Master studije
GEOINFORMATIKA

Mašinsko učenje - metoda najstrmijeg spusta, generalizacija i regularizacija

Tatjana Jakšić Krüger

tatjana@turing.mi.sanu.ac.rs

16. novembar 2021.

Građevinski fakultet Univerziteta u Beogradu

Podsetnik

- Linearni model i estimatori.
- Metoda maksimalne verodostojnosti.
- Uvod u metodu najstrmijeg spusta.

Cilj za danas

- Nastavak za metodu najstrmijeg spusta.
- Izbor modela i generalizacija.
- Regularizacija.

Osnovna terminologija

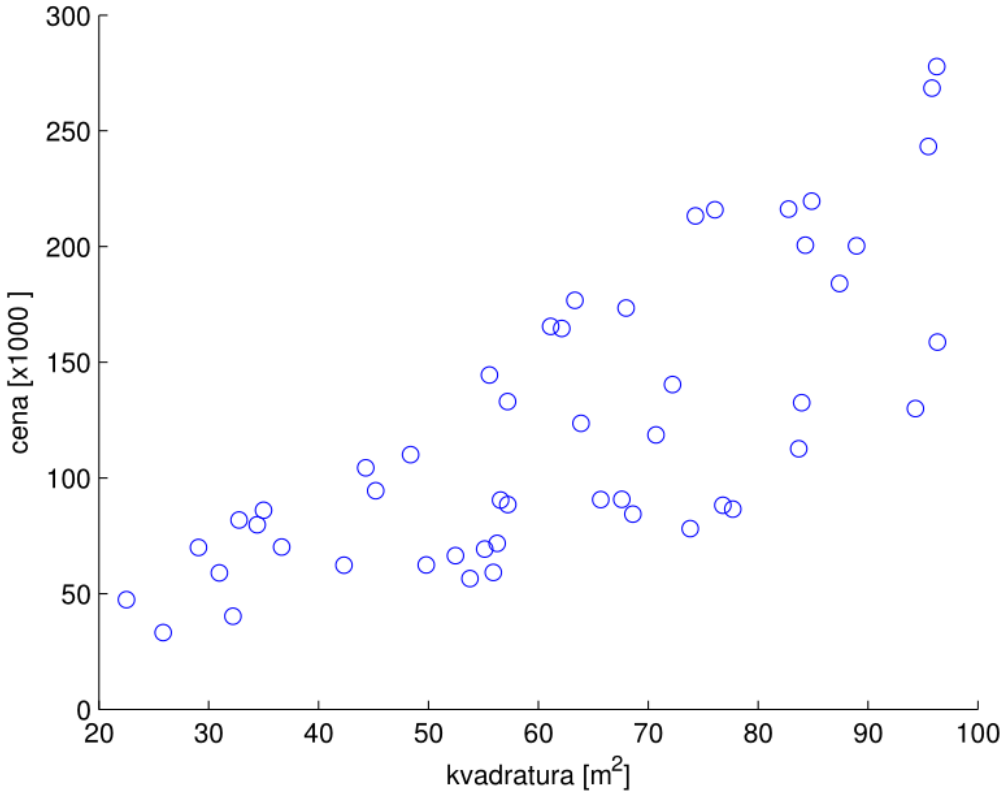
- Notacija:

x	ulaz, svojstvo (*eng* feature, atribut).
y	izlaz, ciljna promenljiva.
n	broj atributa/svojstava.
m	broj primera/instanci.
$(x^{(i)}, y^{(i)})$	i-ti primer iz skupa za obučavanje.
$\{x^{(i)}, y^{(i)}\}_{i-1}^{m}$	skup za obučavanje (trening skup).
$h(x) = y$	hipoteza (model).
$\epsilon_i \vert e_i$	pravi|empirijski reziduali.

