

# 1 Nadstavba

## 1.1 Spracovanie všeobecnej triedy pre $L^1$ a $L^\infty$ lineárnu regresiu

Vypracovali sme modul `Model` pre počítanie  $L^1$  a  $L^\infty$  lineárnej regresie z ľubovoľných číselných dát, ktorý využíva LP formulácie popísané vyššie. Konkrétne `L1Model` využíva formuláciu na minimalizovanie  $L^1$  normy a `LInfModel` minimalizuje  $L^\infty$  normu. Príklad použitia tohto modelu sa nachádza v `model_demonstration.ipynb`. Následne opíšeme jednotlivé metódy jednotlivých modelov.

`Model.__init__(dependent_vect, independent_vect)`

Konštruktor triedy, spoločný pre oba modely, vytvorí inštanciu, ktorá si drží dáta a vie na nich vykonávať operácie popísané nižšie.

Argumenty:

- `dependent_vect`: `np.ndarray` - vektor závislých premenných
- `independent_vect`: `np.ndarray` - matica, ktorej riadky sú vektory nezávislých premenných

`Model.solve()`

Metóda, ktorá vyrieši lineárnu regresnú LP úlohu na daných dátach. `L1Model.solve()` rieši minimalizáciou  $L^1$  normy a `LInfModel.solve()`, rieši minimalizáciou  $L^\infty$  normy.

Vracia:

- `np.ndarray` - vektor optimálnych  $\beta$  premenných

Po zavolaní tejto metódy si inštancia uloží vektor optimálnych  $\beta$  premenných do atribútu `self._beta`, potrebné pre metódy popísané nižšie.

`Model.r2()`

Vypočíta  $R^2$  koeficient pre dané dáta a vypočítaný vektor  $\beta$ .

Vracia:

- `float` - výsledný  $R^2$  koeficient

`Model.visualize()`

Ak je počet nezávislých premenných 1 alebo 2, táto metóda vykreslí graf dát spolu s vypočítanou regresnou priamkou, resp. rovinou.

Vracia:

- `bool` - úspešnosť vizualizácie, kde `False` označuje, že nezávislých premenných je viac ako 2, čiže nie je možné vykresliť graf

## 1.2 Porovnanie použitia $L^1$ a $L^\infty$ lineárnej regresie

Nasledujúce tvrdenia popisujú len naše pozorovania správania sa jednotlivých lineárnych regresii na generovaných dátach

Vyššie v sekcii ?? sme ukázali, že implementácie lineárnej regresie pomocou merania vzdialenosti  $L^1$  a  $L^\infty$  normou majú optimálne riešenie, pre ľubovoľné vstupné dáta. Snažili sme sa odpozorovať, ako sa jednotlivé prístupy odlišujú pre nejaké konkrétne dáta.

V dátach, v ktorých je výrazná lineárna závislosť, minimalizovanie  $L^1$  normy veľmi dobre zachytáva práve tento lineárny vzťah, aj v prítomnosti odľahlých dát - outlierov. Toto správanie vie ale viesť aj k tzv. *overfittingu*. Model príliš tesne zachytáva takéto správanie, čo môže viesť k horším odhadom pre budúce pozorovania.

Na druhej strane minimalizovanie  $L^\infty$  normy je veľmi ovplyvňované outliermi. Aj pre „jasné“ lineárne dáta s nejakými chybnými pozorovaniami, tieto dátové body výrazne odklonia regresnú priamku/nadrovinu.

### 1.2.1 Minimalizácia váženého súčtu

Toto správanie  $L^\infty$  lineárnej regresie môžeme využiť na zníženie *overfittingu*  $L^1$  lineárnej regresie. Jeden z možných prístupov môže byť napríklad pomocou minimalizácie váženej sumy  $\omega \|y - \hat{y}\|_1 + (1 - \omega) \|y - \hat{y}\|_\infty$ ,  $\omega \in [0; 1]$ . Formulovaná LP úloha vyzerá nasledovne (značenie sme prebrali z (??) a (??)):

$$\begin{aligned} \min \quad & \left( \begin{array}{c|c|c} \mathbf{0}_{k+1}^T & \omega \mathbf{1}_n^T & (1 - \omega) \end{array} \right) \left( \begin{array}{c} \beta \\ t \\ \gamma \end{array} \right) \\ & \left( \begin{array}{c|c|c} \mathbf{A} & \mathbb{I}_n & \mathbf{0}_n \\ -\mathbf{A} & \mathbb{I}_n & \mathbf{0}_n \\ \hline \mathbf{A} & \mathbf{0}_{n \times n} & \mathbf{1}_n \\ -\mathbf{A} & \mathbf{0}_{n \times n} & \mathbf{1}_n \end{array} \right) \left( \begin{array}{c} \beta \\ t \\ \gamma \end{array} \right) \geq \left( \begin{array}{c} y \\ -y \\ y \\ -y \end{array} \right) \\ & \beta \in \mathbb{R}^{k+1}, t \geq \mathbf{0}_n, \gamma \geq 0 \end{aligned}$$

