A04 – Predikcia kvality vína, lineárna regresia pomocou L^1, L^∞

Piati proti optimalizácii

Tomáš Antal, Erik Božík, Róbert Kendereš,

Teo Pazera, Andrej Špitalský

2DAV

Január 2024

Predstavenie projektu – lineárna regresia

- lineárna regresia – predikcia $y \in \mathbb{R}^n$ lineárnou kombináciou $x_1, \ldots, x_k \in \mathbb{R}^n$

$$\min ||y - \hat{y}||$$

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

atribúty	x_1	x_2	...	x_k	y
pozorovanie 1	1	0.84	...	121	4.25
⋮	⋮	⋮	⋮	⋮	⋮
pozorovanie n	4	0.12	...	117	5.68

- vyjadriteľné ako úloha lineárneho programovania – L^1, L^∞

Predstavenie projektu – obsah

- formulácia LP úloh
- implementácia v `Python`-e a predikcia kvality vína
- počítanie a interpretácia R^2 koeficientu
- implementácia všeobecnej triedy na počítanie L^1 a L^∞ lineárnej regresie
- minimalizácia váženej sumy noriem

Formulácia úloh lineárneho programovania

Úloha

Nájsť koeficienty $\beta_0, \beta_1, \ldots, \beta_k$ tak, aby predikovaný vektor

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{1}$$

bol čo najbližšie k výstupu y, kde y označuje závislú premennú a $x_1, x_2, \ldots, x_k \in \mathbb{R}^n$ označujú nezávislé premenné. Túto vzdialenosť $||y - \hat{y}||$ sme minimalizovali L^1 a L^∞ normami

Minimalizovanie L^1 normy

Chceme minimalizovať normu $||y - \hat{y}||_1$

označíme:

$$\begin{aligned} \mathbf{A} &:= (\mathbf{1_n}, x_1, \ldots, x_k) \\ \beta &:= (\beta_0, \beta_1, \ldots, \beta_k)^T \end{aligned} \tag{2}$$

Problém prevedieme do tvaru:

$$\text{min } c^T x$$

$$\mathbf{A}x \geq b$$

Minimalizovanie L^∞ normy

Zavedieme nový vektor $t \in \mathbb{R}^n$, ktorým ohraničíme $y - \mathbf{A}\beta$

Minimalizovanie L^1 normy ako úloha lineárneho programovania:

$$\min \left(\mathbf{0}_{k+1}^T \,\middle|\, \mathbf{1}_n^T \right) \begin{pmatrix} \beta \\ \hline t \end{pmatrix}$$

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbb{I}_n \\ \hline -\mathbf{A} & \mathbb{I}_n \end{array}\right) \begin{pmatrix} \beta \\ \hline t \end{pmatrix} \geq \begin{pmatrix} y \\ \hline -y \end{pmatrix}$$

$$\beta \in \mathbb{R}^{k+1},\ t \geq \mathbf{0}_n$$

Minimalizovanie L^∞ normy

Chceme minimalizovať normu $||y - \hat{y}||_\infty$

Zavedieme skalárnu premennú $\gamma \in \mathbb{R}$, prevedieme na úlohu LP

$$-\gamma \mathbf{1}_n \leq y - \mathbf{A}\beta \leq \gamma \mathbf{1}_n$$

Pomocou značenia z (2), výsledná úloha:

$$\min \left(\mathbf{0}_{k+1}^T \;\middle|\; 1 \right) \begin{pmatrix} \beta \\ \hline \gamma \end{pmatrix}$$

$$\left(\begin{array}{c|c} \mathbf{A} & \mathbf{1}_n \\ \hline -\mathbf{A} & \mathbf{1}_n \end{array}\right) \begin{pmatrix} \beta \\ \hline \gamma \end{pmatrix} \geq \begin{pmatrix} y \\ \hline -y \end{pmatrix}$$

$$\beta \in \mathbb{R}^{k+1}, \; \gamma \geq 0$$

Implementácia

- upravený tvar úlohy pre solver

$$\min\ c^T x$$

$$A_{ub} x \leq b_{ub}$$

$$A_{eq} x = b_{eq}$$

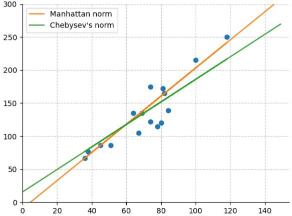
$$x \in [l, u] \qquad\qquad l \leq u;\ l, u \in (\mathbb{R} \cup \{-\infty, \infty\})^n$$

