Regularizacija

- Modifikacija funkcije greške: umesto direktne minimizacije gubitka

$$L\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

minimizujemo regularizovanu grešku:

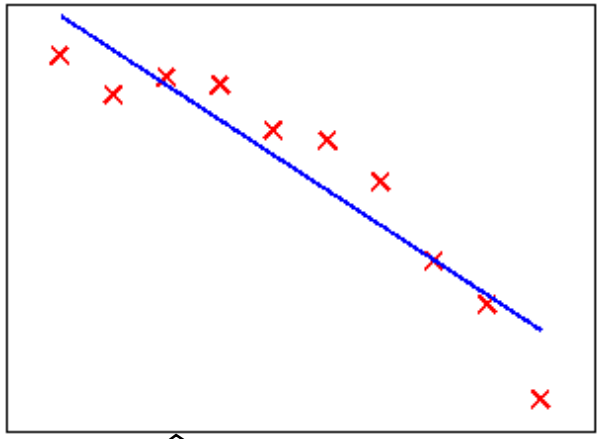
$$\min_\theta \frac{1}{N}\sum_{i=1}^{N} L\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \lambda\Omega(\theta)$$

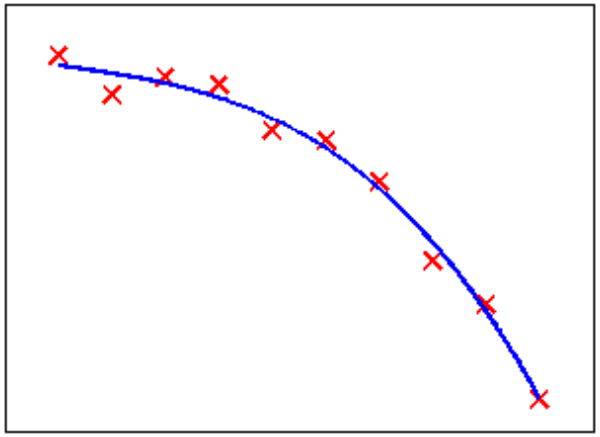
Regularizacioni parametar

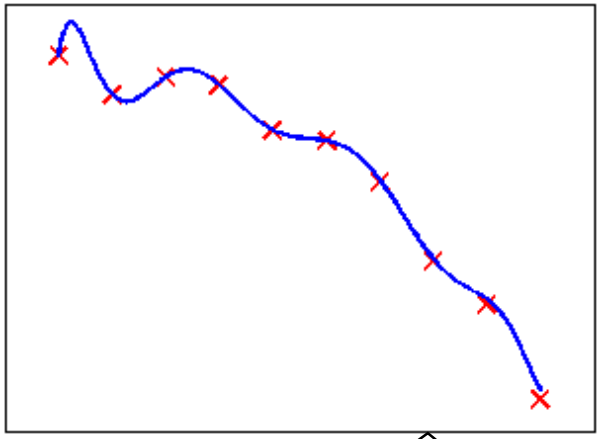
Regularizacioni izraz

Zašto model greši?

1. i 3. model imaju veliku generalizacionu grešku, ali su razlozi različiti







Veliko sistematsko odstupanje

Za dobru generalizaciju ključno je upravljanje prilagodljivošću modela

Velika varijansa

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \lambda\Omega(\theta)$$

Regularizacija

- Regularizacioni parametar λ služi za fino podešavanje prilagodljivosti modela tako što kontroliše nagodbu dva cilja:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \lambda \Omega(\theta)$$

Minimizacija gubitka rezultuje prediktivnim modelom.

Čini da se dobro uklopimo u trening podatke

Minimizacija Ω rezultuje jednostavnim modelom.

