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 závislej premennej $y \in \mathbb{R}^n$

  pomocou nezávislých $x_1, \dots, x_k \in \mathbb{R}^n$

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

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

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

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

Predstavenie projektu – obsah

- formulácia LP úloh a dokázanie optimality
- 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:

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

Problém prevedieme do tvaru:

$$\min\ c^T x$$

$$Ax \geq b$$

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

Minimalizovanie l_1 normy ako úloha lineárneho programovania:

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

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

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

Minimalizovanie l_∞ normy

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

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

$$-\gamma 1_n \leq y - A\beta \leq \gamma 1_n$$

Pomocou značenia z (2), výsledná úloha:

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

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

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

B

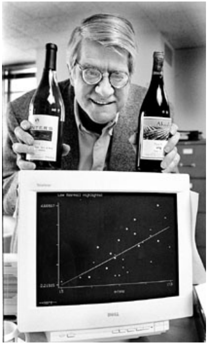
...

Predikcia kvality vína

dáta o víne

- množstvo dážďa v zime
- priemerna teplota počas zretia vína
- množstvo dažďa počas zberu
- vek vína
- populácia Francúzska
- cena

Orley Ashenfelter



Výsledky predikcie

L^1

- + vplyv teplota počas zretia
- + vplyv veku vína
- - vplyv dážď počas zberu
- + vplyv dážd počas zimy
- - vplyv populácie Francúzska

L^∞

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

D

...

Nadstavba

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

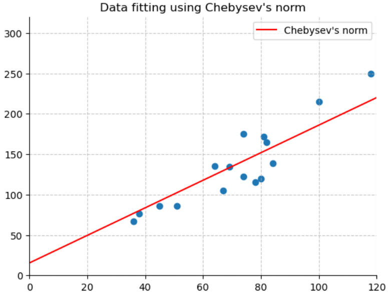
- Zovšeobecnenie problému
- Voľnosť dimenzionality
- Vstupný vektor y a matica X
- Hodnoty β výstupom

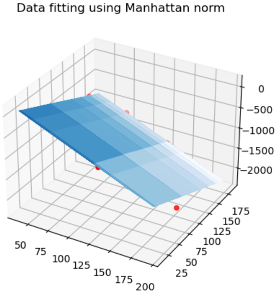
```python
# inicializacia
model1 = L1Model(y, X)
model2 = LInfModel(y, X)
# riesenie
beta1 = model1.solve()
beta2 = model2.solve()
```

Nadstavba

- Hodnota R^2
- Vizualizácia pre 2D a 3D

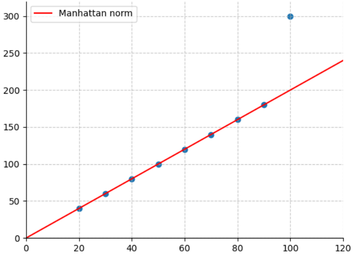
```
model.r2()
model.visualize()
```



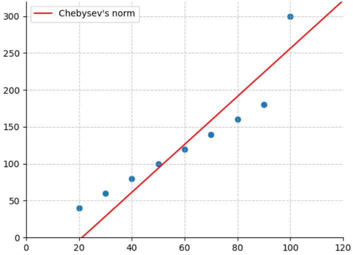


Porovnanie L^1 a L^∞ lineárnej regresie

- 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

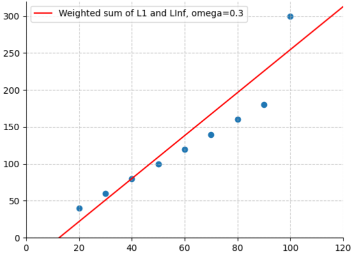
$$\min \ \left(0_{k+1}^T \,\middle|\, \omega 1_n^T \,\middle|\, (1-\omega) \right) \left(\begin{array}{c} \beta \\ \hline t \\ \hline \gamma \end{array} \right), \ \omega \in [0;1]$$

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

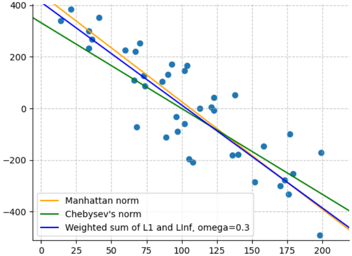
$$\beta \in \mathbb{R}^{k+1}, \ t \geq 0_n, \ \gamma \geq 0$$

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