Linearna i logistička regresija

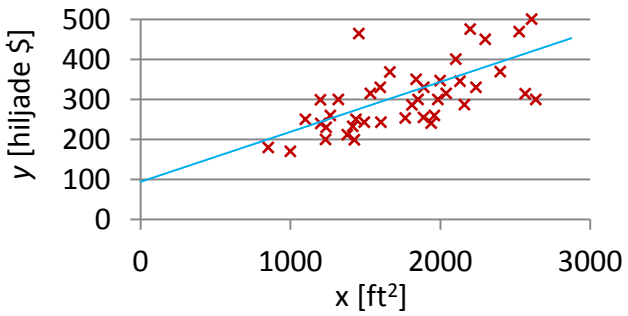
- Linearna regresija
  - Regresija za slučaj jednog obeležja
  - Metod gradijentnog silaska
  - Regresija za slučaj više obeležja
  - Analitička minimizacija funkcije cene
- Logistička regresija

Regresija

- *Regresija* je tehnika za predviđanje vrednosti kontinualne izlazne promenljive koja zavisi od određenih obeležja uzorka
  - Promenljiva ne mora biti stvarno kontinualna, dovoljno je da bude pogodno posmatrati je kao kontinualnu
- Kod linearne regresije za slučaj jednog obeležja pretpostavljamo da je veza između obeležja *x* i izlazne promenljive *y* linearna, odnosno, predviđamo vrednost izlazne promenljive *y* na osnovu *hipoteze* da je ta veza linearna:

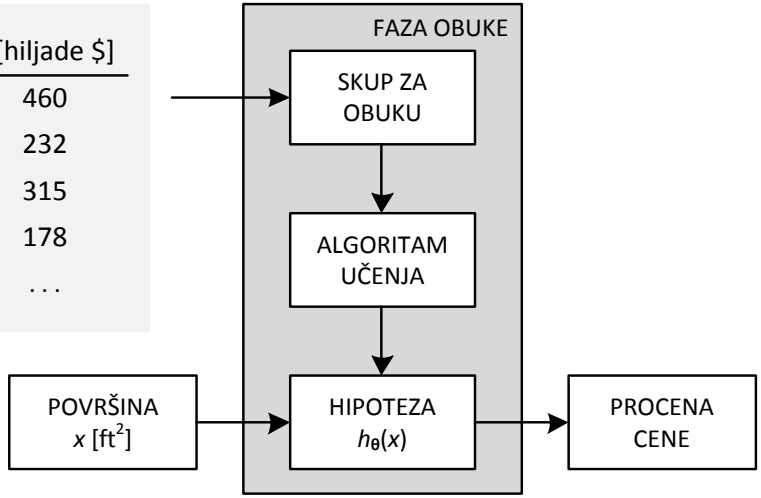
$$h_\theta(x) = \vartheta_0 + \vartheta_1 x$$

  - Pitanje je kako naći optimalne vrednosti parametara ϑ_0 i ϑ_1



Primer: predviđanje cene stana na osnovu kvadrature

x [ft^2]	y [hiljade $]
2104	460
1416	232
1534	315
852	178
...	...



Linearna regresija za slučaj jednog obeležja

- Potrebno je postaviti pravu liniju koja u najmanjoj meri odstupa od uzoraka iz skupa za obuku
- Kao mera odstupanja može se usvojiti *srednja kvadratna greška* na svim uzorcima*
  - Ovu meru nazivamo *funkcija cene*

$$J(\vartheta_0,\vartheta_1)=\frac{1}{2N}\sum_{i=1}^{N}(h_{\boldsymbol{\theta}}(x^{(i)})-y^{(i)})^2$$

gde je: N – ukupan broj uzoraka

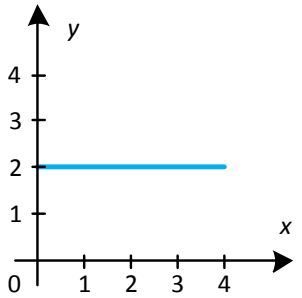
$x^{(i)}$ – vrednost x kod i-tog uzorka

$y^{(i)}$ – vrednost y kod i-tog uzorka

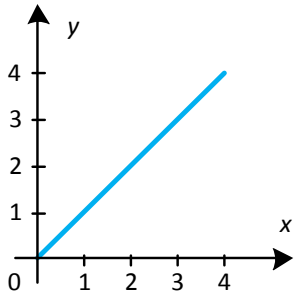
$h_{\boldsymbol{\theta}}(x^{(i)})$ – prognoza vrednosti obeležja y na osnovu hipoteze $h_{\boldsymbol{\theta}}$ za i-ti uzorak

  - Matematički, problem se svodi na minimizaciju funkcije cene:

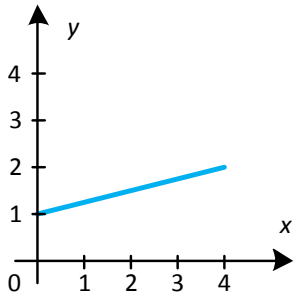
$$(\hat{\vartheta}_0,\hat{\vartheta}_1)=\underset{\vartheta_0,\vartheta_1}{\arg\min}\,J(\vartheta_0,\vartheta_1)$$



$\vartheta_0 = 2$
$\vartheta_1 = 0$
$y = 2$



$\vartheta_0 = 0$
$\vartheta_1 = 1$
$y = x$



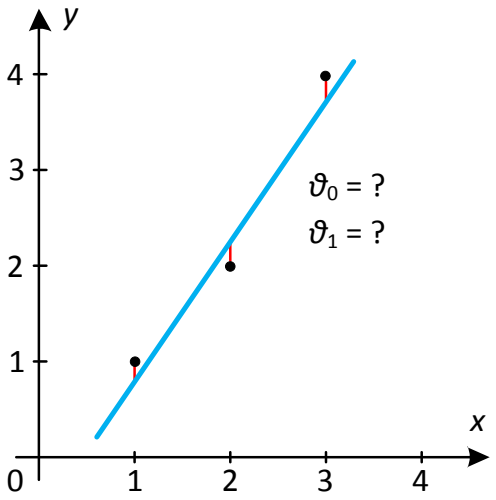
$\vartheta_0 = 1$
$\vartheta_1 = 1/4$
$y = 1 + x/4$

* Zapravo, polovina srednje kvadratne greške, zbog faktora $1/2N$ umesto $1/N$, ali pozitivan konstantan faktor svakako ne utiče na rezultat minimizacije

Linearna regresija (primer)

- Neka su data tri uzorka: {(1, 1), (2, 2), (3, 4)}. Odrediti pravu koja najmanje odstupa od njih u smislu srednje kvadratne greške.

$$x^{(1)} = 1, \quad y^{(1)} = 1$$
$$x^{(2)} = 2, \quad y^{(2)} = 2$$
$$x^{(3)} = 3, \quad y^{(3)} = 4$$

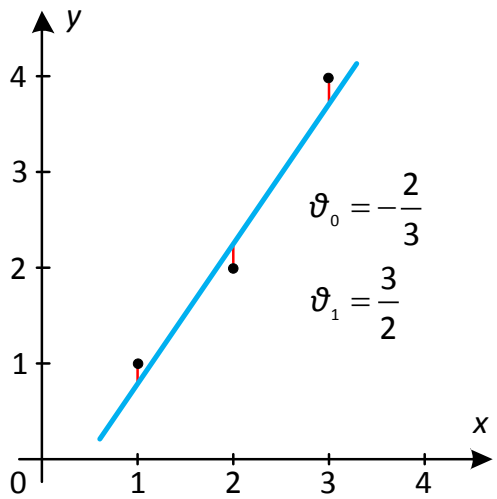


Linearna regresija (primer)

- Neka su data tri uzorka: {(1, 1), (2, 2), (3, 4)}. Odrediti pravu koja najmanje odstupa od njih u smislu srednje kvadratne greške.

$$x^{(1)}=1,\quad y^{(1)}=1 \qquad \hat{y}^{(1)}=\vartheta_0+\vartheta_1 x^{(1)}=\vartheta_0+\vartheta_1$$
$$x^{(2)}=2,\quad y^{(2)}=2 \qquad \hat{y}^{(2)}=\vartheta_0+\vartheta_1 x^{(2)}=\vartheta_0+2\vartheta_1$$
$$x^{(3)}=3,\quad y^{(3)}=4 \qquad \hat{y}^{(3)}=\vartheta_0+\vartheta_1 x^{(3)}=\vartheta_0+3\vartheta_1$$

$$J(\boldsymbol{\theta})=\frac{1}{2N}\sum_{i=1}^{N}(h_{\boldsymbol{\theta}}(x^{(i)})-y^{(i)})^2$$
$$=\frac{1}{6}\left((\vartheta_0+\vartheta_1-1)^2+(\vartheta_0+2\vartheta_1-2)^2+(\vartheta_0+3\vartheta_1-4)^2\right)$$