Jednostavniji modeli imaju manju varijansu u predikcijama i stabilniji su

$\lambda = 0 \Rightarrow$ nema regularizacije

Velika varijansa, ali malo sistematsko odstupanje

$\lambda = \infty \Rightarrow$ parametri θ će biti 0

Veliko sistematsko odstupanje, ali mala varijansa

Regularizacioni izraz Ω

1) Suma apsolutnih vrednosti (L_1 norma): Lasso regression/L_1 regularizacija

$$\Omega = \sum_{d=1}^{D} |\theta_d| = \|\theta\|_1$$

2) Suma kvadrata vrednosti (kvadrirana L_2 norma): Ridge regression/L_2 regularizacija

$$\Omega = \sum_{j=1}^{d} {\theta_j}^2 = \|\theta\|_2^2$$

Ridge regression (L_2)

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\theta_j^2$$

RSS: model treba što više da odgovara trening skupu

Magnitude koeficijenata θ treba da su što manje

θ_0 nije uključen u regularizacioni izraz

Ridge regression (L_2)

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\theta_j^2$$

RSS: model treba što više da odgovara trening skupu

Magnitude koeficijenata θ treba da su što manje

λ kontroliše nagodbu ova dva cilja

$\lambda = 0$

OLS metod (bez regularizacije)

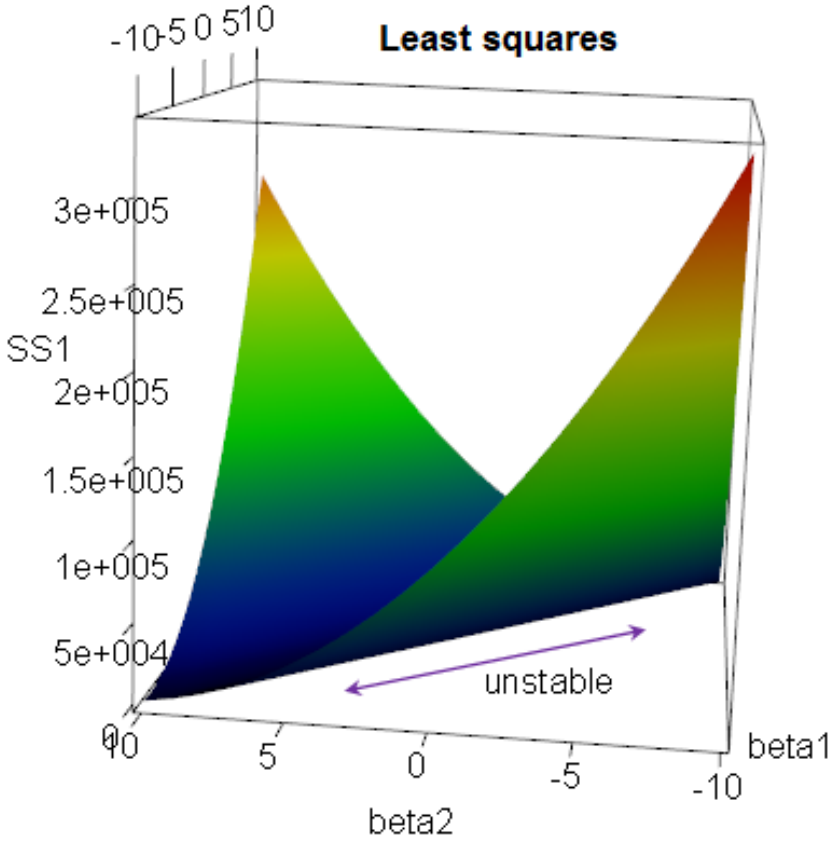
$\lambda = \infty \Rightarrow \theta = 0$