- Ulaz može biti:
  - vektor $\vec{x} = \boldsymbol{x} = [x_1\ x_2\ \ldots\ x_n]^T$.
  - matrica dimenzije $m \times n$.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

Problem predviđanja - model regresije



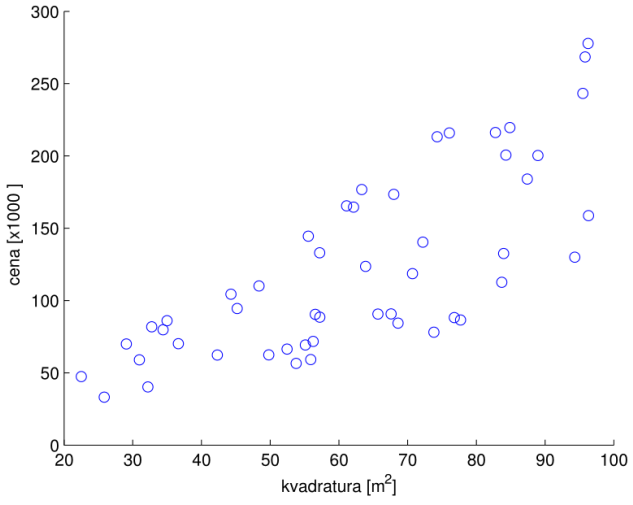
Zadatak: predvidi cenu na osnovu date kvadrature.

Izvor grafika:
http://automatika.etf.bg.ac.rs/images/FAJLOVI_srpski/predmeti/master_studije/MU/01%20Linearna%20regresija.pdf

Hipoteza u $\mathbb{R}^2$

- Ulazna promenljiva x je skalarna veličina.
- Ciljna promenljiva y je skalarna veličina.
- Parametri linearnog modela su θ_0 (odsečak na y-osi) i θ_1 (koeficijent pravca).
- Funkcija hipoteze (model) našeg predviđanja je oblika:

$$h(x) = \theta_0 + \theta_1 x.$$



Zadatak: predvidi cenu na osnovu date kvadrature.

Hipoteza u $\mathbb{R}^3$

- Ulazne promenljive x_1 i x_2 definišu vektor:

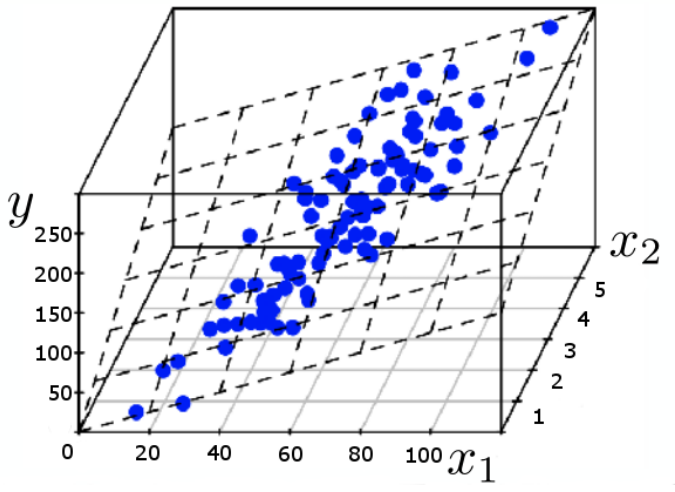
$$\boldsymbol{x} = [x_0\ x_1\ x_2]^T, \quad x_0 = 1$$

- Ciljna promenljiva y je skalarna veličina.
- Parametri linearnog modela su $\theta_0, \theta_1, \theta_2$.
- Hipoteza (model) našeg predviđanja je:

$$h(x; \boldsymbol{\theta}, \theta_0) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$
$$= \sum_{i=0}^{3} \theta_i x_i,$$

pri čemu parametri modela čine vektor

$$\boldsymbol{\theta} = [\theta_0\ \theta_1\ \theta_2]^T.$$



Zadatak: predvidi cenu na osnovu kvadrature (x_1) i broja spavaćih soba (x_2).