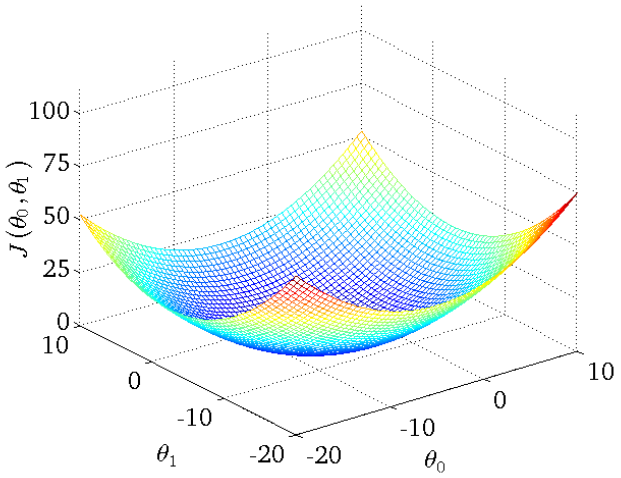


$$\partial J(\boldsymbol{\theta})/\partial\vartheta_0=6\vartheta_0+12\vartheta_1-14=0$$
$$\partial J(\boldsymbol{\theta})/\partial\vartheta_1=12\vartheta_0+28\vartheta_1-34=0$$
$$\Rightarrow\quad \vartheta_0=-\frac{2}{3},\ \vartheta_1=\frac{3}{2} \qquad h_{\boldsymbol{\theta}}(x)=-\frac{2}{3}+\frac{3}{2}x$$

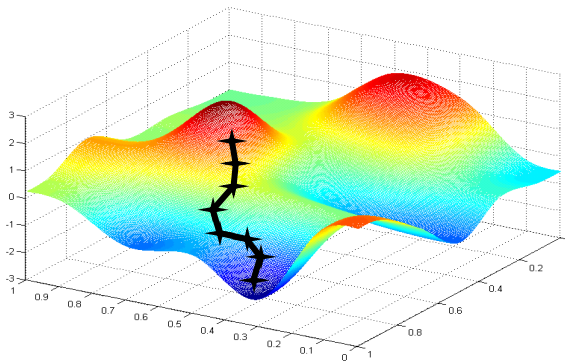
- Šta bi bilo drugačije da je postojalo npr. ograničenje da mora biti $h_{\boldsymbol{\theta}}(0)=0$?

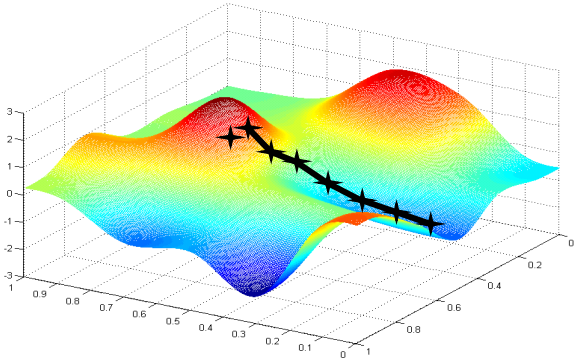
Dva načina rešavanja problema linearne regresije

- Funkcija cene je konveksna kvadratna funkcija parametara ϑ_0 i ϑ_1 sa jedinstvenim minimumom
  - Minimum $J(\vartheta_0, \vartheta_1)$ u ovom jednostavnom slučaju može se naći i **analitički**
  - Alternativa je iterativni **metod gradijentnog silaska** (eng. *gradient descent*)
    - Poći od proizvoljnih vrednosti ϑ_0 i ϑ_1
    - Menjati ϑ_0 i ϑ_1 u malim koracima u pravcu smanjenja $J(\vartheta_0, \vartheta_1)$ dok se ne dostigne minimum



  - Ovo je opšti metod za nalaženje minimuma funkcije više promenljivih
    - Kada površ $J(\vartheta_0, \vartheta_1, \ldots, \vartheta_d)$ ima složeniji oblik, dostizanje globalnog minimuma zavisi od izbora inicijalnih vrednosti $\vartheta_0, \vartheta_1, \ldots, \vartheta_d$





Metod gradijentnog silaska

- Početi od proizvoljnih vrednosti ϑ_0 i ϑ_1
- Istovremeno promeniti ϑ_0 i ϑ_1 prema pravilu

$$\vartheta_0 \leftarrow \vartheta_0 - \alpha \frac{\partial}{\partial \vartheta_0} J(\vartheta_0, \vartheta_1)$$

$$\vartheta_1 \leftarrow \vartheta_1 - \alpha \frac{\partial}{\partial \vartheta_1} J(\vartheta_0, \vartheta_1)$$

gde je α brzina učenja (fiksni mali broj)

- Ponavljati prethodni korak do konvergencije

- Izloženi algoritam u opštem slučaju konvergira ka lokalnom minimumu
  - Ako je brzina učenja premala, algoritam sporo konvergira, ali ako je prevelika, konvergencija može biti ugrožena
  - Kako se trenutna vrednost (ϑ_0, ϑ_1) približava lokalnom minimumu, tako koraci promene postaju sve manji, tako da nema potrebe da se brzina učenja α menja tokom izvršavanja algoritma

Metod gradijentnog silaska (opšti slučaj)

- Početi od proizvoljnih vrednosti $\vartheta_0, \vartheta_1, \ldots \vartheta_d$
- Istovremeno promeniti $\vartheta_0, \vartheta_1, \ldots \vartheta_d$ prema pravilu

$$\vartheta_0 \leftarrow \vartheta_0 - \alpha \frac{\partial}{\partial \vartheta_0} J(\vartheta_0, \vartheta_1, \ldots \vartheta_d)$$

$$\vartheta_1 \leftarrow \vartheta_1 - \alpha \frac{\partial}{\partial \vartheta_1} J(\vartheta_0, \vartheta_1, \ldots \vartheta_d)$$

$$\ldots$$

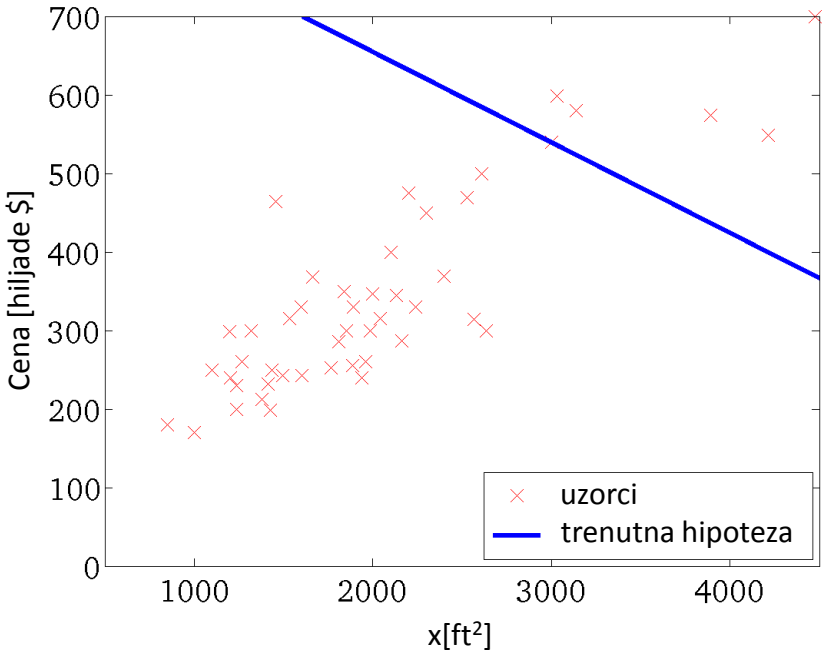
$$\vartheta_d \leftarrow \vartheta_d - \alpha \frac{\partial}{\partial \vartheta_d} J(\vartheta_0, \vartheta_1, \ldots \vartheta_d)$$

gde je α brzina učenja (fiksni mali broj)

- Ponavljati prethodni korak do konvergencije

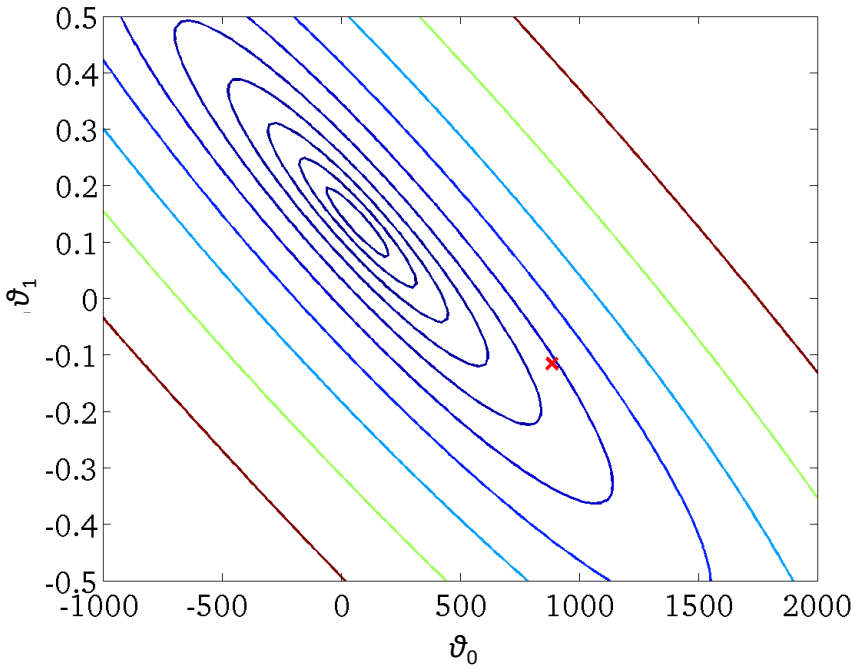
- Iako u opštem slučaju algoritam ne konvergira ka globalnom minimumu, u slučaju linearne regresije to se ipak dešava jer $J(\boldsymbol{\theta})$ predstavlja kvadratnu funkciju $\boldsymbol{\theta}$ tako da površ $J(\boldsymbol{\theta})$ ima jedinstveni minimum