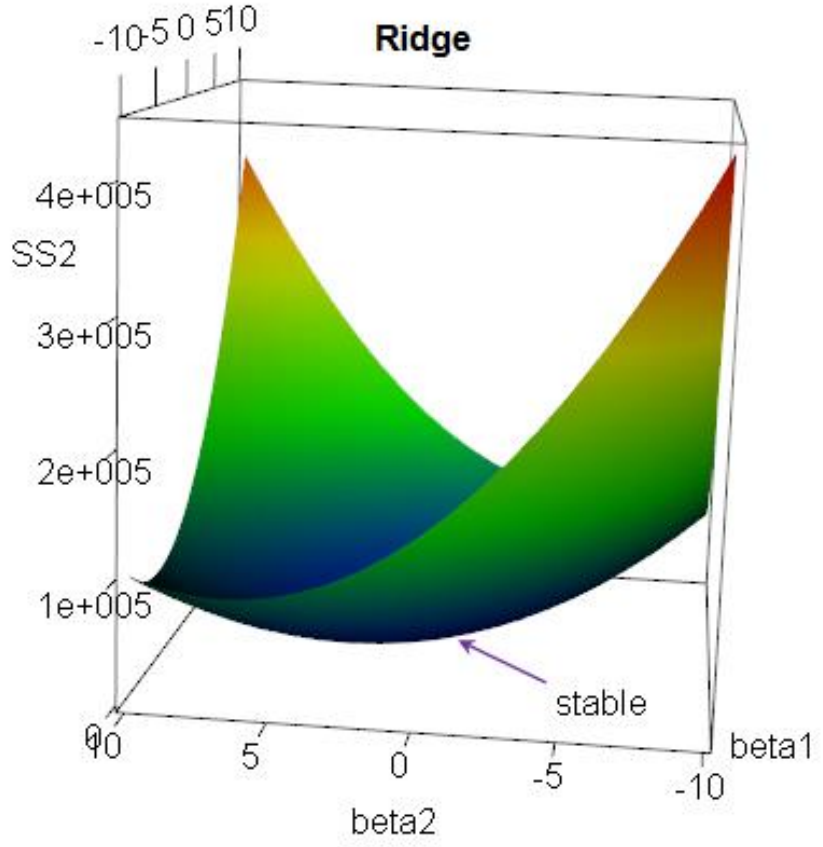
naš model postaje konstanta $h_\theta(x) = \theta_0$

Zašto se zove *ridge* (grebena)?

Multikolinearnost: funkcija cilja nema jedinstven minimum, već greben

Ridge će zameniti „greben“ minimumom





Ridge regression gradient descent

Bez regularizacije:

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \frac{\alpha}{N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j{}^{(i)}$$

Sa regularizacijom:

$$\theta_0^{(t+1)} = \theta_0^{(t)} - \frac{\alpha}{N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0{}^{(i)}$$

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \frac{\alpha}{N}\sum_{i=1}^{N}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j{}^{(i)} + \alpha\lambda\theta_j$$

$$\theta_j^{(t+1)} = \theta_j^{(t)}\,(1 - \alpha\lambda) - \frac{\alpha}{N}\sum_{i=1}^{N}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j{}^{(i)}$$

U svakom koraku smanjujemo θ

Ridge regression closed form solution

Bez regularizacije: $\theta = (X^T X)^{-1} X^T \mathrm{y}$ $\quad X = \begin{bmatrix} 1 & x_1{}^{(1)} & \dots & x_D{}^{(1)} \\ 1 & x_1{}^{(2)} & \dots & x_D{}^{(2)} \\ \dots & \dots & & \\ 1 & x_1{}^{(N)} & \dots & x_D{}^{(N)} \end{bmatrix}$

U slučaju $N \leq D$ (*N* – broj primera, *D* – broj obeležja) $X^T X$ je singularna matrica (nije invertibilna)

Sa regularizacijom (za λ > 0):

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \dots & \\ & & & 1 \end{bmatrix} \right)^{-1} X^T y$$

I u slučaju $N \leq D$ ova matrica se može invertovati

Priprema podataka za regularizaciju

Podatke je neophodno normalizovati

$$x_j^{(i)} = \frac{x_j^{(i)} - \mu_j}{\sigma_j}$$

- Neka se sve vrednosti x kreću u istom opsegu $[-1, 1]$
- Uzmimo obeležje x_j i pomnožimo ga se 10^{-6}
- Bez regularizacije: $\theta_j^{\text{new}} = \theta_j^{\text{old}} \cdot 10^6$
- Regularizacija: tretira sva obeležja jednako – praktično bi uticala samo na θ_j

Tretiranje θ_0 (intercept)

