

NMFM308 - Zadání domácí úlohy

RNDr. Jan Vávra, Ph.D.

2025-09-25

Všeobecné pokyny

Zápočet pro předmět NMSA230 bude udělen za odevzdání uspokojivého řešení této domácí úlohy. Při neuspokojivém řešení budete požádáni o přepracování. Spolupráce na řešení je sice vítaná, nicméně **každý vytvoří své vlastní originální řešení, očividné kopírování řešení nebude tolerováno.**

Kvalita vypracování nemusí být vysoká, neočekává se kvalita seminárních či bakalářských prací. Nicméně by bylo dobré každý provedený krok řádně okomentovat: proč definujeme takto, co tímto získáme, co můžeme vyčíst z tabulek či obrázků apod. Snažte se ukazovat jen podstatné části kódu (definice klíčových proměnných, atd.), méně důležité části R-kového kódu můžete zatajit (kreslení obrázků, apod.).

Řešení zašlete ve formátu PDF stvořeného (ideálně) pomocí R markdownu na e-mailovou adresu **vavraj@karlin.mff.cuni.cz** do **xx. ledna 2026** (bude upřesněno).

Načtení dat a základní informace o nich

Data si můžete stáhnout ve formátu csv z mé webové stránky:

```
web_domaci <- "https://www.karlin.mff.cuni.cz/~vavraj/Rko/domaci/"
data <- read.csv(paste0(web_domaci, "california_house_price.csv"))
```

Jedná se o data zaměřená na cenu bydlení v Kalifornii (proměnná `median_house_value`). Každý z řádků odpovídá jednomu bloku, pro něj jsou pozorované veličiny určitým způsobem agregovány přes domácnosti v bloku. Význam a definice proměnných zahrnutých v datasetu naleznete v následujícím textovém dokumentu:

```
download.file(paste0(web_domaci, "info_california_house_price.txt"))
```

Cílem bude explorativní analýzou nalézt, které veličiny obsažené v datasetu mají zřejmě vliv na cenu domácnosti.

Podrobnější instrukce

Jednotlivé body jsou uspořádané vzestupně zhruba podle toho, jak budeme postupovat ve výuce. Doporučuji tedy na úloze **pracovat průběžně** a nedělat vše naráz na poslední chvíli. I když jsou zadání bodů 5. a 6. krátká, tak se jedná o stěžejní část úkolu, kterou není radno odbýt.

1. Seznamte se s proměnnými. Roztřídte si je na numerické a kategoriální. Numerické si převedte do vhodných jednotek. Pro kategoriální zaveďte vhodné kategorizace. Nadefinujte si pro ně hezké popisky včetně jednotek či kategorií. Odstraňte řádky s chybějícími hodnotami. Uznáte-li za vhodné, nadefinujte si nové smysluplné veličiny (např. počet lidí na ložnici či místnost).
2. Zapřemýšlejte, jak ze sloupců udávajících polohu napočítat vzdálenost od center známých měst (Los Angeles, San Francisco). Zaveďte si binární veličinu, zda se blok nachází blíže Los Angeles či San Francisku. Zaveďte si kategoriální veličinu udávající, zda je blízko či daleko od těchto měst.
3. Číselně popište a graficky znázorněte rozdělení mediánové ceny domácnosti. Jaká je její maximální hodnota? Které bloky jí dosahují?