$h_\theta(x)$



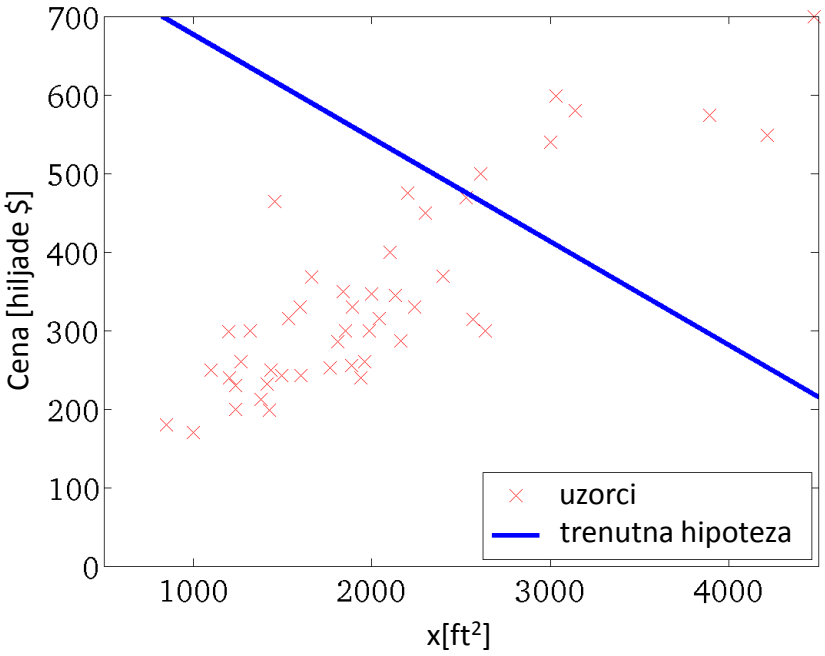
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



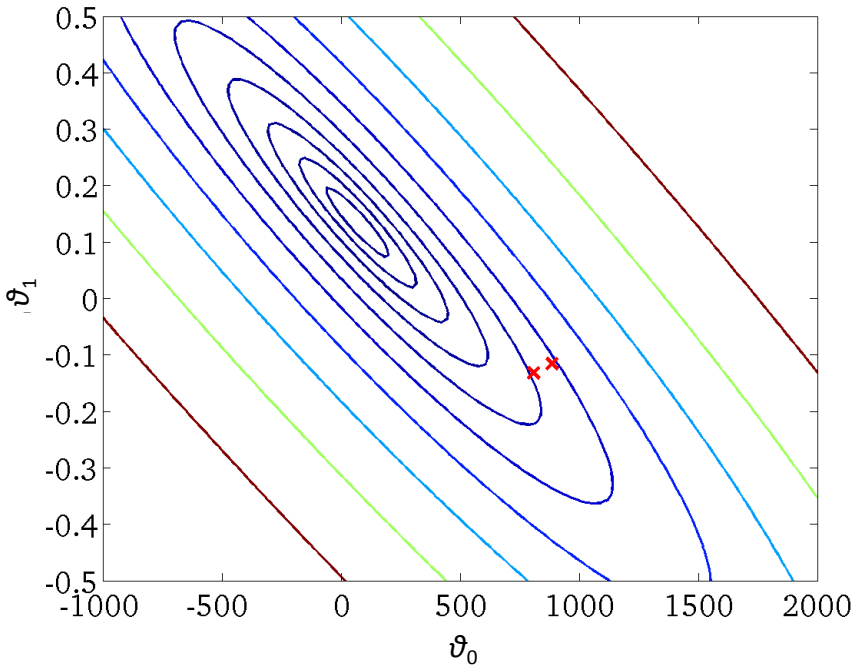
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



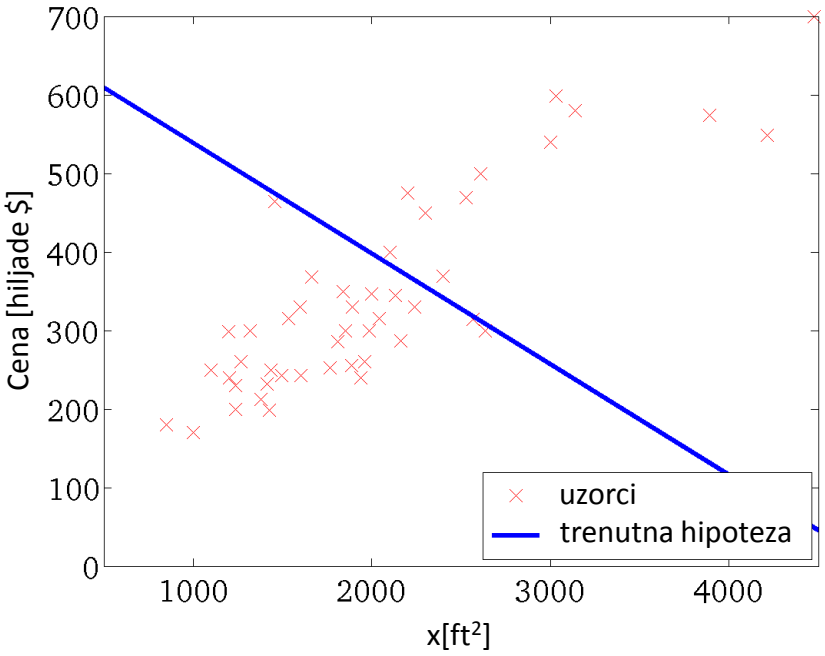
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



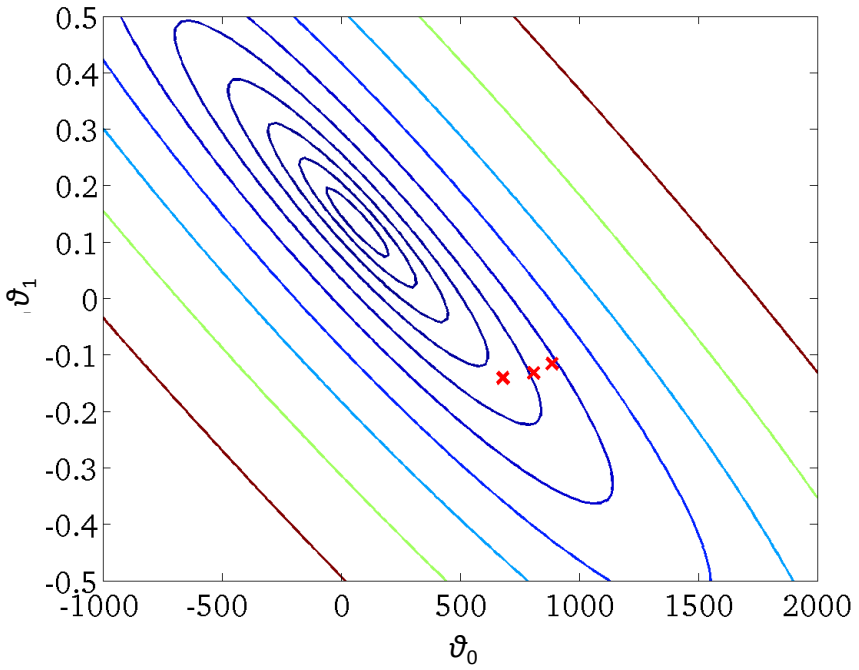
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_{\theta}(x)$



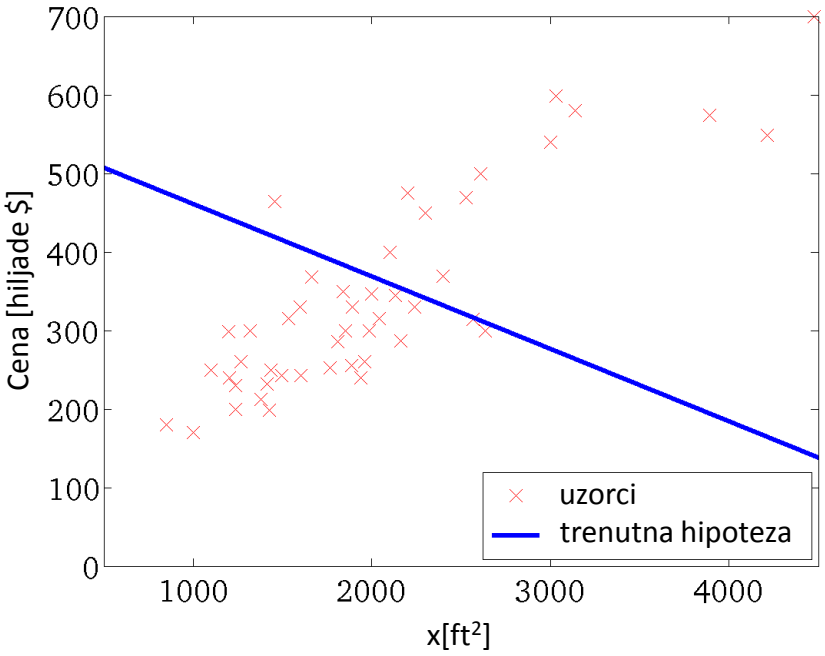
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_{\theta}(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