Postavka problema u $\mathbb{R}^{n+1}$

- Vektor svojstava je:

$$\boldsymbol{x} = [x_0\ x_1\ x_2\ \ldots\ x_n]^T, \quad x_0 = 1.$$

- Vektor parametara je:

$$\boldsymbol{\theta} = [\theta_0\ \theta_1\ \theta_2\ \ldots\ \theta_n]^T.$$

- Usvajamo linearnu hipotezu:

$$\begin{aligned} h(\boldsymbol{x}; \boldsymbol{\theta}, \theta_0) = h_\theta(\boldsymbol{x}) &= \theta_0 + \theta_1 x_1 + \ \ldots\ + \theta_n x_n. \\ &= \boldsymbol{\theta}^T \boldsymbol{x}. \end{aligned}$$

- Kako biramo $\boldsymbol{\theta}$?

$$h_\theta(\boldsymbol{x}) \approx y,$$

tj. takvo da je rizik najmanji:

$$\min_\theta \frac{1}{2} \sum_{i=1}^{m} (h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)})^2 = \min_\theta J(\theta).$$

Algoritmi optimizacije

1. Metoda najstrmijeg spusta (*eng.* gradient descent).
2. "Šaržni" gradijentni spust (*eng.* batch gradient descent).
3. Stohastički gradijentni spust.
4. Lokalno ponderisana linearna regresija.
5. Njutnova metoda.

Parcijalni izvod u $\mathbb{R}^2$

- Neka je $f = f(x, y)$.
- Parcijalni izvod po x u nekoj tački (x_0, y_0):

$$\frac{\partial f(x_0, y_0)}{\partial x} = \lim_{l \to 0} \frac{f(x_0 + l, y_0) - f(x_0, y_0)}{l} = \partial_x f(x_0, y_0)$$

- Parcijalni izvod po y u nekoj tački (x_0, y_0):

$$\frac{\partial f(x_0, y_0)}{\partial y} = \lim_{l \to 0} \frac{f(x_0, y_0 + l) - f(x_0, y_0)}{l} = \partial_y f(x_0, y_0)$$

- Ukoliko $\partial_x f(x_0, y_0)$ postoji, funkcija f je diferencijabilna po x u tački (x_0, y_0).

Parijalni izvod i gradijent u $\mathbb{R}^d$

- Neka je $f : \mathbb{R}^d \to \mathbb{R}$.
- Neka je f definisana i diferencijabilna u okolini U_a tačke $\vec{a} = (a_1, a_2, \ldots, a_d)$.
- *Gradijent* je vektor ∇f čije su koordinate:

$$\nabla f(a_1, \ldots, a_d) = (\frac{\partial f}{\partial x_1}(a_1, \ldots, a_d), \ldots, \frac{\partial f}{\partial x_n}(a_1, \ldots, a_d))$$

$$\nabla f(\boldsymbol{a}) = (\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{a})).$$

- Gradijent u tački $\boldsymbol{a} \in \mathbb{R}^d$ predstavlja vektor u čijem smeru funkcija f najbrže raste/opada u okolini tačke $\boldsymbol{a}$.
- Gradijent se koristi za odredjivanje maksimalne/minimalne vrednosti f u okolini tačke $\boldsymbol{a}$.

Metoda najstrmijeg spusta

Input: f, $\boldsymbol{a_0}$: analitički izraz sa f i polazna tačka
Output: $\nabla f(\boldsymbol{a_k})$

```
repeat
    Izračunaj ∇f(a_k);
    Pomeri se ka tački a_{k+1} u smeru gradijenta −∇f(a_k);
    k = k + 1 ;
until |f(a_k − f'(a_{k−1}))| ≤ ϵ |f(a_{k−1})|;
```

- Nedostaje nam korak koji kaže za koliko moramo da se pomeramo.
- Parametar λ pomaže da naučimo dužinu pomeraja iz $\boldsymbol{a_k}$ u $\boldsymbol{a_{k+1}}$:

$$\boldsymbol{a_{k+1}} = \boldsymbol{a_k} - \lambda \nabla f(\boldsymbol{a_k}).$$

- Izbor vrednosti za λ može biti određen:
  - teorijski: $\sum_{i=0}^{\infty} \lambda_i = \infty$ i $\sum_{i=0}^{\infty} \lambda^2 < \infty$.
  - empirijski: $\lambda = 0.01$.

