Projekt z Lineárneho Programovania:
A04 Predikcia kvality vína, lineárna regresia pomocou l_1, l_∞

Lineárna regresia je jedna z najbežnejších metód na predikovanie. Pomocou nej vieme určiť ako „najlepšie“ možno vyjadriť výstupnú (závislú) premennú ako lineárnu (resp. afinnú) kombináciu známych vstupných (nezávislých) premenných. Závislá premenná sa zvyčajne označuje y a nezávislé premenné sa zvyčajne označujú $x_1, x_2, \ldots, x_k$, pričom k je počet nezávislých premenných. Vo všeobecnosti sú $y, x_1, \ldots, x_k \in \mathbb{R}^n$. Vašou úlohou bude nájsť koeficienty $\beta_0, \beta_1, \ldots, \beta_k$ tak, aby predikovaný vektor hodnôt

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

bol čo najbližšie k výstupu y. Vektor $y - \hat{y}$ nazývame vektorom chýb merania alebo vektorom rezíduí.

Na meranie „blízkosti“ vektorov y a $\hat{y}$, resp. dĺžky vektora rezíduí možno použiť rôzne normy. Napr. aplikovanie euklidovskej normy by viedlo k nelineárnemu problému najmenších štvorcov. My budeme pracovať s l_1 a l_∞ normami, pretože vtedy sa dá problém previesť na úlohu lineárneho programovania. Budeme teda uvažovať úlohy typu

$$\min \|y - \hat{y}\|_1, \quad \min \|y - \hat{y}\|_\infty.$$

A **Formulácia úloh lineárneho programovania.** Predpokladajte, že $y, x_1, \ldots, x_k$ sú dané. Naformulujte vyššie uvedené minimalizačné úlohy ako úlohy lineárneho programovania s premennými $\beta_0, \beta_1, \ldots, \beta_k$.

B **Implementácia a grafické znázornenie.** Implementujte LP formulácie z časti A. Otestujte ich pre $k = 1$ a dátové body x, y zo súboru `A04 plotregres.npz`, t.j. vyriešte obe úlohy a nájdite optimálne β_0, β_1. Vykreslite obrázok s dátovými bodmi a oboma „regresnými priamkami“ $y = \beta_0 + \beta_1 x$.

C **Predikcia kvality vína.** Orley Ashenfelter[1] predikoval cenu červeného vína z regiónu Bordeaux. Tieto vína chutia lepšie, keď sú staršie, preto sa mladé vína uskladňujú, kým dozrejú. Je však ťažké určiť budúcu kvalitu Bordeaux ochutnaním mladého vína, keďže časom sa chuť a vlastnosti signifikantne zmenia – dokážu to len experti. Ashenfelter však vypozoroval, že zlé vína sú často nadcenené a dobré vína sú niekedy podcenené a rady expertov na víno robia trh s vínom neefektívnym. Preto zostavil vlastný model pre oceňovanie vína, ktorý zohľadňuje aj niekoľko aspektov súvisiacich s počasím v regióne Bordeaux.[2]

Pomocou dát zo súboru `A04 wine.csv` a vašich l_1 a l_∞ modelov z časti A sa pokúste predikovať kvalitu vína podobne ako Ashenfelter. Nezávislými premennými budú: *Winter Rain, Average Growing Season temperature (AGST)*, *Harvest Rain* a *Age of Vintage*. ~~Ne~~závislou premennou bude *Price*.

[1] https://www.youtube.com/watch?v=8WMRj9mTQtI

[2] V čase publikovania modelu si Ashenfelter vyslúžil veľa kritiky avšak neskôr sa jeho predikcie ukázali byť správne.

D **Počítanie R^2.** Pre oba modely spočítajte tzv. koeficient determinácie (R-kvadrát)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

kde $\bar{y}$ je priemer závislých premenných. R^2 nadobúda hodnoty v intervale $[0, 1]$ a hovorí, aký podiel rozptylu závislej premennej je vysvetlený nezávislými premennými Čím sú hodnoty R^2 bližšie k 1, tým viac sa spravidla považuje za lepší. Ashenfelterov model mal hodnotu 0,83.

E **Nadstavba.** Vymyslite rozšírenie alebo modifikáciu projektu, napr. porovnajte výsledky s metódou najmenších štvorcov, alebo experimentujte s účelovou funkciou.