Písomka z Princípov dátovej vedy, 18.11.2024 **Meno:**

Poznámka: Pre výberové otázky (typu a), b), c)...) je správna vždy práve jedna odpoveď. Za správnu odpoveď v takých otázkach je $+x$ bodov a za nesprávnu odpoveď $-x$ bodov.
Poznámka: Namiesto desatinných čiarok sú kvôli prehľadnosti v zadaní vždy desatinné bodky.

Príklad 1. Máme údaje o 50 ľuďoch a logistickou regresiou modelujeme, či mali infarkt za posledný rok, v závislosti od veku (v rokoch) a výšky (v cm):

$$g(\pi_i) = \beta_0 + \beta_1 vek_i + \beta_2 vyska_i.$$

Vypočítali sme $\widehat{\beta} = (-6;\ 0.03;\ 0.01)^T$.
(i) (*1 bod*) Model teda hovorí, že čím je človek vyšší, tým je jeho pravdepodobnosť infarktu:

a) vyššia b) nižšia

(ii) (*1 bod*) Akú pravdepodobnosť infarktu za rok model predikuje pre 50-ročného človeka, ktorý má 180 centimetrov?

Príklad 2 (*1 bod*)**.** Skonštruovali sme klasifikačný model na predikovanie, či firma zbankrotuje. V testovacej vzorke máme 80 firiem:

- 30 z nich zbankrotovalo, pričom pre 22 z nich model predikoval bankrot a zvyšným 8 predikoval "nebankrot"
- 50 nezbankrotovalo, pričom 10 z nich model predikoval bankrot a 40 predikoval "nebankrot".

Vypočítajte accuracy (presnosť) modelu (na testovacej vzorke).

Príklad 3. Máme model

$$vynos_i = \beta_0 + \beta_1 maj_i + \beta_2 zam_i + \varepsilon_i,$$

kde $vynos_i$, maj_i a zam_i sú ročné výnosy, majetky a počty zamestnancov 135 firiem. Napíšte predpis funkcie $J(\beta)$, ktorej minimalizáciou získavame odhady parametrov pre tento model pre:
(i) (*1 bod*) lineárnu regresiu, (ii) (*1 bod*) hrebeňovú (ridge) regresiu.
(Nestačí napísať všeobecný predpis $J(\beta)$ pre všeobecný model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$.)

Príklad 4. Chceme zistiť, či liek *Všehodžús* pomáha znižovať (diastolický) tlak. Preto sme zobrali 50 ľudí s vysokým krvným tlakom a mesiac im dávali Všehodžús. Pred začiatkom podávania lieku mali tlak v priemere 98 mmHg a po mesiaci podávania Všehodžúsu bol ich priemerný tlak 84 mmHg.
(i) (*1 bod*) Chceme otestovať, či Všehodžús ozaj pomáha. Napíšte, akú teda treba zvoliť hypotézu H_0 a akú H_1 (hypotézy stačí zapísať slovne, netreba "matematicky").

(ii) (*1 bod*) p-hodnota príslušného testu vyšla 0.035. Môžeme povedať, že Všehodžus štatisticky signifikantne pomáha znižovať tlak?

a) áno b) nie

Príklad 5 (*3 body*)**.** Nižšie sú dva obrázky rozhodovacích hraníc predikčných modelov pre dáta typu $(\mathbf{x}, y)$, kde $\mathbf{x}$ je zo štvorca $[0,1]^2$ a $y \in \{-1, +1\}$. Tmavošedá oblasť zodpovedá $\widehat{y} = -1$, všade inde v rámci $[0,1]^2$ je $\widehat{y} = +1$.
Pre každý z obrázkov určte, či rozhodovacia hranica môže zodpovedať klasifikačnému stromu. Ak nie, napíšte prečo. Ak áno, zapíšte príslušný rozhodovací strom (teda uzly, podmienky...).