Implementácia

```
c = np.concatenate(([0]*(k + 1), np.ones(n)))
A = np.block([np.ones((n, 1)), np.array(x.values)])
I = np.identity(n)

A_ub = np.block([[-A, -I], [A, -I]])
b_ub = np.concatenate([-y, y])
bounds = [(None, None)]*(k + 1) + [(0, None)] * n
```

Riešenie úlohy a vizualizácia

```
solve = linprog(c, A_ub, b_ub,
                bounds=bounds)
betas = solve.x[:k+1]
```



priamky L^1 a L^∞ lineárnych regresií pre arbitrárne dáta

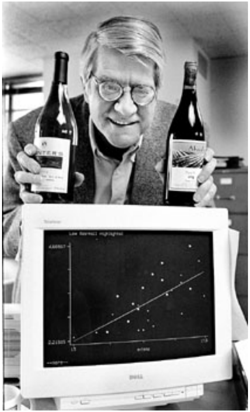
Predikcia kvality vína - dáta

Nezávislé premenné

- množstvo dažďa v zime
- priemerná teplota počas zretia vína
- množstvo dažďa počas zberu
- vek vína
- populácia Francúzska

Závislá premenná

- cena



Orley Ashenfelter

Predikcia kvality vína - výsledky

L^1

- $+$ vplyv teploty počas zretia
- $+$ vplyv veku vína
- $-$ vplyv dažďu počas zberu
- $+$ vplyv dažďu počas zimy
- $-$ vplyv populácie Francúzska

L^∞

- rovnaké poradie ako L^1
- ale $-$ vplyv veku vína

$$\beta_0^{(1)} \approx -8.88 \cdot 10^{-1} \quad \beta_1^{(1)} \approx 1.58 \cdot 10^{-3} \quad \beta_2^{(1)} \approx 5.21 \cdot 10^{-1}$$

$$\beta_3^{(1)} \approx -4.51 \cdot 10^{-3} \quad \beta_4^{(1)} \approx 1.13 \cdot 10^{-2} \quad \beta_5^{(1)} \approx -2.21 \cdot 10^{-5}$$

R^2 – koeficient determinácie

- typicky hodnota z intervalu [0, 1]
- podiel rozptylu závislej premennej zachytený modelom
- čím bližšie k 1, tým lepšie vysvetľuje rozptyl

R^2 – koeficient determinácie

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

- rozdiely medzi skutočnými hodnotami y a predpovedanými
- rozdiely medzi skutočnými hodnotami y a priemerom (rozptyl)

- ukazuje, aký podiel rozptylu závislej premennej je vysvetlený nezávislými premennými.

Výsledky pre naše predikcie

- regresie L^1, L^∞
- koeficienty pre obe normy:

$$R^2_{(1)} \approx 0.78813$$
$$R^2_{(\infty)} \approx 0.80649$$

- obe dostatočne zachytávajú rozptyl

Všeobecná trieda pre L^1 a L^∞ lineárnu regresiu

```
from models.models import L1Model, LInfModel
```

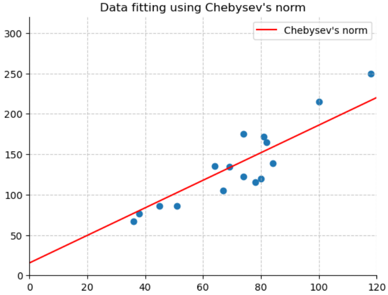
- zovšeobecnenie problému
- voľnosť dimenzionality
- vstupný vektor y a matica $\mathbf{X}$
- hodnoty β výstupom