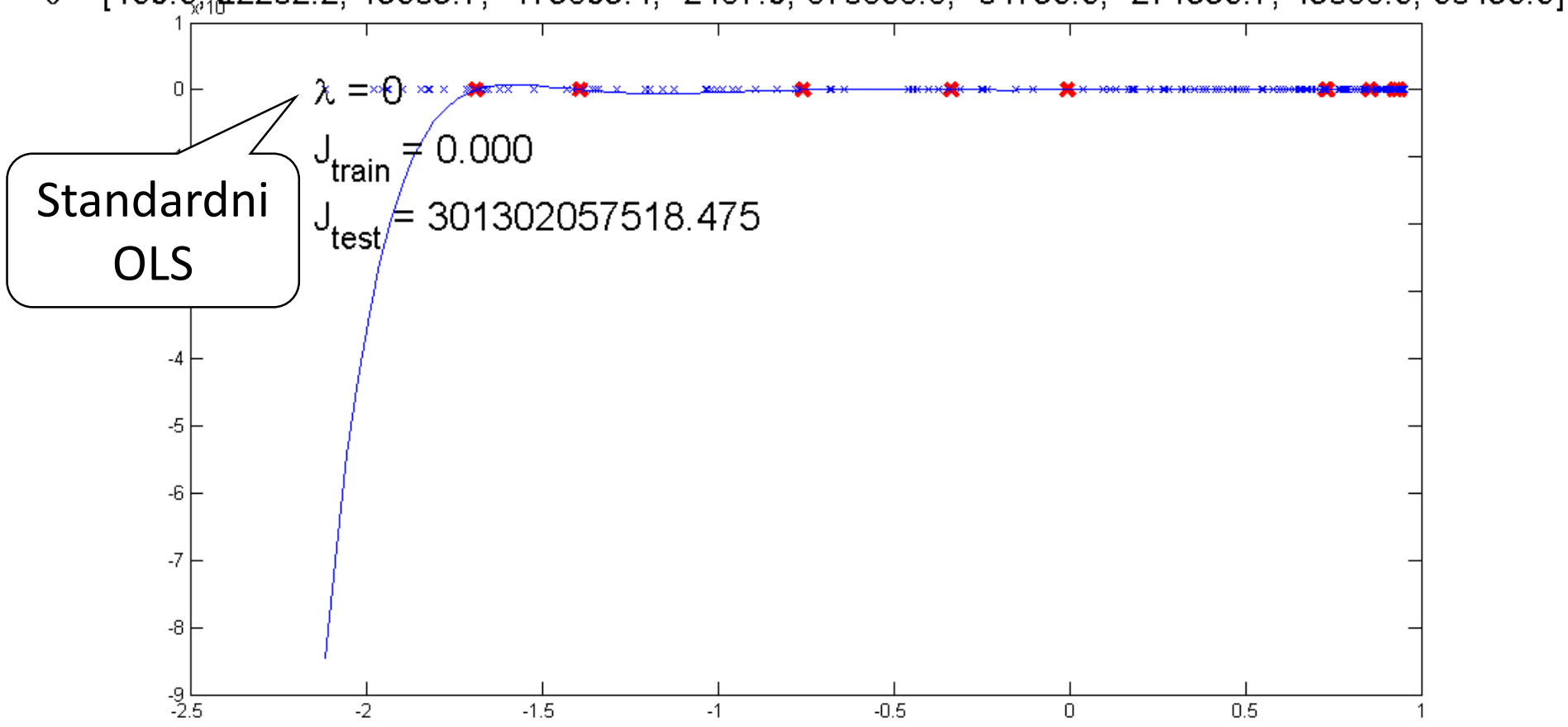
- θ_0 je očekivana vrednosti y kada su svi ulazi 0

- Nema smisla da θ_0 forsiramo da bude malo jer konceptualno nije indikator preprilagođavanja

- Da ne bismo u jednačinama zasebno tretirali θ_0 možemo:
  1. *centrirati podatke oko 0* (transformisati y da ima srednju vrednost 0)
  2. Primeniti ridge regresiju tretirajući sve θ jednako