Gradijentna metoda za odredjivanje θ_k

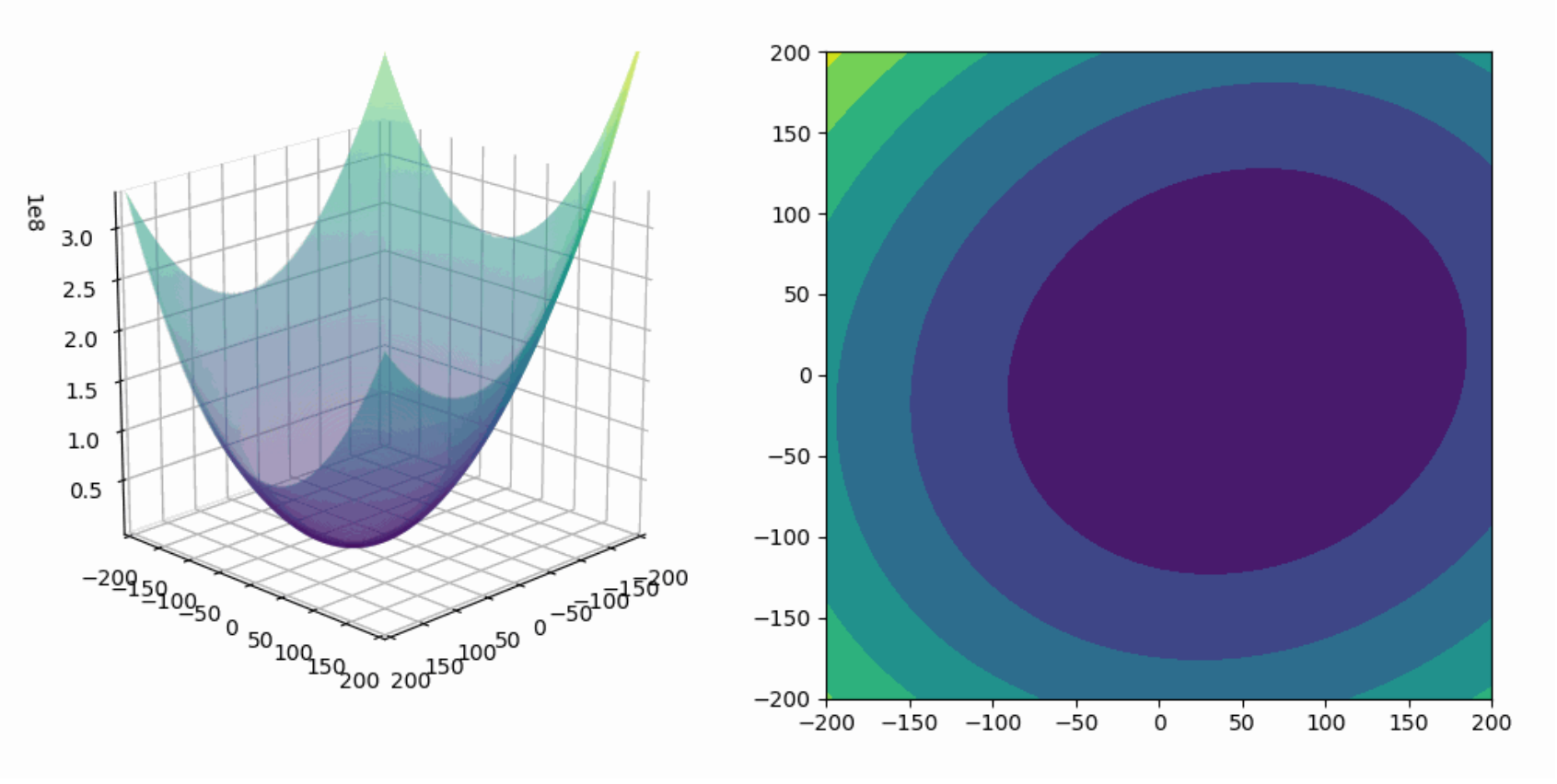
- Neka je θ_k k-ta komponenta vektora $\boldsymbol{\theta}$ vektor u $\mathbb{R}^{n+1}$ prostoru.
- Primenom gradijentnog spusta θ_k ažuriramo:

$$\theta_k := \theta_k - \lambda \frac{\partial J(\theta)}{\partial \theta_k}.$$

- Parcijalni izvod od $J(\boldsymbol{\theta})$ po θ_k:

$$\begin{aligned}
\frac{\partial J(\boldsymbol{\theta})}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k}\left(\frac{1}{2}\sum_{i=1}^{m}(h_\theta(\boldsymbol{x}^{(i)}) - y^{(i)})^2\right) \\
&= \frac{\partial}{\partial \theta_k}\left(\frac{1}{2}\sum_{i=1}^{m}\left(\sum_{j=0}^{n}\theta_j x_j^{(i)} - y^{(i)}\right)^2\right) \\
&= \frac{\partial}{\partial \theta_k}\left(\frac{1}{2}\sum_{i=1}^{m}(\theta_k x_k^{(i)} - y^{(i)})^2\right) \\
&= \sum_{i=1}^{m}(\theta_k x_k^{(i)} - y^{(i)})x_k^{(i)}.
\end{aligned} \tag{1}$$

Grafik funkcije $J(\theta)$



prikazani gif je moguće pokrenuti u Adobe Reader pdf editoru.

Minimizacija nad celim skupom za obučavanje za izračunavanje θ

- Imamo matricu svojstava X, dimenzije $m \times n$:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_{n^{(2)}} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

- Neka je $\boldsymbol{y}$ m-dimenzioni vektor koji sadrži ciljne vrednosti (labele) iz skupa za obučavanje:

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = [y^{(1)}\ y^{(2)} \dots y^{(m)}]^T$$

- Koristićemo sledeću osobinu:

$$h_\theta(\boldsymbol{x}^{(i)}) = \sum_{j=0}^{n} \theta_j x_j^{(i)} \iff h_\theta(\boldsymbol{x}^{(i)}) = (\boldsymbol{x}^{(i)})^T(\boldsymbol{\theta})$$

Minimizacija nad celim skupom za obučavanje

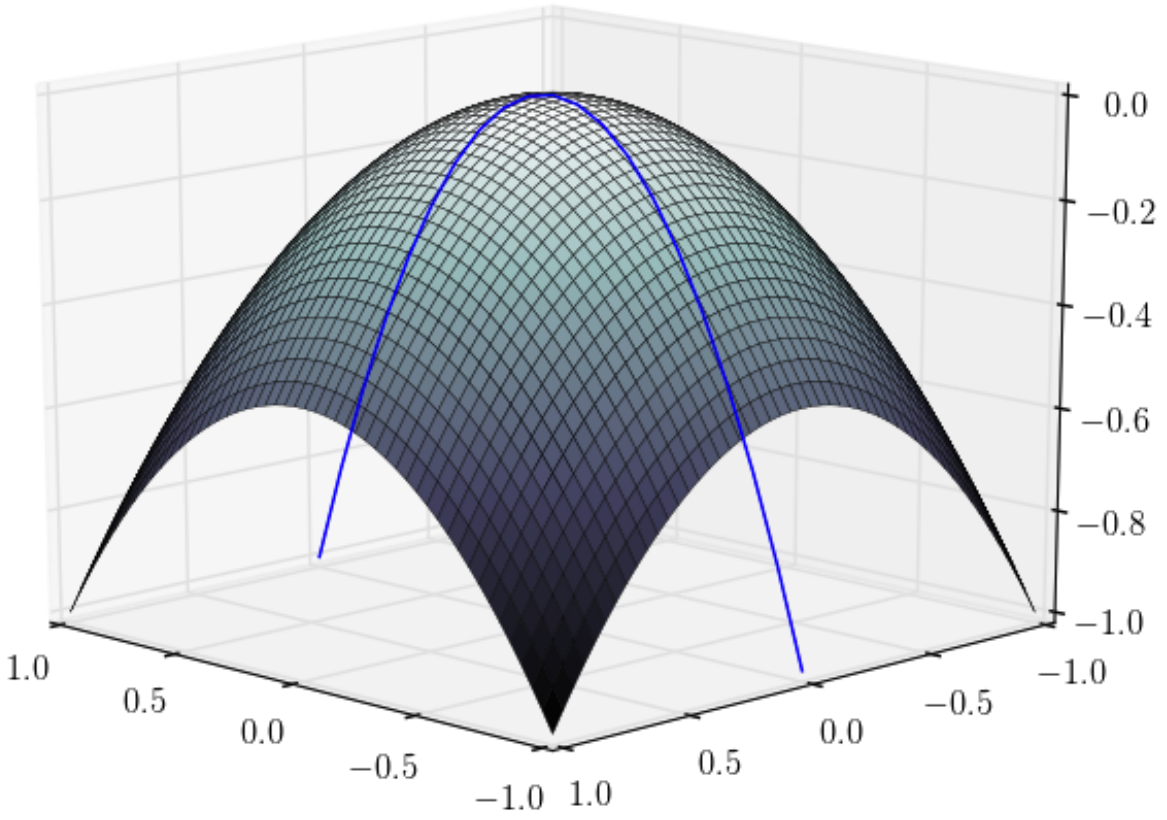
$$X\boldsymbol{\theta} - \boldsymbol{y} = \begin{bmatrix} (x^{(1)})^T\boldsymbol{\theta} \\ \vdots \\ (x^{(m)})^T\boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} (x^{(1)})^T\boldsymbol{\theta} - y^{(1)} \\ \vdots \\ (x^{(m)})^T\boldsymbol{\theta} - y^{(m)} \end{bmatrix}$$