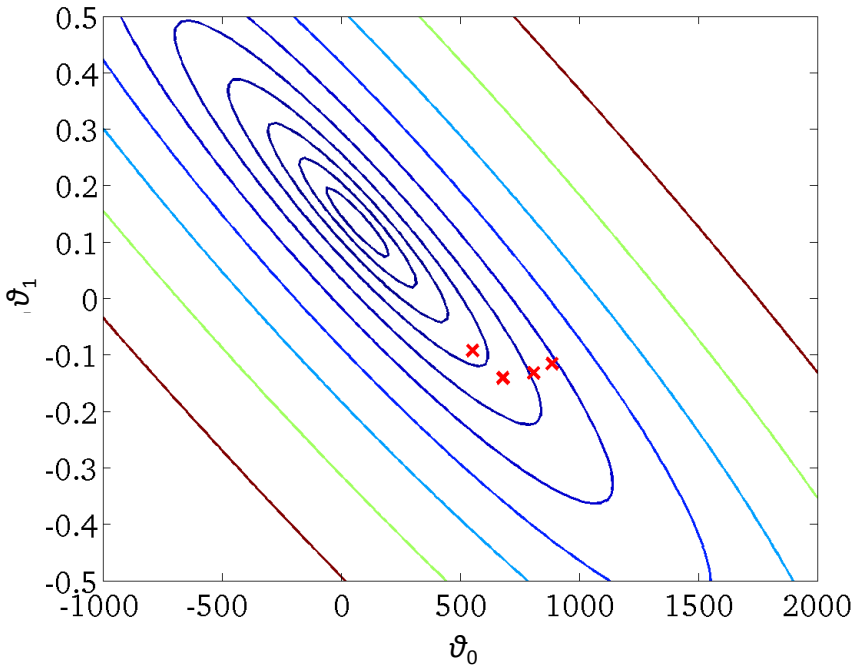
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



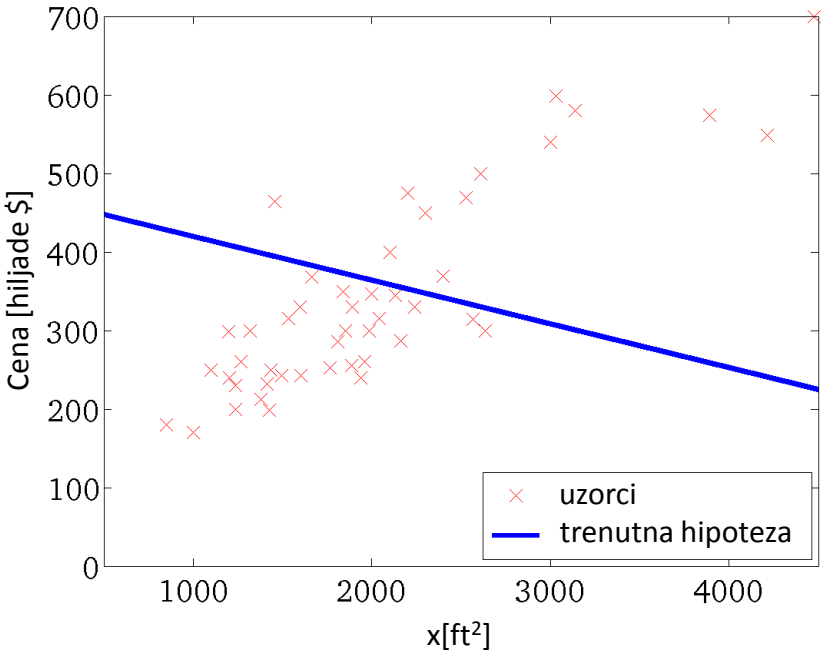
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



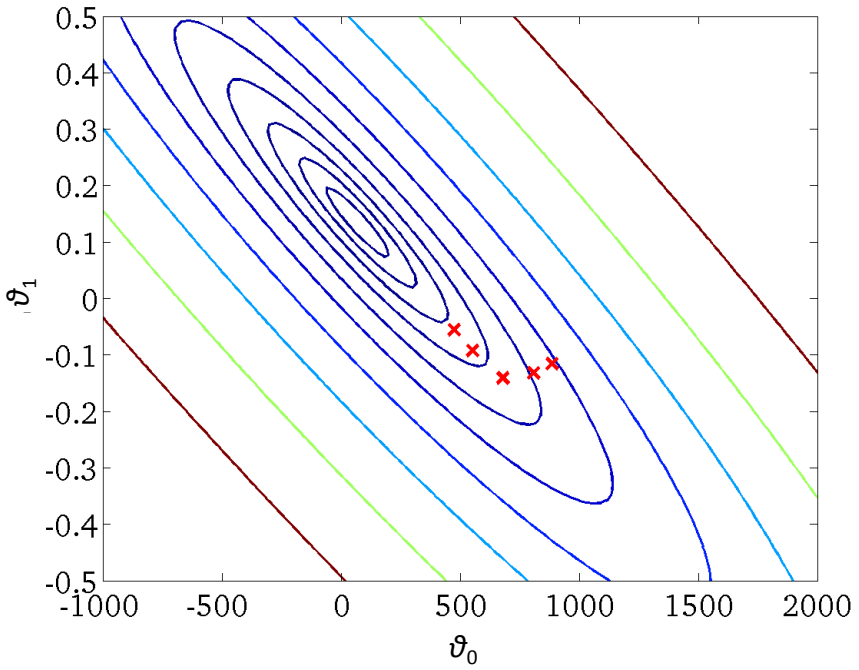
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



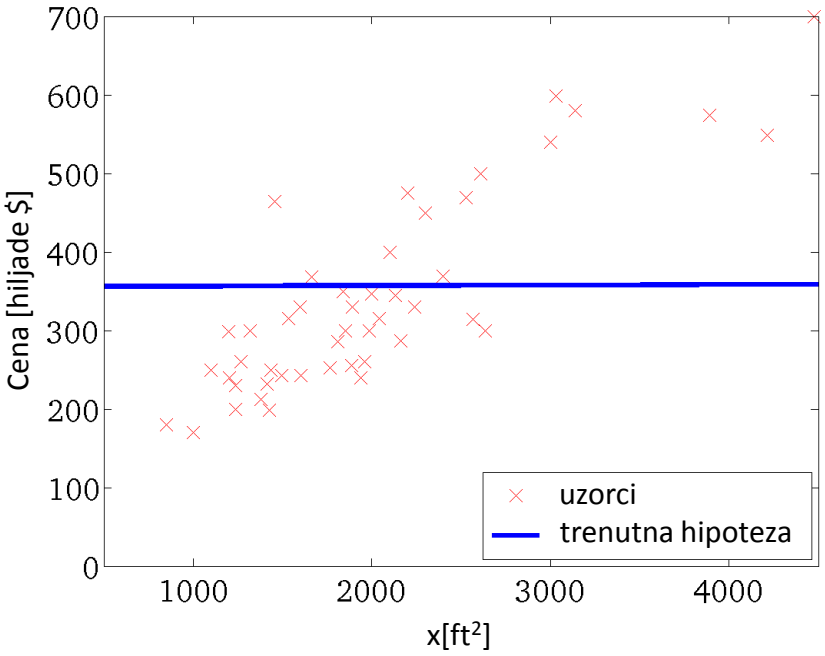
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



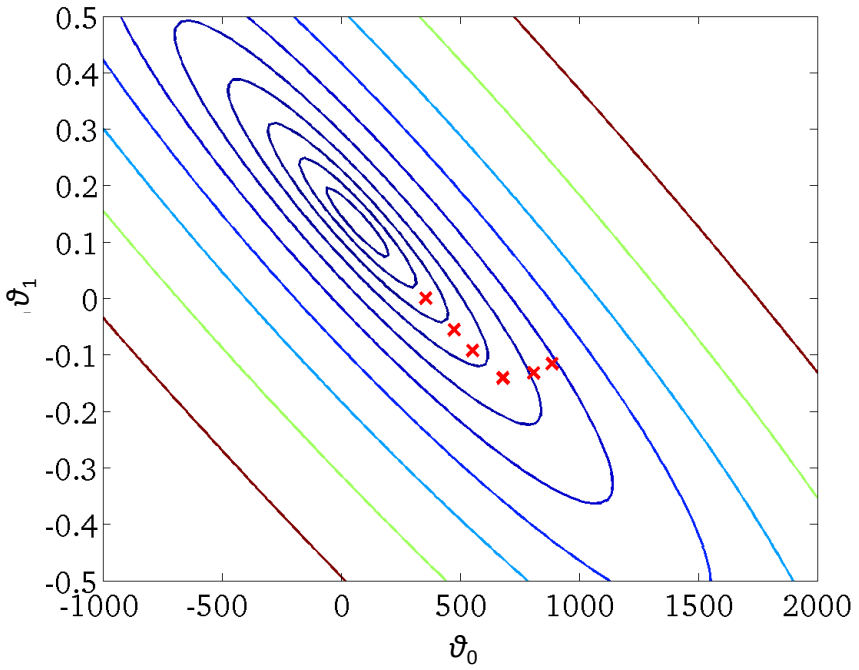
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