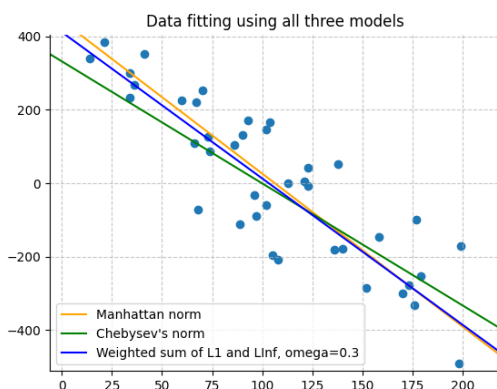
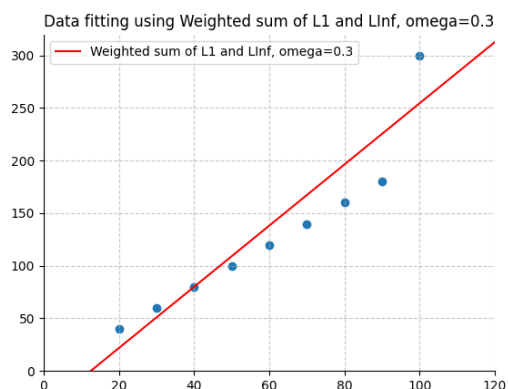
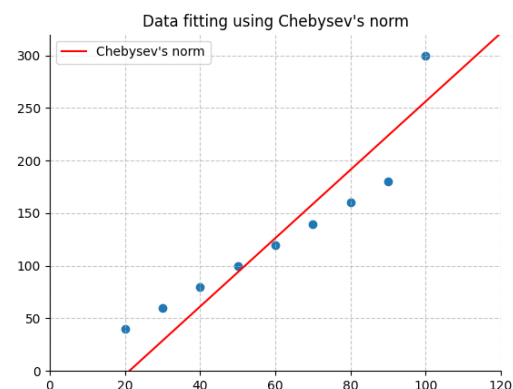
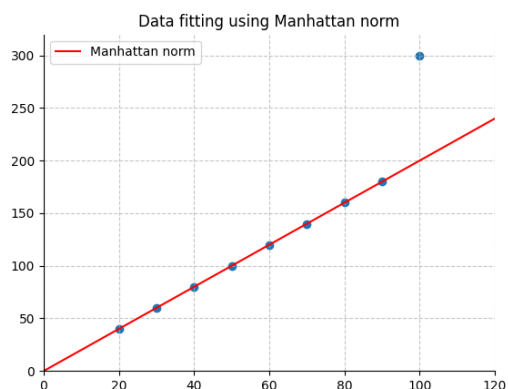
Podobným spôsobom ukážeme, že táto úloha nadobúda optimálne riešenie. Sformulujme duálnu úlohu:

$$\begin{aligned} \text{Nech } \alpha &= \left( \begin{array}{c} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{array} \right), \alpha_{1,2,3,4} \in \mathbb{R}^n \\ \max \quad & \left( \begin{array}{c|c|c|c} y^T & -y^T & y^T & -y^T \end{array} \right) \alpha \\ & \left( \begin{array}{c|c|c|c} \mathbf{A}^T & -\mathbf{A}^T & \mathbf{A}^T & -\mathbf{A}^T \end{array} \right) \alpha = \mathbf{0}_{k+1} \\ & \left( \begin{array}{c|c|c|c} \mathbb{I}_n & \mathbb{I}_n & \mathbf{0}_{n \times n} & \mathbf{0}_{n \times n} \end{array} \right) \alpha \leq \omega \mathbf{1}_n \\ & \left( \begin{array}{c|c|c|c} \mathbf{0}_n^T & \mathbf{0}_n^T & \mathbf{1}_n^T & \mathbf{1}_n^T \end{array} \right) \alpha \leq 1 - \omega \\ & \alpha \geq \mathbf{0}_{4n} \end{aligned}$$

Vidíme, že primárna úloha je prípustná pre  $\beta = \mathbf{0}_{k+1}$ ,  $t = |y|$ ,  $\gamma = |\hat{y}|$  (využitím značenia ako v ?? a ??) a duálna úloha je prípustná pre  $\alpha = \mathbf{0}_{4n}$ , teda zo silnej duality obe riešenia nadobúdajú optimálne riešenie.

## 1.2.2 Implementácia `WeightedL1LInfModel`

Takáto lineárna regresia je implementovaná v triede `WeightedL1LInfModel`. Jej používanie je rovnaké ako pri predchádzajúcich implementáciách. Jediná zmena je pre metódu `WeightedL1LInfModel.solve(omega)`, ktorá teraz očakáva parameter `omega` : `float` v intervale  $[0; 1]$ .



Porovnanie správania sa jednotlivých lineárnych regresii, prvé tri grafy zobrazujú rovnaké lineárne dáta s jedným outlierom