- Na osnovu osobine $\boldsymbol{z}$ važi $\boldsymbol{z}^T\boldsymbol{z} = \sum_i z_i^2$ dobijamo.

$$\begin{aligned} \frac{1}{2}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) &= \frac{1}{2}\sum_{i=1}^{m}((\boldsymbol{x}^{(i)})^T\boldsymbol{\theta} - \boldsymbol{y}^{(i)})^2 \\ &= \sum_{i=1}^{m}((h(\boldsymbol{x}^{(i)}) - \boldsymbol{y}^{(i)})^2 \\ &= J(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

- Cilj je minimizacija ovako zapisane funkcije cilja $J(\boldsymbol{\theta})$.

Primeri kvadratne povrsi bez minimalne vrednosti



Konkavna površ

Globanlni minimum u jednom koraku

- Dimenzija gradijenta je iste dimenzije kao i vektor θ.
- Pomoćne formule:

$$(AB)^T = B^T A^T$$

$$trA = trA^T$$

$$\nabla_A trABA^T C = CAB + C^T AB^T$$

- dovode do sledećeg:

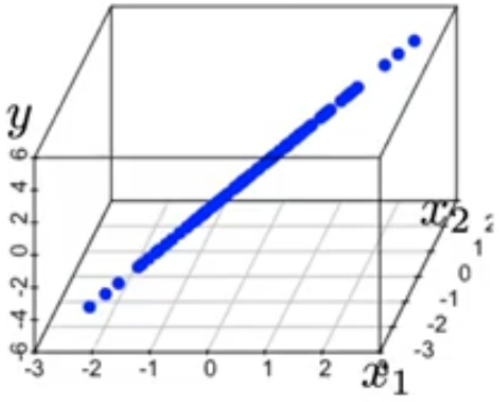
$$\nabla J(\boldsymbol{\theta}) = \nabla \frac{1}{2}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) = \ldots = X^T X\boldsymbol{\theta} - X^T\boldsymbol{y}. \qquad (3)$$

- Globalni optimum nalazimo kada je $\nabla J(\boldsymbol{\theta}) = 0$ tj:

$$\boldsymbol{\theta} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Kada formula za θ ne radi?

- Ponekada ne postoji jedinstvena hipperravan.