4. Vezměte si podmnožinu těchto domácností a zakreslete jejich polohu, barvou odlište kategoriální veličiny z bodu 2. Zvýrazněte blok sestávající se z jediné domácnosti. Co dalšího o tomto bloku víte? Porovnejte tuto podmnožinu od zbytku, co se týče rozdělení mediánového příjmu a vzdálenosti od moře.
5. Vhodnými statistikami, tabulkami a obrázky popište celý soubor dat. Především se zaměřte na vztah proměnných k mediánové ceně domácnosti (vynechte `longitude` a `latitude`).
6. Nalezněte dvě podmnožiny celého souboru, které se liší ve vlivu nějaké proměnné na mediánovou cenu domácnosti. Doložte obrázky či tabulkami.
7. Do jednoho obrázku zakreslete polohu všech bloků. Barevně odstupňujte mediánovou cenu domácnosti (zkuste transparentní barvy `alpha < 1`). Velikost symbolu (v intervalu 0.3 až 2) nastavte lineárně od logaritmu celkové populace v bloku. Typ symbolu zvolte dle nějaké kategoriální proměnné. Přidejte legendu.
8. Definujte a vykreslete následující funkci dvou proměnných - zeměpisné délky a šířky. Pro dané x a y je $f(x, y)$ rovna průměru mediánové ceny domácnosti v blocích do vzdálenosti z od polohy $[x, y]$. Hodnotu funkce dodefinujte nulou, pakliže poblíž není žádná domácnost. Hodnotu parametru z si vhodně nastavte dle svého uvážení. *Kdyby výpočet trval dlouho, tak uvažujte jen náhodnou podmnožinu dostupných bloků.*
9. Nastavte si generátor náhodných čísel svým studentským číslem. Rozdělte dataset náhodně na trénovací (80 %) a testovací (zbylých 20 %).
10. Vezměte data z trénovací sady. Nalezněte přímkou a jinou parametricky zadanou křivku, která *nejlépe* prokládá vztah mezi mediánovou cenou a příjmem domácnosti. Pro zpestření (tedy nepovinné) zkuste proložit různými přímkami (křivkami) dle nějaké binární proměnné. Odhadněte si *průměrné rozptýlení okolo křivky* pomocí

$$\frac{1}{n} \sum_{i=1}^n \left(Y_i - g(X_i, \hat{\beta}) \right)^2,$$

kde Y_i je mediánová cena domácnosti, X_i je mediánový příjem domácnosti, $g(\cdot, \hat{\beta})$ je ona optimální křivka a $i = 1, \dots, n$ jsou indexy domácností v trénovacím datasetu. Která křivka dosahuje nejnižšího rozptýlení?

11. Vezměte data z testovací sady. Pro danou hodnotu příjmu (a kategoriální proměnné) predikujte pomocí odhadnuté přímky (křivky) cenu domácnosti. Zvolte vhodnou souhrnou metriku měřící vzdálenost vaší predikce od skutečné ceny domácnosti. Která křivka dle této metriky vychází nejlépe? Jak by obstálo řešení, ve kterém byste predikovali pouze průměrnou hodnotou mediánové ceny domácnosti (tj. vodorovnou přímkou)?
12. Spojte data opět dohromady (trénovací a testovací) a odhadněte ten nejlepší model (dle předchozího bodu) znovu na celých datech. Vezměte odhadnuté parametry křivky $\hat{\beta}$ a hodnoty X_i a nagenerte si náhodně nové $Y_i^{(1)}$ z normálního rozdělení se střední hodnotou $g(X_i, \hat{\beta})$ a rozptylem odhadnutým dle bodu 10 (nyní je již index i chápán pro celý dataset). Odhadněte tu samou křivku znovu ale pomocí generovaných hodnot $\mathbf{Y}^{(1)}$. Zakreslete (a okomentujte) vedle sebe bodové diagramy (s proloženými křivkami) pro skutečné dvojice (X_i, Y_i) a generované dvojice $(X_i, Y_i^{(1)})$.
13. Postup z předchozího bodu zopakujte $1000 \times$ (vždy stejné $\mathbf{X}$ a $\hat{\beta}$), čímž získáte 1000 hodnot parametru $\beta^{(1)}, \dots, \beta^{(1000)}$. Jak dlouho Vám trval tento výpočet? Graficky znázorněte tyto nasimulované hodnoty (marginálně i sdruženě). Najděte (a přikreslete) krajní kvantily takové, aby utvořily interval pokrývající 95 % simulovaných hodnot.