1 Nadstavba

1.1 Spracovanie všeobecnej triedy pre L^1 a L^∞ lineárnu regresiu

Vypracovali sme modul `Model` pre počítanie L^1 a L^∞ lineárnej regresie pre ľubovoľné číselné dáta, ktorý využíva LP formulácie popísané v sekciách vyššie. Konkrétne `L1Model` využíva formuláciu na minimalizovanie L^1 normy a `LInfModel` minimalizuje L^∞ 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^∞ normy.

Vracia:

- `np.ndarray` - vektor optimálnych β premenných

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

```
Model.r2()
```

Vypočíta R^2 koeficient pre dané dáta a vypočítaný vektor β.

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^∞ lineárnej regresie

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

Vyššie v sekcii **??** sme ukázali, že implementácie lineárnej regresie pomocou merania vzdialenosti L^1 a L^∞ 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^∞ normy je veľmi ovplyvňované outliermi. Aj pre „jasne“ lineárne dáta s nejakými chybnými pozorovaniami, tieto dátové body výrazne odklonia regresnú priamku/nadrovinu.

1.2.1 Minimalizácia váženého súčtu

Toto správanie L^∞ lineárnej regresie sa môžeme pokúsiť 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 (**??**)):

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

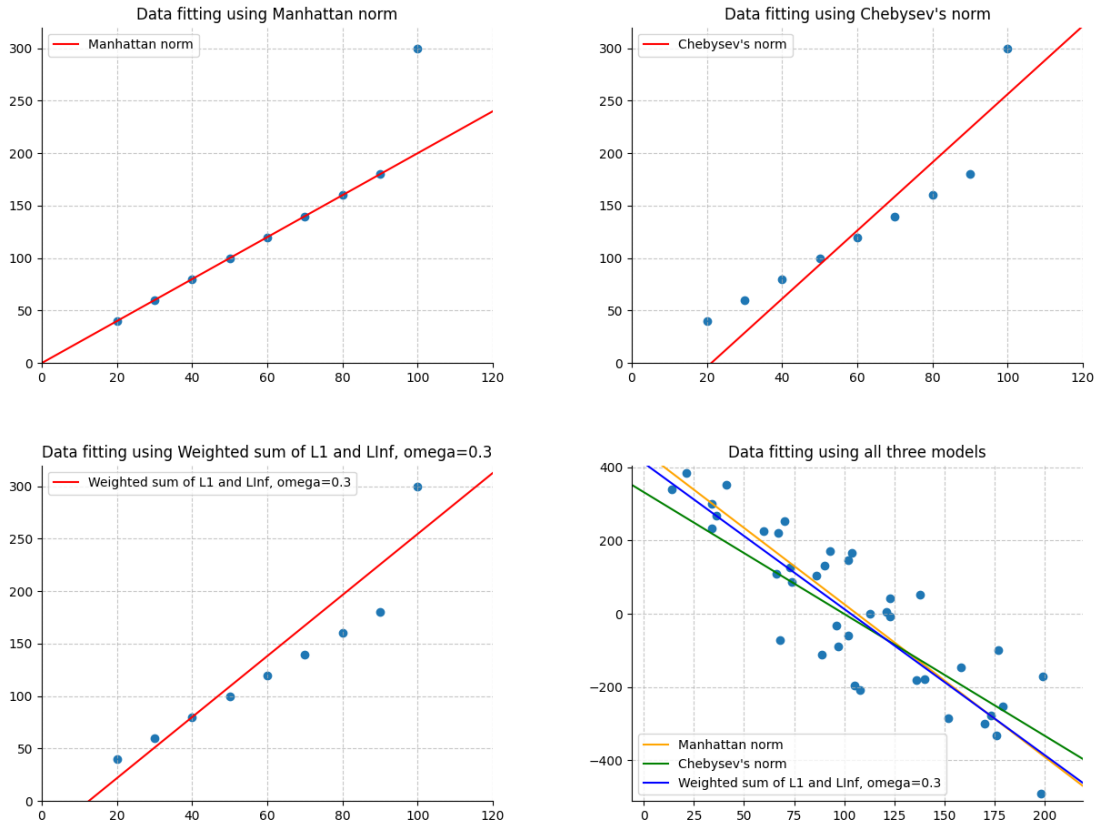
Podobným spôsobom ukážeme, že táto úloha nadobúda optimálne riešenie. Sformulujme duálnu úlohu:

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

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 slabej 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 očakáva parameter `omega:float`, pričom akceptuje iba `omega`$\in [0; 1]$.



Porovnanie správania sa jednotlivých lineárnych regresii, prvé tri grafy zobrazujú rovnaké lineárne dáta s jedným outlierom, štvrtý zobrazuje použitie všetkých troch lineárnych regresií na lineárnych dátach s náhodným šumom