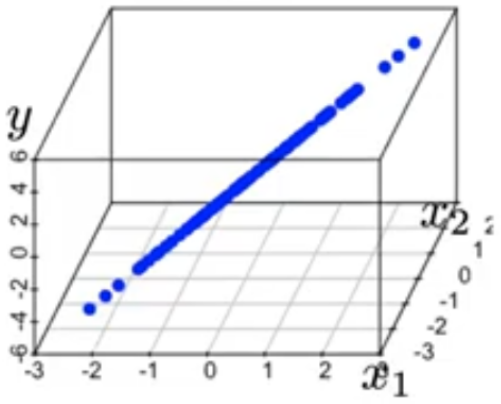
Podaci čine liniju.

- Ponekada usled šuma deluje da imamo dobro rešenje, ali ono nije značajno bolje od drugih.

- Visoko korelisani atributi.
- Preveliki broj atributa.

Kada formula za θ ne radi?

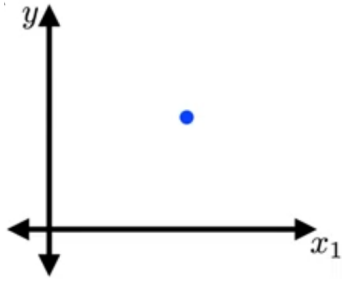
- Ponekada ne postoji jedinstvena hipperravan.



Podaci čine liniju.

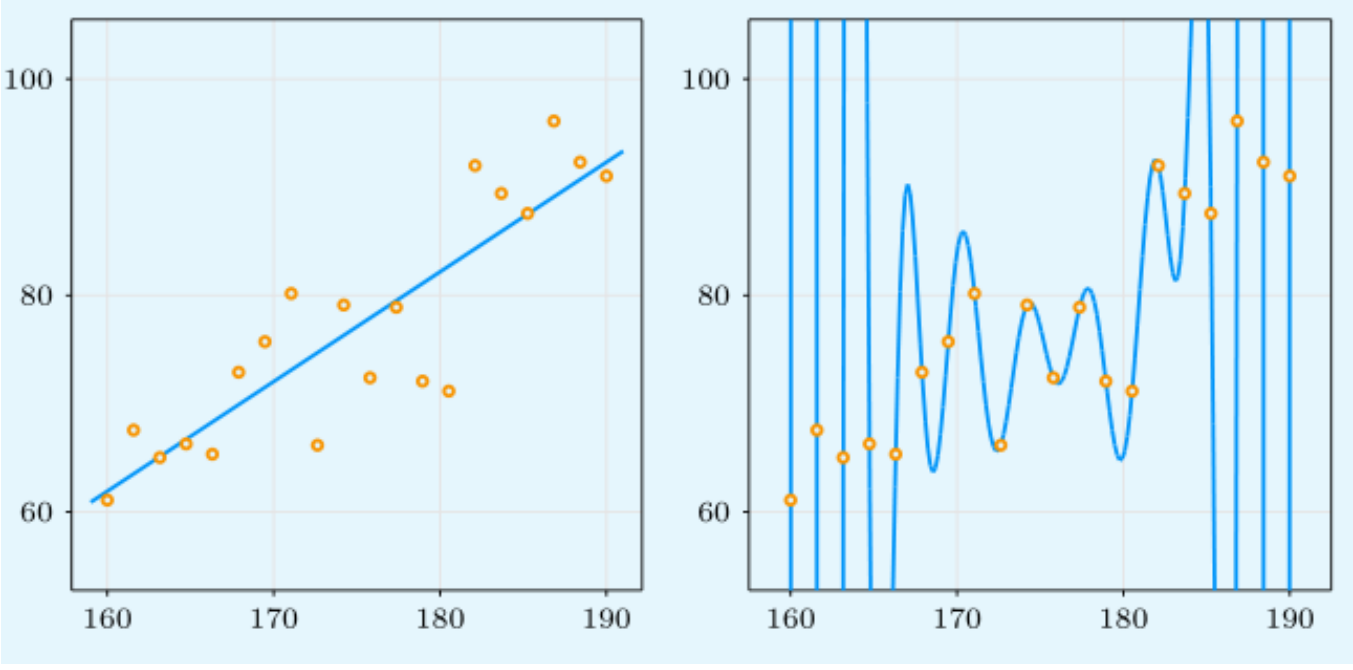
- Ponekada usled šuma deluje da imamo dobro rešenje, ali ono nije značajno bolje od drugih.