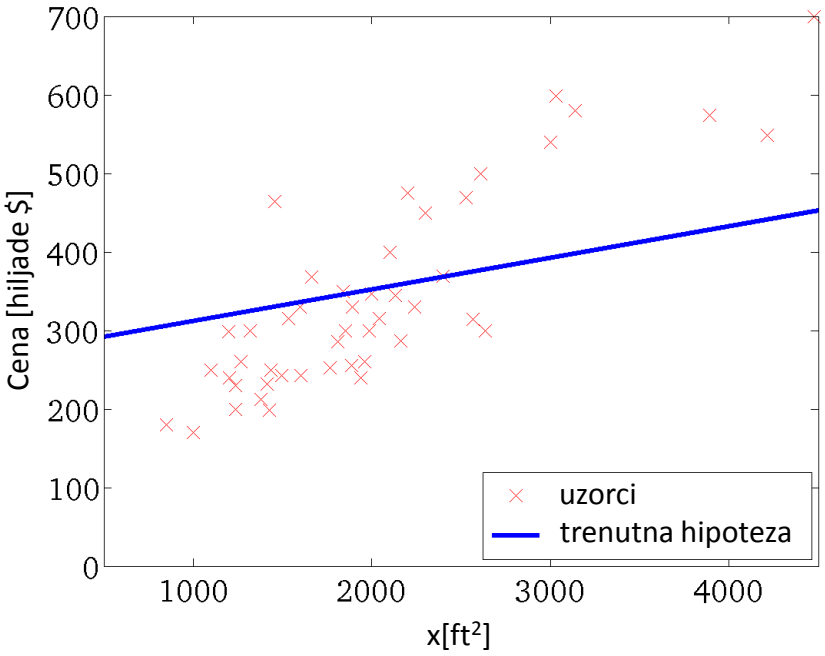
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



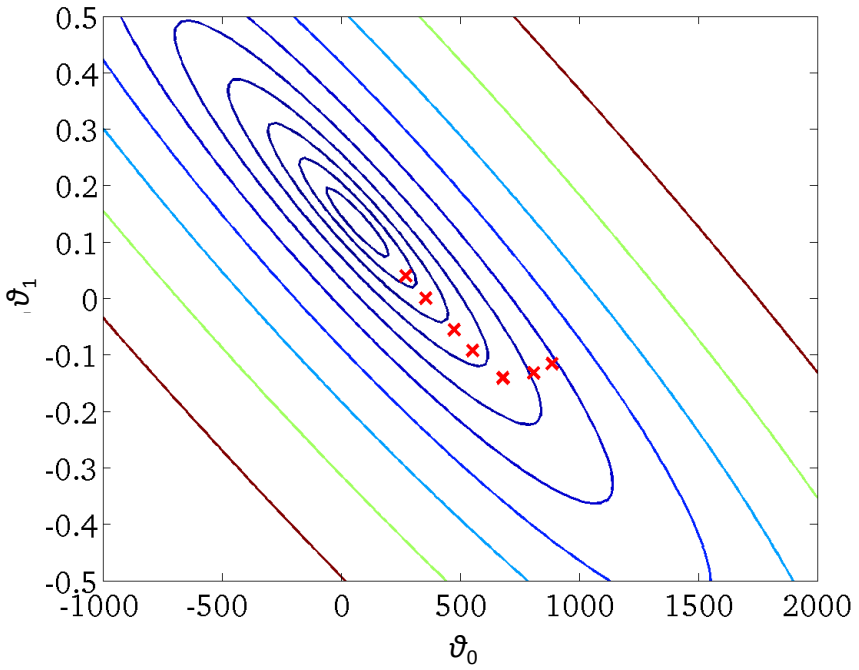
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



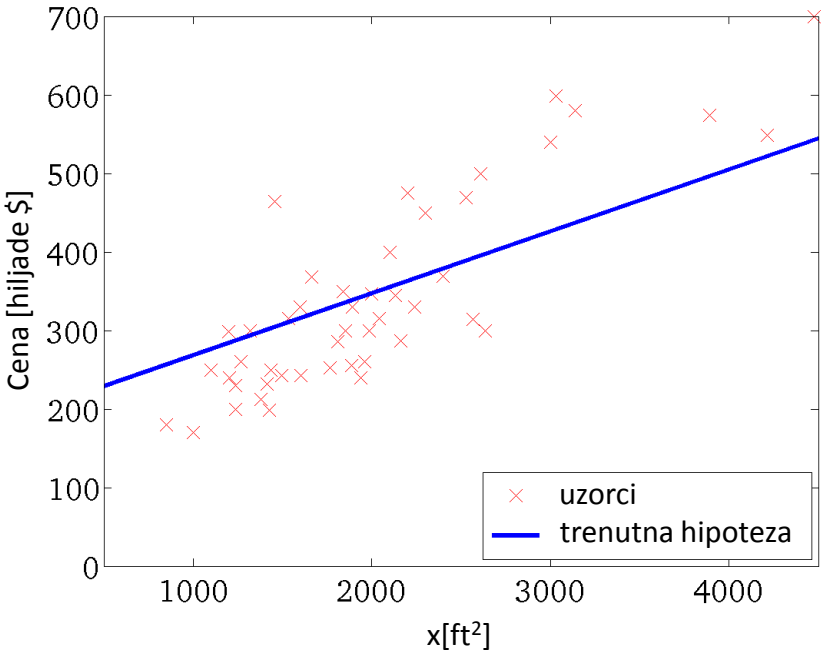
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



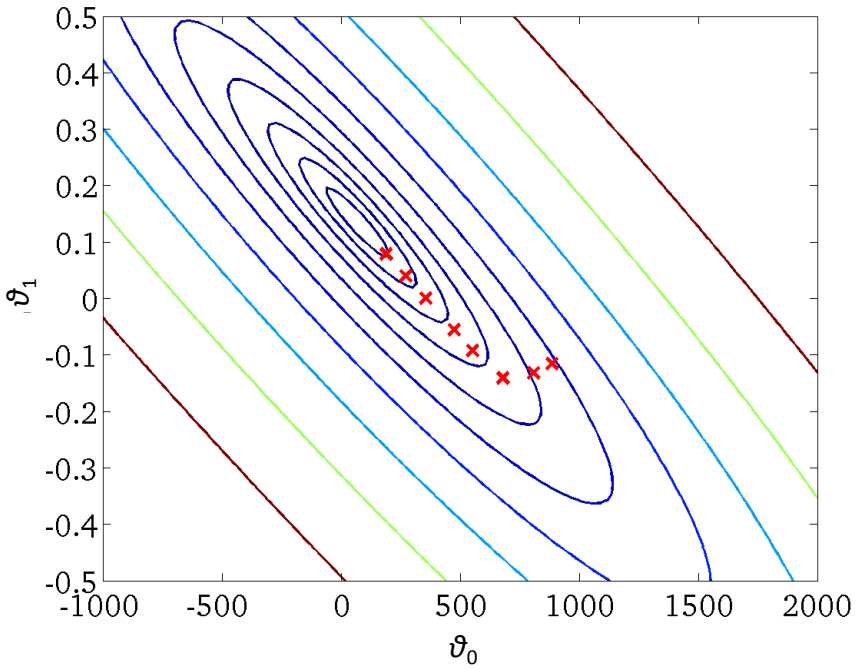
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



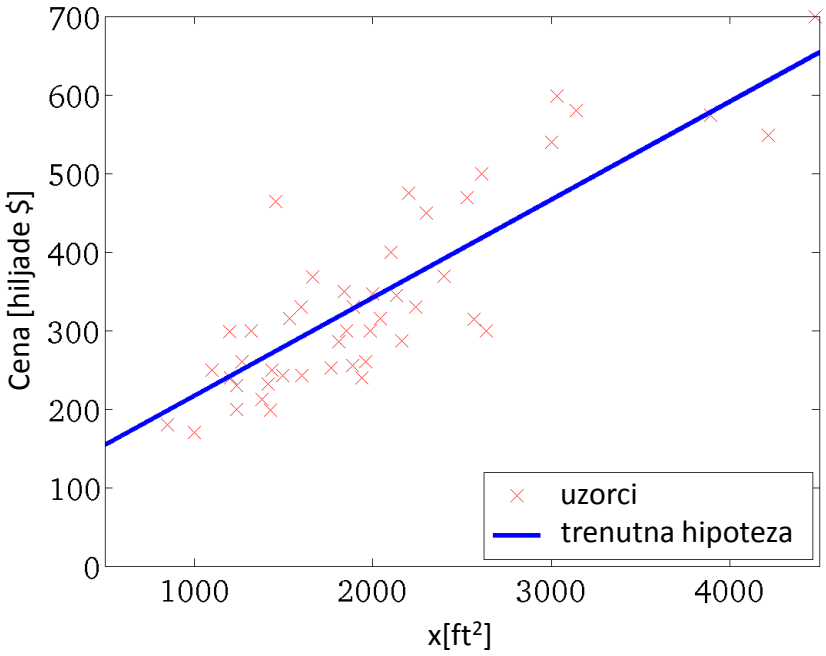
Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$