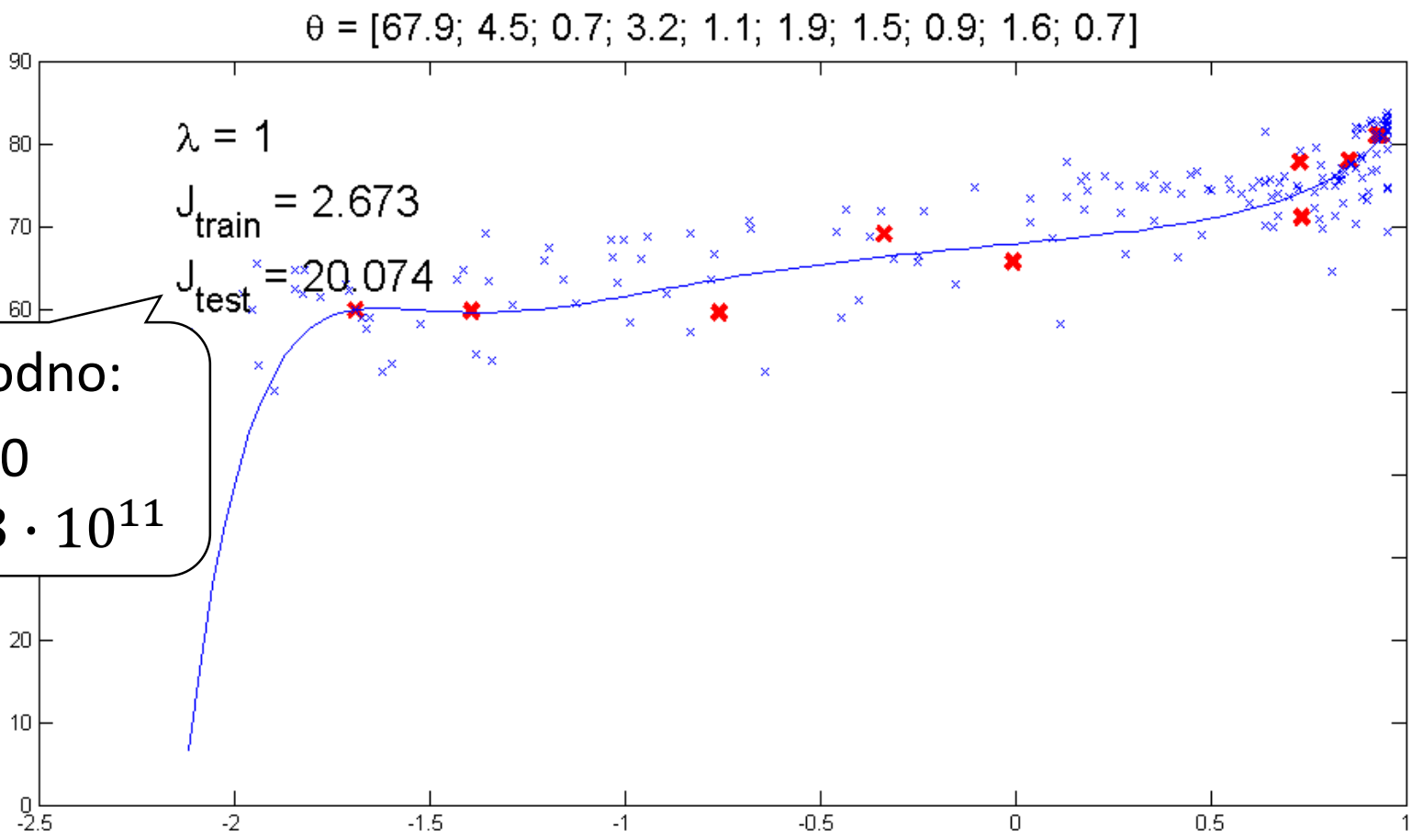
Kako odabrati λ?