- Visoko korelisani atributi.
- Preveliki broj atributa.



Kada minimizacija ne radi? Preprilagođavanje. Generalizacija.

- Slika levo polinom prvog stepena. Slika desno je polinom višeg stepena. Deluje da druga slika bolje fituje jer je greška nula.
- Preprilagođavanje (*eng*. overfitting) je fenomen koji se javlja kada sa složenošću modela raste nepouzdanost.



- *Generalizacija* je mera koliko dobro će istrenirani model predvideti nove rezultate.

Složenost modela

- Složenost odgovara na pitanje koliko je dobro imati veliku ili premalu složenost.
- Moguće je da je složenost manja od instinske funkcije - *eng*. underfitting.

Izvor https://www.barelysignificant.com/slides/vendredi_quanti_2021/vendredi_quantis#32

Regularizacija.

- Analiza atributa, potraga za najuticajnijim svojstvima.
- Regularizacija - smanjivanje složenosti modela.
- Poznati načini regularizacije kod linearnih modela je penalizovanje:
  - Lasso regression.
  - Ridge regression.
- Ridž regresija penalizuje na taj način da neki parametri dobijaju vrednosti bliske nuli.
- Koristi se l_2 norma (euklidsko rastojanje)

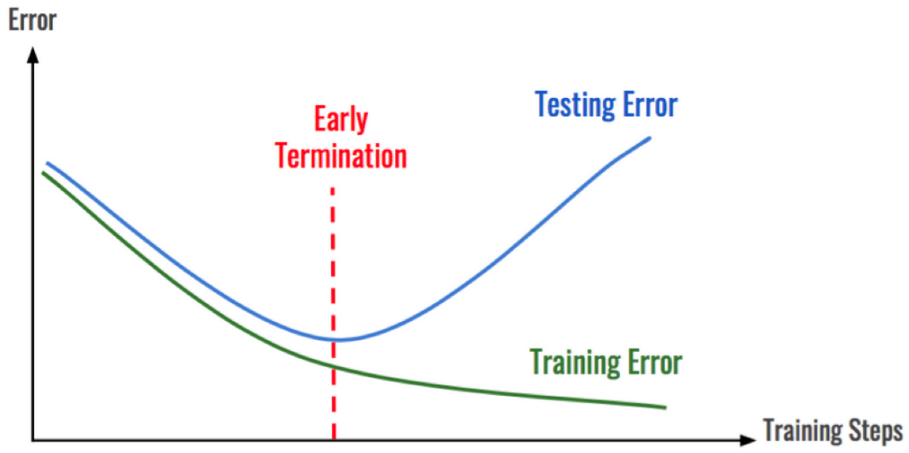
$$J_{ridge}(\theta) = \frac{1}{2}\sum_{i=1}^{m}((x^{(i)})^T - y^{(i)})^2 + \lambda||\theta||^2, \quad \lambda > 0.$$

- Ukoliko želimo da neke komponente vektora budu nula, koristimo Laso regresiju koja se zasniva na l_1-meri.

$$J_{lasso}(\theta) = \frac{1}{2}\sum_{i=1}^{m}((x^{(i)})^T - y^{(i)})^2 + \lambda\sum_{i=1}^{p}|\theta_i|.$$

"Early stopping" regularizacija

- Regularizacija u vidu ranog zaustavljanja (*eng.* early stopping).
- Obučavanje se zaustavlja pre nego što se javi problem sa preprilagođavanjem.



- Neophodno je balansiranje između:
  1. Složenosti modela (hipoteze).
  2. Veličine skupa za obučavanje.
  3. Greške generalizacije nad skupom za testiranje.

Korisni linkovi

- Predavanja prof. Predraga Tadića, `http://automatika.etf.bg.ac.rs/sr/13m051mu`.

U nastavku...

- Podela skupa na 3 dela (obučavanje, validacija, testiranej).
- Cross-validation.

Hvala na pažnji.

Sada idemo na prikaz komandi.

Da li imate pitanja?