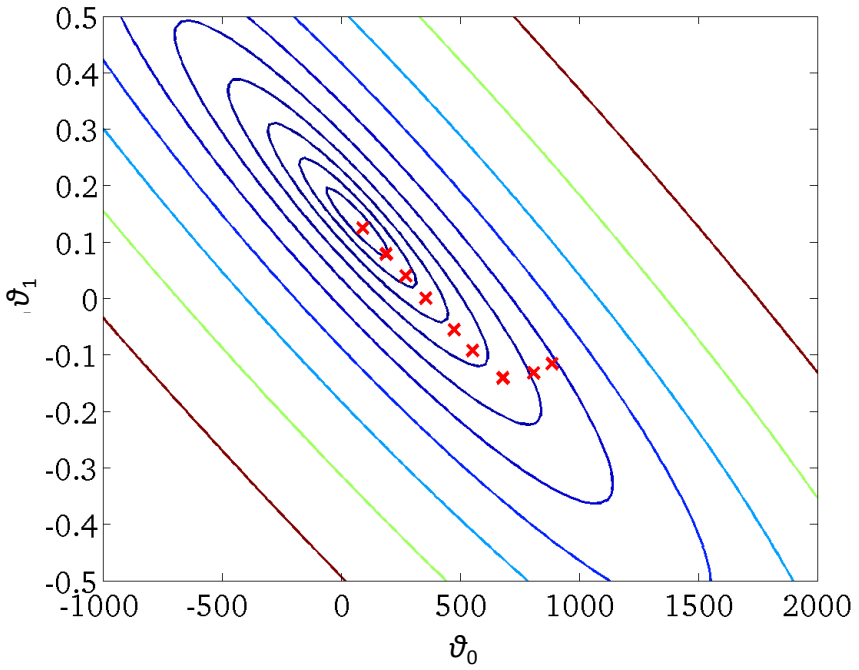
Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

$h_\theta(x)$



Za fiksne vrednosti ϑ_0 i ϑ_1
$h_\theta(x)$ je funkcija x

$J(\vartheta_0, \vartheta_1)$

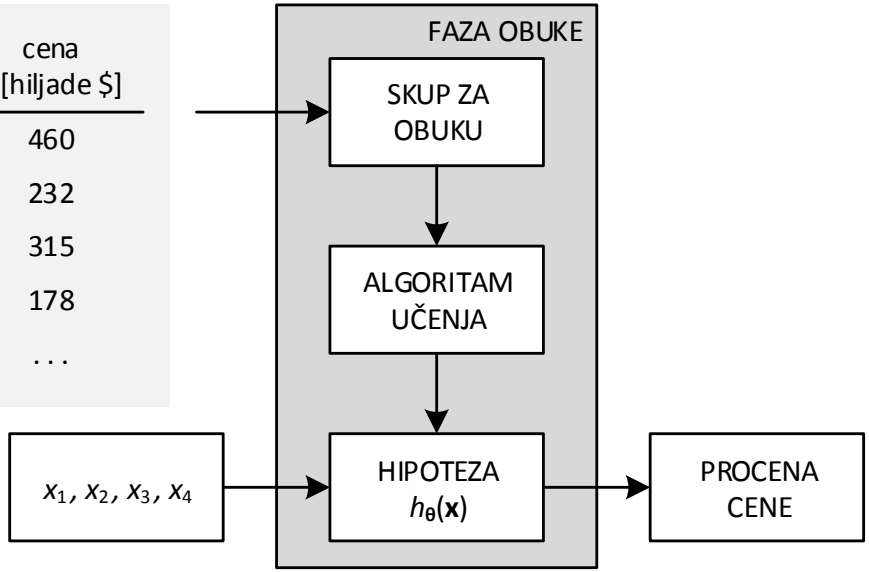


Funkcija cene je funkcija parametara ϑ_0 i ϑ_1

Linearna regresija za slučaj više obeležja

- Veći broj obeležja pruža više informacija i omogućava da se izlazna veličina tačnije predvidi
  - Ovaj slučaj se ne može vizuelizovati kao prethodni, ali se u suštini ni u čemu ne razlikuje od njega

površina x_1 [ft^2]	sobnost x_2	sprat x_3	starost x_4 [godina]	cena y [hiljade \$]
2104	6	1	45	460
1416	4	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...



  - Opšti oblik hipoteze je sada:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d$$

Linearna regresija za slučaj više obeležja

- Za kompaktniji prikaz izračunavanja koristi se vektorska notacija
  - Vektor obeležja treba proširiti pomoćnim obeležjem x_0 koje je *uvek jednako* 1 da bi se i parametar ϑ_0 mogao tretirati na isti način kao i ostali:

$$\begin{aligned} h_{\boldsymbol{\theta}}(\mathbf{x}) &= \vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d \\ &= \vartheta_0 \cdot 1 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d \\ &= \begin{bmatrix} \vartheta_0 & \vartheta_1 & \ldots & \vartheta_d \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} = \boldsymbol{\theta}^{\mathrm{T}} \mathbf{x} \end{aligned}$$

  - Funkcija cene jednaka je (kao i u slučaju sa jednim obeležjem):

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

    a algoritam gradijentnog silaska izvodi se na identičan način kao i u slučaju jednog obeležja, sa ovako definisanom funkcijom cene ($\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$)

Metod gradijentnog silaska za slučaj više obeležja

- Početi od proizvoljnih vrednosti $\vartheta_0, \vartheta_1, \ldots, \vartheta_d$
- Istovremeno promeniti $\vartheta_0, \vartheta_1 \ldots, \vartheta_d$ prema pravilu

$$\vartheta_0 \leftarrow \vartheta_0 - \alpha \frac{\partial}{\partial \vartheta_0} J(\vartheta_0, \vartheta_1, \ldots, \vartheta_d)$$

$$\vartheta_1 \leftarrow \vartheta_1 - \alpha \frac{\partial}{\partial \vartheta_1} J(\vartheta_0, \vartheta_1, \ldots, \vartheta_d)$$

$$\vdots$$

$$\vartheta_d \leftarrow \vartheta_d - \alpha \frac{\partial}{\partial \vartheta_d} J(\vartheta_0, \vartheta_1, \ldots, \vartheta_d)$$

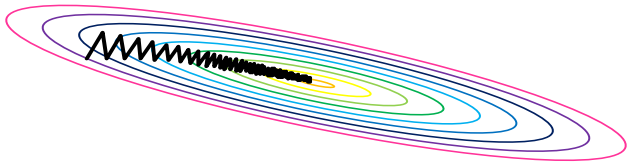
gde je α brzina učenja (fiksni mali broj)

- Ponavljati prethodni korak do konvergencije

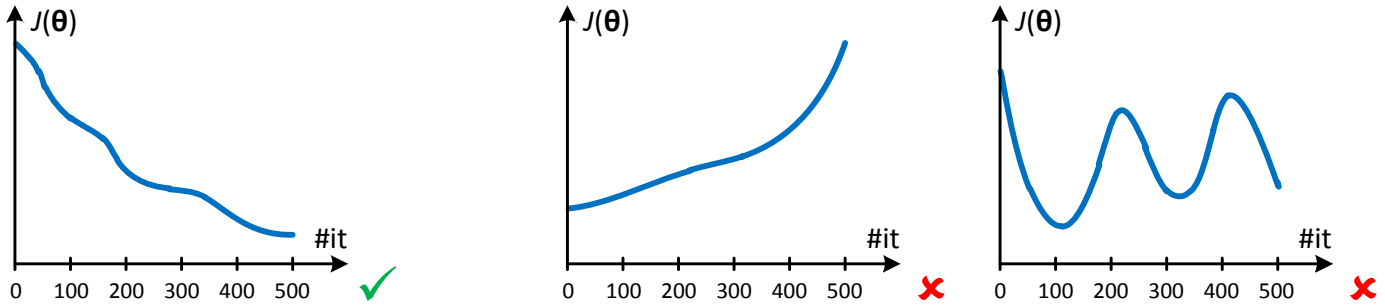
- Izloženi algoritam u opštem slučaju konvergira ka lokalnom minimumu
  - Ako je brzina učenja premala, algoritam sporo konvergira, ali ako je prevelika, konvergencija može biti ugrožena
  - Kako se trenutna vrednost ($\vartheta_0, \vartheta_1 \ldots, \vartheta_d$) približava lokalnom minimumu, tako koraci promene postaju sve manji, tako da nema potrebe da se brzina učenja α menja tokom izvršavanja algoritma

Metod gradijentnog silaska – korisni saveti

- Uzorke u skupu za obuku treba normalizovati
  - U suprotnom može nastupiti spora cik-cak konvergencija
  - Srednja vrednost oduzima se svim obeležjima osim x_0



- U zavisnosti od problema broj iteracija može biti znatno različit, pa treba uočiti kada konvergira
  - Zaustavni kriterijum može biti da promena $J(\boldsymbol{\theta})$ između dve iteracije bude manja od nekog veoma malog broja ε (npr. 10^{-3})
  - Veoma je teško naći adekvatnu vrednost za ε automatski
- Treba pratiti izgled funkcije $J(\boldsymbol{\theta})$ (npr. na svakih 100 iteracija)



- Izostanak konvergencije je često znak da treba smanjiti brzinu učenja α
  - Dobra praksa je prikazati grafik $J(\boldsymbol{\theta})$ za više brzina učenja α – npr. početi od 0.001 pa povećavati 3 puta u svakom koraku

Analitička minimizacija funkcije cene

- U jednostavnijim slučajevima (za ne suviše velik broj obeležja) problem minimizacije funkcije cene može se rešiti i analitički
- Kao funkcija cene u ovom slučaju se po pravilu koristi *kvadratna greška*:

$$J_{\text{MSE}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2 = \sum_{i=1}^{N} (\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}^{(i)} - y^{(i)})^2 = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2$$

pri čemu je:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(N)} \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_d^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_0^{(N)} & x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \\ \vdots \\ \vartheta_d \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

- Gradijent funkcije cene iznosi:

$$\nabla_{\boldsymbol{\theta}} J_{\text{MSE}}(\boldsymbol{\theta}) = 2\sum_{i=1}^{N} (\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}^{(i)} - y^{(i)})\mathbf{x}^{(i)} = 2\mathbf{X}^{\mathsf{T}}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

i njegovim izjednačavanjem sa nulom dobija se:

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$$

Analitička minimizacija funkcije cene

- Matrica $\mathbf{X}^T\mathbf{X}$ je *kvadratna* (bez obzira što $\mathbf{X}$ po pravilu nije), dimenzija $(d+1)\times(d+1)$
- Ako je $\mathbf{X}^T\mathbf{X}$ regularna (invertibilna) matrica, rešenje za $\boldsymbol{\theta}$ dato je izrazom:

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

  - $\mathbf{X}^+$ je pseudoinverzna matrica matrice $\mathbf{X}$
  - Za pseudoinverznu matricu $\mathbf{X}^+$ važi $\mathbf{X}^+\mathbf{X} = \mathbf{I}$, ali u opštem slučaju $\mathbf{X}\mathbf{X}^+ \neq \mathbf{I}$
- Nalaženje ovog rešenja može biti problematično u nekim slučajevima
  - Kada je broj obeležja prevelik (tipično preko 10.000), inverzija matrice $\mathbf{X}^T\mathbf{X}$ predstavlja računski izuzetno zahtevan zadatak
    - moguće rešenje je *redukcija dimenzionalnosti* (npr. prosto izdvajanje najbitnijih obeležja)
    - drugo moguće rešenje je primena metoda gradijentnog silaska
  - Kada je skup za obuku izuzetno korelisan (npr. među obeležjima se pojavljuju površina u m^2 i površina u ft^2), matrica $\mathbf{X}^T\mathbf{X}$ postaje *skoro singularna*
    - pri izračunavanju $(\mathbf{X}^T\mathbf{X})^{-1}$ dominiraju manje sopstvene vrednosti (šum), što izaziva numeričke probleme
    - pored navedenih pristupa rešavanju ovog problema moguće je primeniti i *regularizaciju*, što je ekvivalentno dodavanju malog umnoška jedinične matrice matrici $\mathbf{X}^T\mathbf{X}$, i naziva se *ridge* regresija (eng. *ridge* = greben)

Ridge regresija

- Modifikovana funkcija cene koja se minimizuje data je izrazom:

$$J_{\mathrm{RR}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{d} \vartheta_i^2 = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2 + \lambda \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\theta}$$

  gde je λ regularizacioni parametar

  - Novi član u $J_{\mathrm{RR}}(\boldsymbol{\theta})$ doprinosi smanjenju procene veličine pojedinih parametara ϑ_i

- Odgovarajuće rešenje za $\boldsymbol{\theta}$ dato je izrazom:

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^{\mathrm{T}} \mathbf{y}$$

  - Za $\lambda = 0$ prethodni izraz svodi se na pseudoinverzno rešenje (pošto se i funkcija cene $J_{\mathrm{RR}}(\boldsymbol{\theta})$ svodi na $J_{\mathrm{MSE}}(\boldsymbol{\theta})$)
  - Optimalna vrednost za λ obično se nalazi unakrsnom validacijom
  - Ako obeležja imaju bitno različite varijanse, umesto jedinične matrice može se koristiti dijagonalna matrica koja sadrži varijanse pojedinih obeležja

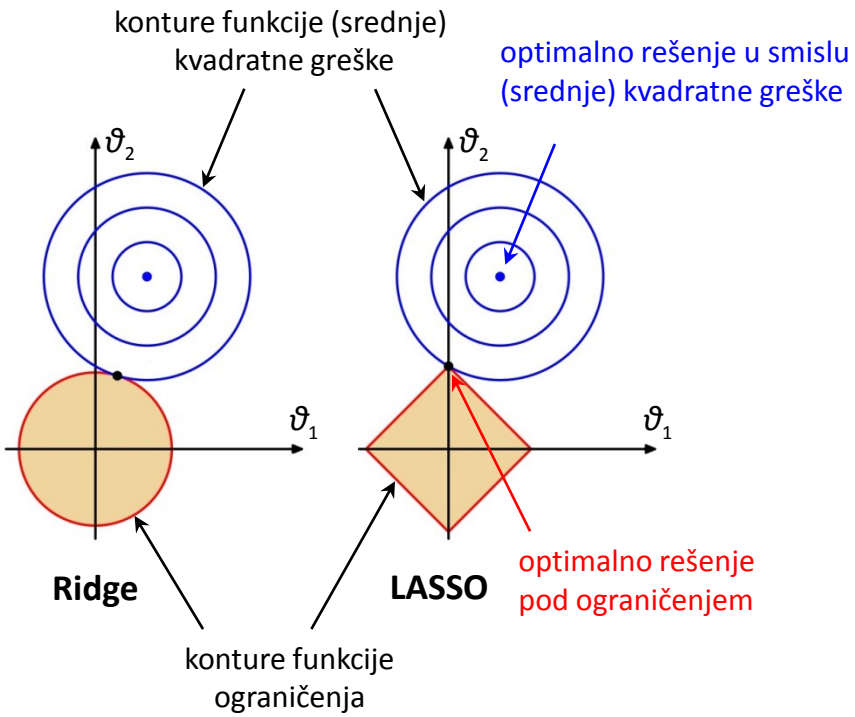
LASSO regresija

- Alternativa *ridge* regresiji kod koje je funkcija cene definisana kao:

$$J_{\text{LR}}(\boldsymbol{\theta}) = \sum_{i=1}^{N} (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2 + \lambda \sum_{i=1}^{d} |\vartheta_i|$$

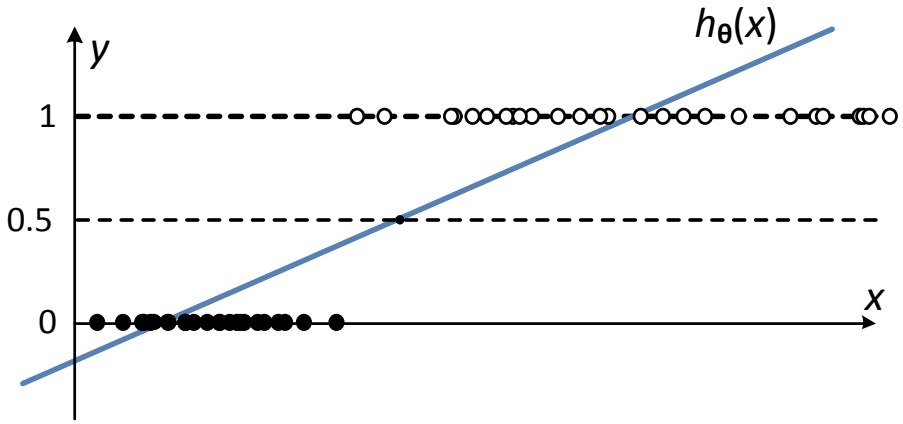
gde je λ regularizacioni parametar

  - Novi član u $J_{\text{LR}}(\boldsymbol{\theta})$ i ovde doprinosi smanjenju procene veličine pojedinih parametara ϑ_i, ali se ovde umesto l_2 norme minimizuje l_1 norma
  - Optimalna vrednost za λ i ovde se nalazi unakrsnom validacijom
  - Pokazuje se da će, uz pogodan izbor parametra λ jedan deo koeficijenata ϑ_i imati vrednost 0, što se suštinski svodi na **izbor jednog podskupa obeležja**
    - LASSO rezultuje retkim (eng. *sparse*) modelom



Primena linearne regresije u klasifikaciji

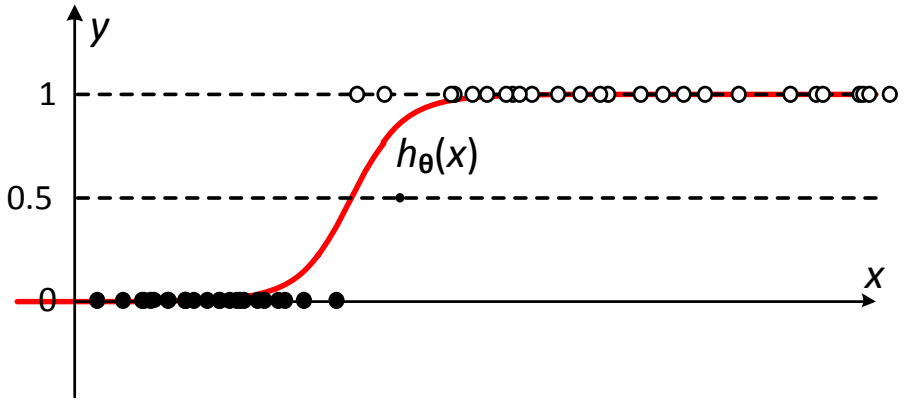
- Linearna regresija mogla bi se direktno primeniti i kao klasifikator, ali tu postoje određeni problemi:
  - Ako bi klase bile 0 i 1, dobija se procena koja nije ograničena na opseg između 0 i 1
  - Položaj praga je vrlo osetljiv na uzorački skup i zavisan od uzoraka koji zapravo nisu bitni



  - Nema procene verovatnoće pripadanja određenoj klasi
    - Poželjno je znati s kojom verovatnoćom neko ima rak ili ne, da li će neko vratiti dug ili ne...
- Bilo bi mnogo logičnije dozvoliti da izlazna veličina bude *nelinearna* funkcija *x*
  - U opštem slučaju, *y* treba da bude nelinearna funkcija linearne kombinacije ulaznih promenljivih

Primena linearne regresije u klasifikaciji

- Linearna regresija mogla bi se direktno primeniti i kao klasifikator, ali tu postoje određeni problemi:
  - Ako bi klase bile 0 i 1, dobija se procena koja nije ograničena na opseg između 0 i 1
  - Položaj praga je vrlo osetljiv na uzorački skup i zavisan od uzoraka koji zapravo nisu bitni



