Cvičení 8

PŘÍZNAKOVÉ METODY ROZPOZNÁVÁNÍ

Implementujte klasifikátor v 2D prostoru. Vstupem bude textový soubor s trénovacími vzorky (příznaky x, y a třída) a testovacími vzorky (příznaky x, y). Cílem je určit třídu testovacích vzorků.

1. Klasifikátor s kritériem minimální vzdálenosti – Nearest Centroid [**2 body**]
   Pro každou třídu je z trénovacích dat vytvořen jeden zástupce (centroid/těžiště). Neznámý vzorek je poté klasifikován dle nejbližšího zástupce. Použijte Eukleidovskou metriku.

2. Klasifikátor dle nejbližšího souseda – K-Nearest Neighbors [**2 body**]
   Třída neznámého vzorku je určena na základě K nejbližších trénovacích vzorků. Pro výpočet volte $K = 1$. Použijte Eukleidovskou metriku.

3. Rozhodovací strom - Decision Tree [**3 body**]
   Využijte tzv. information gain (IG) – množství informace získané o náhodné proměnné (T) na základě pozorování jiné náhodné proměnné (a). Náhodná proměnná T v našem případě představuje třídu a pozorování a představuje příznak.

   K výpočtu lze využít entropie (H) následujícím způsobem:

$$IG(T, a) = \mathrm{H}(T) - \mathrm{H}(T|a) \tag{1}$$

$$H(T) = -\sum_{t \in \mathcal{T}} p(t) \log p(t) \tag{2}$$

$$H(T|a) = \sum_{v \in \mathrm{vals}(a)} \frac{|S_a(v)|}{|T|} \cdot \mathrm{H}\left(S_a(v)\right); \qquad S_a(v) = \{\mathbf{x} \in T | x_a = v\} \tag{3}$$

   Uzel rozhodovacího stromu bude obsahovat podmínku na vybraný příznak (např. příznak a je menší rovno než hodnota t) a skupinu trénovacích dat rozdělí vždy na 2 podskupiny. Uzel může mít tedy nejvýše 2 potomky. To umožní zjednodušit rovnici 3 na:

$$H(T|a, t) = \frac{|T_{a \leq t}|}{|T|} \cdot H(T_{a \leq t}) + \frac{|T_{a > t}|}{|T|} \cdot H(T_{a > t}) \tag{4}$$

   Podmínka bude zvolena ze všech možných v rámci daného uzlu na základě nejlepšího IG. Postup tvorby rozhodovacího stromu během trénování tedy bude:

   (a) výpočet IG dle rovnic 1, 2 a 4 pro všechny možné kombinace a a t v daném uzlu
   (b) ukončovací podmínka (nejlepší možný $IG \leq 0$)
   (c) výběr nejlepší podmínky dle IG
   (d) rozdělení dle podmínky a vytvoření odpovídajících uzlů (podskupin)
   (e) rekurze