0.1 Klasifikace

Klasifikace je úloha, ve které jednotlivé datové položky zařazujeme do předem daných tříd označených indexem $c = 1, 2, \cdots, n_c$, kde n_c je počet tříd.

Přístupů k řešení úlohy klasifikace je celá řada. Jsou to např. algoritmy K-means, DBSCAN, COBWEB a další, shrnuté v systému Weka, volně ke stažení na

http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.

Bayesovský přístup ke klasifikaci nejčastěji využívá modely ve tvaru směsi komponent, o kterých budeme mluvit v Kapitole **??**. Nicméně každý model $f(y_t|\psi_t)$ s diskrétním výstupem $y_t \in \{1, 2, \cdots, n_c\}$ lze využít pro klasifikaci. Hodnota předpovědi $\hat{y}_t$ může být považována za index třídy, do které patří regresní vektor ψ_t. Model tak rozděluje prostor všech možných regresních vektorů do n_c podprostorů a každý z těchto podprostorů tvoří jednu třídu klasifikace. Využití diskrétního modelu pro úlohu klasifikace ukážeme v následujícím příkladě.

Příklad [klasifikace s diskrétním modelem]

Pokračujeme v příkladu **??**.

V našem příkladě je $y_t \in \{1, 2\}$. První hodnota odhadu výstupu $\hat{y}_t = 1$ indikuje regresní vektor (vektor veličin mající vliv na vážnost nehody) z první třídy (lehké nehody) a druhá hodnota $\hat{y}_t = 2$ zařazuje okolnosti nehody do druhé (těžké nehody).

Pro výpočet odhadu výstupu máme vzorec (**??**)

$$f(y_t|\psi_t, d(t-1)) = \int_{\Theta^*} f(y_t|\psi_t, \Theta) f(\Theta|d(t-1)) \, d\Theta.$$

Jestliže přejdeme k bodovým odhadům parametru, tedy vezmeme aposteriorní hp jako Diracův impulz v bodovém odhadu (**??**) $f(\Theta|d(t-1)) = \delta\left(\Theta - \hat{\Theta}_{t-1}\right)$, kde bodové odhady parametrů jsou prvky tabulky (**??**), platí (**??**)

$$f(y_t|\psi_t, d(t-1)) = \int_{\Theta^*} f(y_t|\psi_t, \Theta) \, \delta\left(\Theta - \hat{\Theta}_{t-1}\right) d\Theta = f\left(y_t|\psi_t, \hat{\Theta}_{t-1}\right).$$

Znamená to, že prediktivní hp je přímo dána bodovým odhadem výstupu s dosazeným bodovým odhadem parametrů.

V našem případě, daném tabulkou**??**, množina všech regresních vektorů (z levé části tabulky) je rozdělena podle pravděpodobností hodnot y_t do dvou skupin: $\{1, 3, 4\}$- kde větší pravděpodobnost má $y_t = 1$ a $\{2, 6, 7, 8\}$- kde pravděpodobnější je $y_t = 0$. Regresní vektor 5 je hraniční a může být přiřazen do libovolné skupiny. Skupiny jsou charakterizovány hodnotou předpovídaného výstupu, který jsme určili jako hodnotu s největší pravděpodobností.

Data vypovídají o tom, že lehké nehody se stávají v případech, kdy rychlost byla normální, a počasí a osvětlení byly většinou dobré. Těžké nehody nastávají většinou za velké rychlosti a při počasí a osvětlení spíše špatném.

Logistický model lze rovněž využít pro úlohu klasifikace. V odstavci **??** jsme ukázali, jak lze na základě odhadu logistické regrese z dat $d(t) = \{y(t), \psi(t)\}$ pro libovolně zvolený regresní vektor ψ určit jemu odpovídající predikci $\hat{y}$. Přitom množina $\psi(t)$ nemusí zdaleka obsahovat (v praxi také neobsahuje) všechny konfigurace hodnot regresního vektoru ψ. Odhad logistické regrese a na něm založená predikce tak provádí klasifikaci v prostoru všech konfigurací regresního vektoru ψ, tj. každý regresní vektor ψ přiřadí do jedné ze dvou skupin: první skupina regresních vektorů dává predikci 1 a druhá predikci 2.

Příklad [klasifikace s logistickou regresí]

Navazujeme na příklad **??**.

Do odhadnuté rovnice logistické regrese (**??**) můžeme za regresní vektor postupně dosadit všechny možné hodnoty regresního vektoru a získat tak predikce i pro ty regresní vektory, které ve skutečnosti nebyly naměřeny. Tím získáme optimální odpověď na otázku: "Co bychom obdrželi, kdyby nastalo $\cdots$?". Situace je v následující tabulce

i	ψ_i	$\hat{y}_i$
1	$[1,1,1]$	1
2	$[1,1,2]$	2
3	$[1,2,1]$	2
4	$[1,2,2]$	2
5	$[2,1,1]$	1
6	$[2,1,2]$	1
7	$[2,2,1]$	1
8	$[2,2,2]$	1

Z této tabulky je vidět, že množina všech regresních vektorů (popisujících určitou situaci) se rozpadne na dvě množiny, vzhledem k hodnotě předpovídaného výstupu (určitého příznaku situace). První množina charakterizovaná hodnotou predikce 0 je množina regresních vektorů $\{1, 5, 6, 7, 8\}$ a druhá množina obsahuje regresní vektory $\{2, 3, 4\}$. Tím jsme provedli klasifikaci regresních vektorů (situací) ψ podle hodnoty predikce výstupu (příznaku) y.