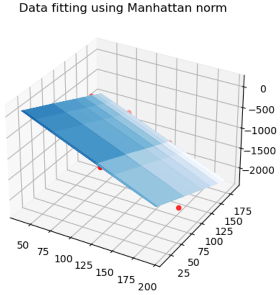
```
# inicializacia
model1 = L1Model(y, X)
model2 = LInfModel(y, X)
# riesenie
beta1 = model1.solve()
beta2 = model2.solve()
```

Všeobecná trieda pre L^1 a L^∞ lineárnu regresiu

- hodnota R^2
- vizualizácia pre 2D a 3D

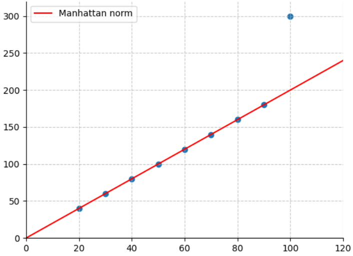
```
model.r2()

model.visualize()
```



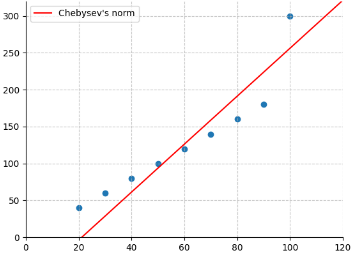


Porovnanie L^1 a L^∞ lineárnej regresie

- L^1 – veľmi dobre zachytáva lineárny vzťah, môže viesť k *overfittingu*
- L^∞ – príliš ovplyňovaná outliermi



regresia L^1 normou



regresia L^∞ normou

Minimalizácia váženého súčtu noriem

- redukcia *overfittingu* L^1 regresie váženým súčtom s L^∞ normou

$$\min\ \omega||y - \hat{y}||_1 + (1-\omega)||y - \hat{y}||_\infty,\ \omega \in [0;1]$$

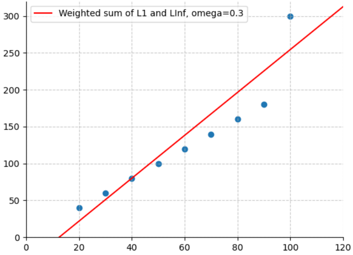
- stále implementovateľné ako úloha lineárneho programovania

$$\text{min}\ \left(\mathbf{0}_{k+1}^T \;\middle|\; \omega\mathbf{1}_n^T \;\middle|\; (1-\omega)\right) \left(\begin{array}{c} \beta \\ \hline t \\ \hline \gamma \end{array}\right),\ \omega \in [0;1]$$

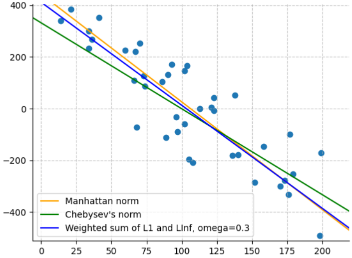
$$\left(\begin{array}{c|c|c} \mathbf{A} & \mathbb{I}_n & \mathbf{0}_n \\ \hline -\mathbf{A} & \mathbb{I}_n & \mathbf{0}_n \\ \hline \mathbf{A} & \mathbf{0}_{n\times n} & \mathbf{1}_n \\ \hline -\mathbf{A} & \mathbf{0}_{n\times n} & \mathbf{1}_n \end{array}\right) \left(\begin{array}{c} \beta \\ \hline t \\ \hline \gamma \end{array}\right) \geq \left(\begin{array}{c} y \\ \hline -y \\ \hline y \\ \hline -y \end{array}\right)$$

Minimalizácia váženého súčtu noriem

- implementované ako `WeightedL1LInfModel`



regresia váženým súčtom noriem



porovnanie troch regresií

Zhrnutie

- formulácia lineárnej regresie ako úlohy LP
- predikcia ceny vín
- jednoduchý framework na počítanie lineárnej regresie pomocou L^1 a L^∞ noriem, resp. ich váženej sumy

Ďalšie kroky

- analýza časovej zložitosti, napr. voči najmenším štvorcom
- porovnanie vhodnosti jednotlivých prístupov podľa vstupných dát