhání žárovek bylo z úsporných důvodů vyřazeno. Žárovky byly napájeny nepřetržitě po dobu $T$ a jedinou informaci, kterou máme k dispozici, je že po této době je $G$ žárovek ještě dobrých (*G*ood) a $F$ žárovek už selhalo (*F*ailed). Najděte maximálně věrohodný odhad $\theta$.

3. (4b) Adaboost. Máme následující jednodimenzionální data:

   a množinu slabých klasifikátorů $h(x) = \text{sign}(ax + b)$ $(a \in \{-1, 1\}, b \in \mathbb{R})$. Vysvětlete na tomto příkladu, jak funguje a co dělá Adaboost (udělejte celou jednu iteraci, končící prvním převálením dat.)

4. (4b) Vysvětlete algoritmus K-means++. Dejte příklad, kde použití K-means++ pomůže.

5. (5b) Rozhodovací stromy.

   (a) (2b) Vysvětlete jednoduchý způsob konstrukce stromu shora-dolů, který používá k výběru rozhodujícího kritéria maximalizaci informačního zisku.

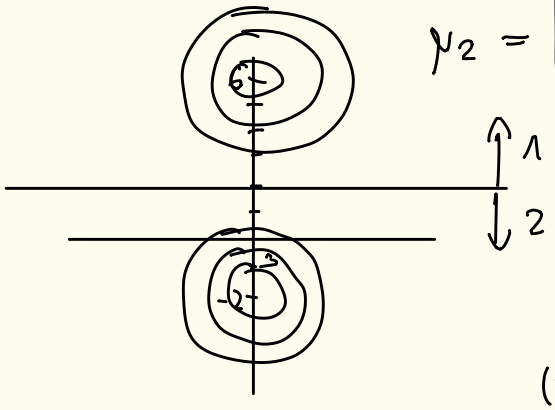
   (b) (3b) Diskutujte výhody a nevýhody rozhodovacích stromů.

(1.) $N(\mu_1, C_1)$
$N(\mu_2, C_2)$

$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$ $\quad C_1 = C_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

$\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

$\begin{pmatrix} 3 \\ 6 \end{pmatrix} - \begin{pmatrix} 3 \\ -2 \end{pmatrix} = \begin{pmatrix} 0 \\ 8 \end{pmatrix} / 2 \Rightarrow \begin{pmatrix} 0 \\ 4 \end{pmatrix} + \begin{pmatrix} 3 \\ -2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$




optimální rozhodnutí pro $y > 2$

$$q([x,y]) \quad (x,y) \in \mathbb{R}^2 \begin{cases} 1 & y > 2 \\ 2 & \text{else} \end{cases}$$

rozhodovací hranice je tedy přímka $\begin{pmatrix} 3 \\ 2 \end{pmatrix} + t \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

(2) (a)
$$p(t) = \frac{1}{\theta} e^{-t/\theta}, \quad t \in [0, \infty)$$

$$T = \{0.0, 1.0, 2.3, 2.7, 6.0, 12.0\}$$

$$L(\theta) = \prod_{t \in T} p(t_i) = \prod_{t \in T} \frac{1}{\theta} e^{-t_i/\theta}$$

$$\ell(\theta) = \sum_{t \in T} \log \frac{1}{\theta} - \frac{t_i}{\theta} = N \log \frac{1}{\theta} - \sum_{t \in T} \frac{t_i}{\theta} =$$

$$= -N \log \theta - \sum_{t \in T} t_i \theta^{-1}$$

$$\frac{\partial \ell}{\partial \theta} = \frac{-N}{\theta} + \sum_{t \in T} t_i \theta^{-2} = \frac{-N}{\theta} + \frac{1}{\theta^2} \sum_{t \in T} t_i \overset{\text{nastavíme}}{=} 0$$