  - Nema procene verovatnoće pripadanja određenoj klasi
    - Poželjno je znati s kojom verovatnoćom neko ima rak ili ne, da li će neko vratiti dug ili ne...
- Bilo bi mnogo logičnije dozvoliti da izlazna veličina bude *nelinearna* funkcija *x*
  - U opštem slučaju, *y* treba da bude nelinearna funkcija linearne kombinacije ulaznih promenljivih

Modifikacije linearne regresije

- Da bi se povećale mogućnosti modela, on se može uopštiti uvođenjem nelinearnih funkcija koje se primenjuju direktno nad vektorom obeležja:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \vartheta_0 + \vartheta_1 f_1(x_1) + \vartheta_2 f_2(x_2) + \ldots + \vartheta_d f_d(x_d)$$

  ili u opštem slučaju:

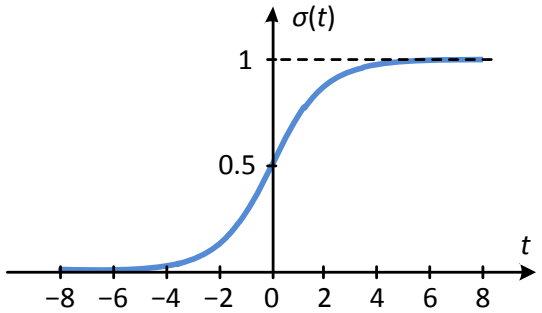
$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \vartheta_0 + \sum_{i=1}^{d} \vartheta_i f_i(\mathbf{x})$$

- Funkcije $f_i(\mathbf{x})$ nazivaju se *baznim funkcijama*
  - Primera radi, za $f_1(x)=x$, $f_2(x)=x^2$,... $f_d(x)=x^d$ (za slučaj jednog obeležja) dobila bi se polinomijalna regresija
  - Na ovaj način u model se može uvesti i interakcija između pojedinih obeležja, npr. $f_1(x)=x_1$, $f_2(x)=x_2$, $f_3(x)=x_1x_2$ (za slučaj dva obeležja)
    - I kada se uvodi interakcija, u modelu treba ostaviti i polazna obeležja
- Uvođenjem baznih funkcija deskriptivnost modela se povećava, ali se povećava i njegova složenost (što otežava obuku)
  - Iako su bazne funkcije nelinearne, model je i dalje linearan jer $h_{\boldsymbol{\theta}}(\mathbf{x})$ linearno zavisi od $\boldsymbol{\theta}$, a $\boldsymbol{\theta}$ je ono što se određuje

Logistička regresija

- Logistička regresija dobija se primenom *sigmoida* (logističke funkcije) na linearnu kombinaciju ulaznih promenljivih **x**

$$h_{\boldsymbol{\vartheta}}(\mathbf{x}) = \sigma(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d), \quad \sigma(t) = \frac{1}{1+e^{-t}}$$



- Dobijena nelinearna funkcija interpretira se kao *matematičko očekivanje vrednosti* izlazne promenljive

$$h_{\boldsymbol{\vartheta}}(\mathbf{x}) = E\{y \mid \mathbf{x}\} = \sigma(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \ldots + \vartheta_d x_d)$$

  - Ovo za posledicu ima da je veza između *y* i **x** izražena kroz vrednosti *verovatnoće*
  - U slučaju binarne klasifikacije, ovom očekivanju odgovara verovatnoća klase *y* = 1

$$E\{y \mid \mathbf{x}\} = 0 \cdot \mathrm{P}(y = 0 \mid \mathbf{x}) + 1 \cdot \mathrm{P}(y = 1 \mid \mathbf{x}) = \mathrm{P}(y = 1 \mid \mathbf{x})$$

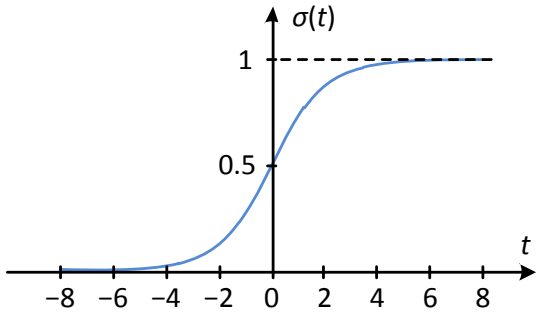
  - $h_{\boldsymbol{\vartheta}}(\mathbf{x})$ se može interpretirati i kao diskriminantna funkcija određene klase (s različitim parametrima $\boldsymbol{\vartheta}$ za svaku klasu), što omogućuje i klasifikaciju u više od dve klase
    - Odlučuje se po tome koja diskriminantna funkcija ima najveći „odziv“ na **x**
    - Vrednost određene diskriminantne funkcije ujedno predstavlja i verovatnoću pripadnosti uzorka **x** odgovarajućoj klasi: $p_k = \mathrm{P}(y = k \mid \mathbf{x})$

Logistička regresija

- Ako je *y* binarna promenljiva, dobija se:

$$\mathrm{P}(y=1\,|\,\mathbf{x}) = \sigma(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}) = \frac{e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}{1+e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}$$

$$\mathrm{P}(y=0\,|\,\mathbf{x}) = 1-\mathrm{P}(y=1\,|\,\mathbf{x}) = \frac{1}{1+e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}$$



  - Ovim se praktično linearno modeluje logaritam *kvote*:

KVOTA: $\dfrac{\mathrm{P}(y=1\,|\,\mathbf{x})}{\mathrm{P}(y=0\,|\,\mathbf{x})} = e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}$

$$\ln\frac{\mathrm{P}(y=1\,|\,\mathbf{x})}{\mathrm{P}(y=0\,|\,\mathbf{x})} = \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x} = \vartheta_0 + \vartheta_1 x_1 + \ldots + \vartheta_d x_d$$

- Slično važi i u opštem slučaju *K* klasa (*K* > 2):

$$\ln\frac{\mathrm{P}(y=k\,|\,\mathbf{x})}{1-\mathrm{P}(y=k\,|\,\mathbf{x})} = \boldsymbol{\theta}^{(k)\mathrm{T}}\mathbf{x} = \vartheta_0^{(k)} + \vartheta_1^{(k)} x_1 + \ldots + \vartheta_d^{(k)} x_d$$

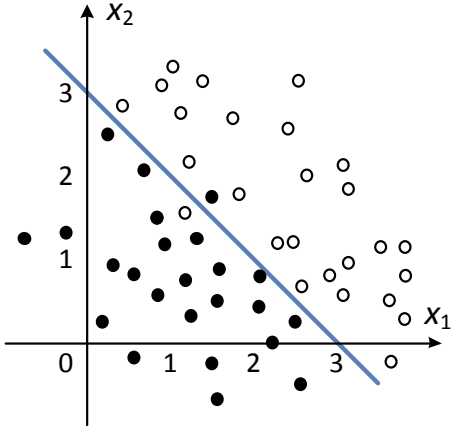
  - Logaritam kvote naziva se *logit* funkcija

Logistička regresija – granice odlučivanja

- Ako je *y* nelinearna funkcija *linearne* kombinacije ulaznih promenljivih:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2)$$

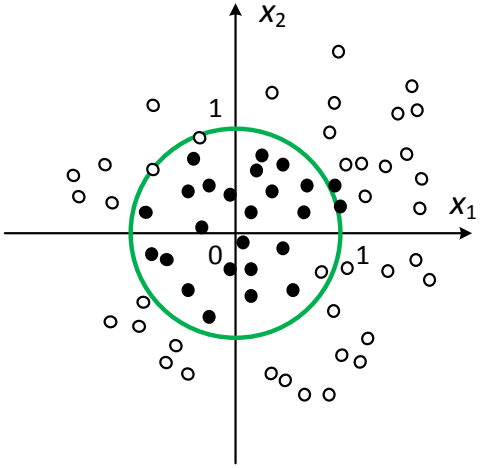
$$y = 1 \quad \text{ako} \quad -3 + x_1 + x_2 \geq 0$$



- Ako je *y* nelinearna funkcija *nelinearne* kombinacije ulaznih promenljivih:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \sigma(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \vartheta_3 x_1^2 + \vartheta_4 x_2^2)$$

$$y = 1 \quad \text{ako} \quad -1 + x_1^2 + x_2^2 \geq 0$$



Logistička regresija – nalaženje $\boldsymbol{\theta}$

- Nalaženje koeficijenata ima sličnosti sa estimacijom parametara gustine raspodele verovatnoće
  - $\boldsymbol{\theta}$ se može naći *metodom maksimalne izglednosti* skupa uzoraka

$$l(\boldsymbol{\theta}) = \prod_{i, y^{(i)}=1} \mathrm{P}(y=1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}) \prod_{i, y^{(i)}=0} (1 - \mathrm{P}(y=1 \mid \mathbf{x}^{(i)}, \boldsymbol{\theta}))$$

  - Maksimalno izgledna vrednost vektora parametara $\boldsymbol{\theta}$ je ona za koju se minimizuje funkcija cene, što se može izvesti metodom gradijentnog silaska

$$J(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Cost}(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = \begin{cases} -\ln h_{\boldsymbol{\theta}}(\mathbf{x}), & y = 1 \\ -\ln(1 - h_{\boldsymbol{\theta}}(\mathbf{x})), & y = 0 \end{cases}$$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$