

# A04 – Predikcia kvality vína, lineárna regresia pomocou $L^1$ , $L^\infty$

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

- ▶ vyjadriteľné ako úloha lineárneho programovania –  $L^1, L^\infty$

# 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  $L^\infty$  lineárnej regresie
- ▶ minimalizácia váženej sumy noriem

# Formulácia úloh lineárneho programovania

## Úloha

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

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (1)$$

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

## Minimalizovanie $L^1$ normy

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

označíme:

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

Problém prevedieme do tvaru:

$$\min c^T x$$

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

## Minimalizovanie $L^\infty$ normy

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

Minimalizovanie  $L^1$  normy ako úloha lineárneho programovania:

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

## Minimalizovanie $L^\infty$ 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:

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

# B

...



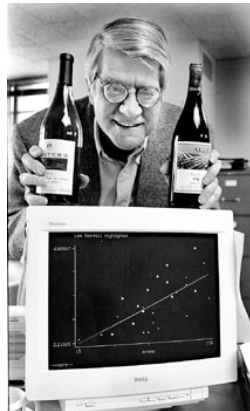
# Predikcia kvality vína - dáta

## 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

$L^\infty$

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

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

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

# D

...

# Všeobecná trieda pre $L^1$ a $L^\infty$ lineárnu regresiu

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

- ▶ zovšeobecnenie problému
- ▶ voľnosť dimenzionality
- ▶ vstupný vektor  $y$  a matica  $X$
- ▶ hodnoty  $\beta$  výstupom

```
# inicializacia
model1 = L1Model(y, X)
model2 = LInfModel(y, X)

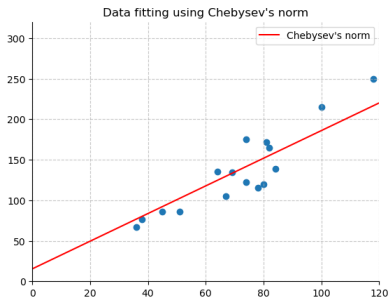
# riesenie
beta1 = model1.solve()
beta2 = model2.solve()
```

# Všeobecná trieda pre $L^1$ a $L^\infty$ lineárnu regresiu

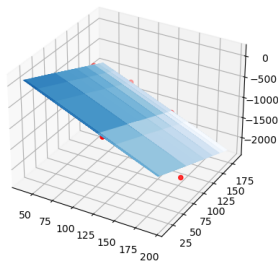
- ▶ hodnota  $R^2$
- ▶ vizualizácia pre 2D a 3D

```
model.r2()
```

```
model.visualize()
```

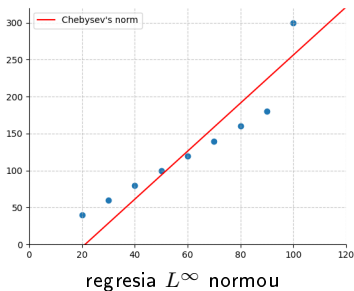
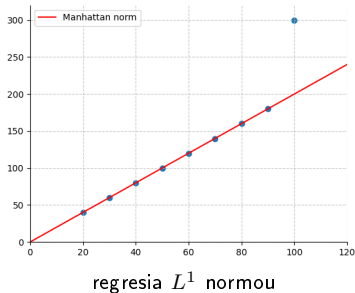


Data fitting using Manhattan norm



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

- ▶  $L^1$  – veľmi dobre zachytáva lineárny vzťah, môže viesť k *overfittingu*
- ▶  $L^\infty$  – príliš ovplyňovaná outliermi



# Minimalizácia váženého súčtu noriem

- ▶ redukcia *overfittingu*  $L^1$  regresie váženým súčtom s  $L^\infty$  normou

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

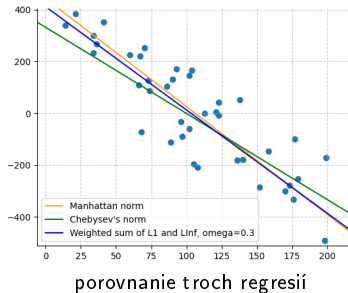
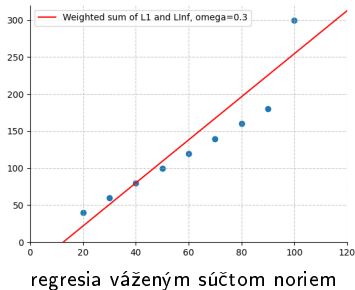
- ▶ stále implementovateľné ako úloha lineárneho programovania

$$\min \left( \mathbf{0}_{k+1}^T \mid \omega \mathbf{1}_n^T \mid (1 - \omega) \right) \begin{pmatrix} \beta \\ t \\ \gamma \end{pmatrix}, \quad \omega \in [0; 1]$$

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

# Minimalizácia váženého súčtu noriem

► implementované ako `WeightedL1LInfModel`





# Zhrnutie

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

## Ďalšie kroky

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