$$-N\theta + \sum_{t \in T} t_i = 0$$

$$\theta = \frac{\sum_{t \in T} t_i}{N} = 24/6 = 4$$

(4) $T = \{6, G, G, B, B, \dots\}$

G good B bad žárovek

alive @ $T = \int_T^\infty p(t)dt =$

$$L(\theta) = \prod_{i=0}^{B} p(t_i) + \prod_{i=0}^{G} e^{-T/\theta} \qquad = \frac{1}{\theta}\int_T^\infty e^{-t/\theta}dt = \left[-e^{-t/\theta}\right]_T^\infty =$$

$$= 0 + e^{-T/\theta}$$

$$l(\theta) = \sum_{i=0}^{B} \log\frac{1}{\theta} - \frac{t_i}{\theta} - \sum_{i=0}^{G}\frac{T}{\theta}$$

$-B\log\theta$

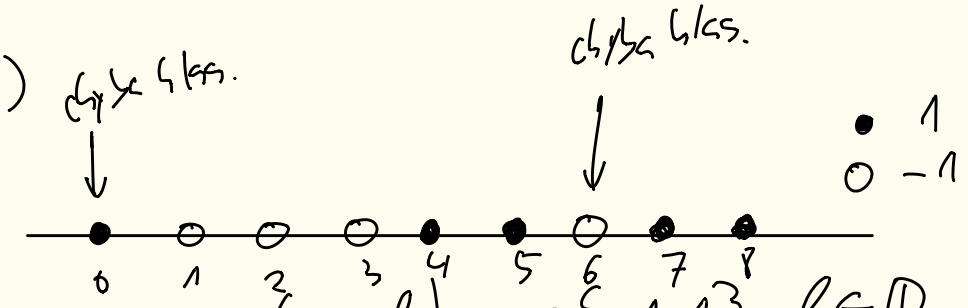
$$= B\log\frac{1}{\theta} - \frac{1}{\theta}\sum_B t_i - \frac{6T}{\theta}$$

$$\frac{\partial l}{\partial\theta} = -\frac{B}{\theta} + \frac{\sum_B t_i}{\theta^2} - \frac{6T}{\theta^2} = 0 \quad /\cdot\theta^2$$

$$= -B\theta + \sum_B t_i - 6T = 0$$

$$\theta = \frac{\sum_B t_i - 6T}{B}$$

(3)




$h(x) = \text{sign}(ax+b) \quad a \in \{-1, 1\} \quad b \in \mathbb{R}$

1. slabý klasifikátor: $a = 1$, $b = 3.5$

Dál = $1/N = 1/9$

$$\varepsilon_1 = \sum_i D_1(i) [\![ h_t(x_i) \neq y_i ]\!]$$

$\varepsilon_1 = 2/9$ → dvě misklasifikace

$$\alpha_1 = 1/2 \ln\left(\frac{1-\varepsilon_1}{\varepsilon_1}\right)$$

$$D_2(i) = D_1(i) \cdot e^{-\alpha_1 y_i h_1(x_i)} \begin{cases} D_1(i) e^{-1/2 \ln 7/2} & \text{pro správně klasifikace} \\ D_1(i) \cdot e^{1/2 \ln 7/2} & \text{pro chybně klasifikace} \end{cases}$$

ok: 0.059 7×

err: 0.208 2×

---

$\sum D_2(i) = 0.829$

$D_2(0,6) = 0.250$

$D_2(1,2,3,4,5,7,8) = 0.071$

---

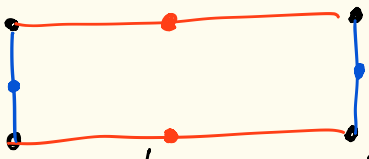
(4) k means ++

funguje stejně jako k means, ale delší chytrou inicializaci centroidů, → vyšší pravděpodobnost centroid dělá do bodů co jsou nejvzdálenější od již vybraných centroidů

$$d_l = \min_{c \in C} \|x_l - c\|$$

$$p(l) = \frac{d_l^2}{\sum_{l \in L} d_l^2}$$

příklad:



při špatné inicializaci nekonverguje k means do glob. minima, ale do lokálního

k means ++ spíše vybere rohové body, které vedou do glob. min.