θ = [160.6; 22282.2; 19689.7; -179309.4; -2457.0; 378365.5; -84796.6; -271996.7; 49835.6; 68496.6]



3_Regularizacija_odabir_lambda\Start.m

Kako odabrati λ?

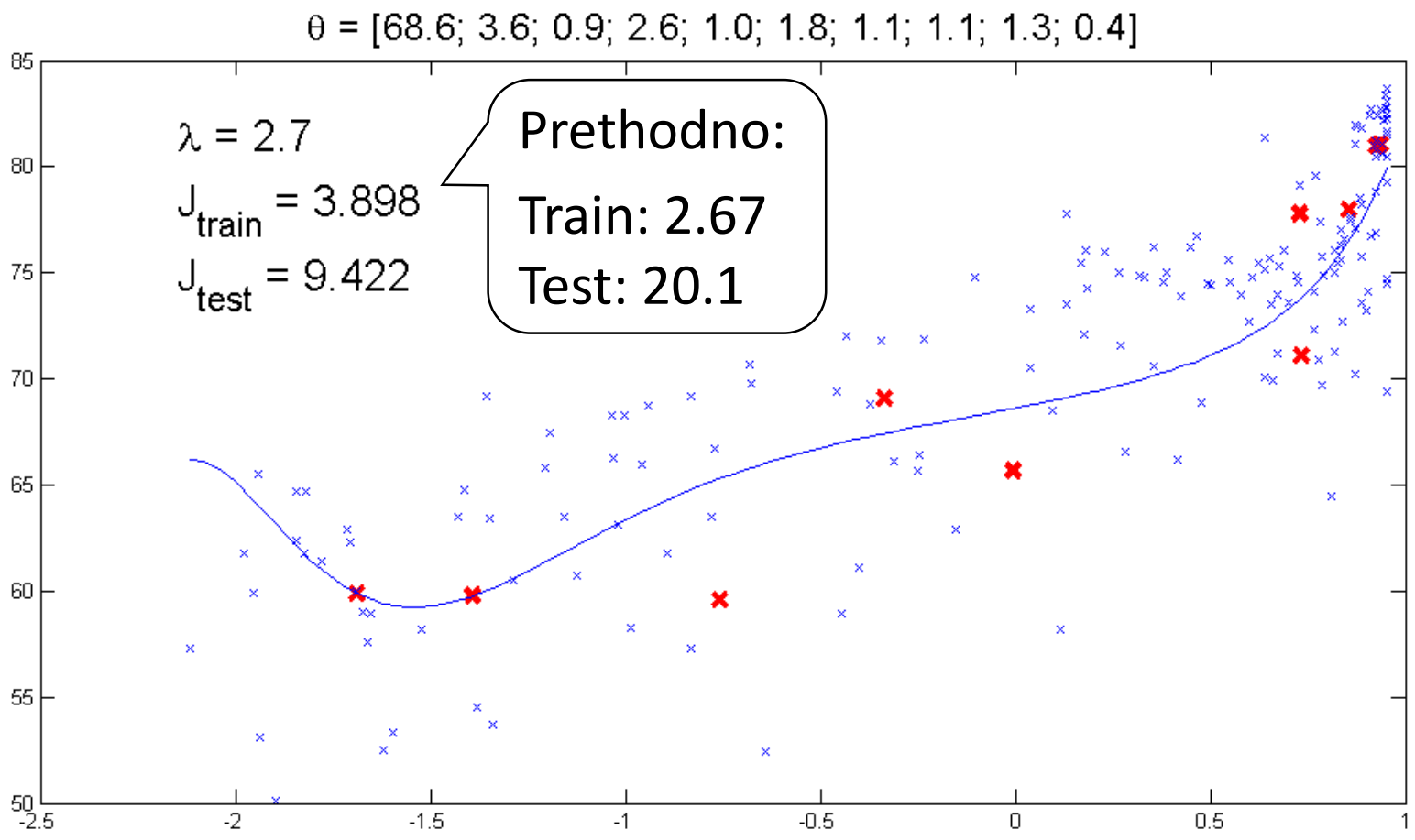


Prethodno:

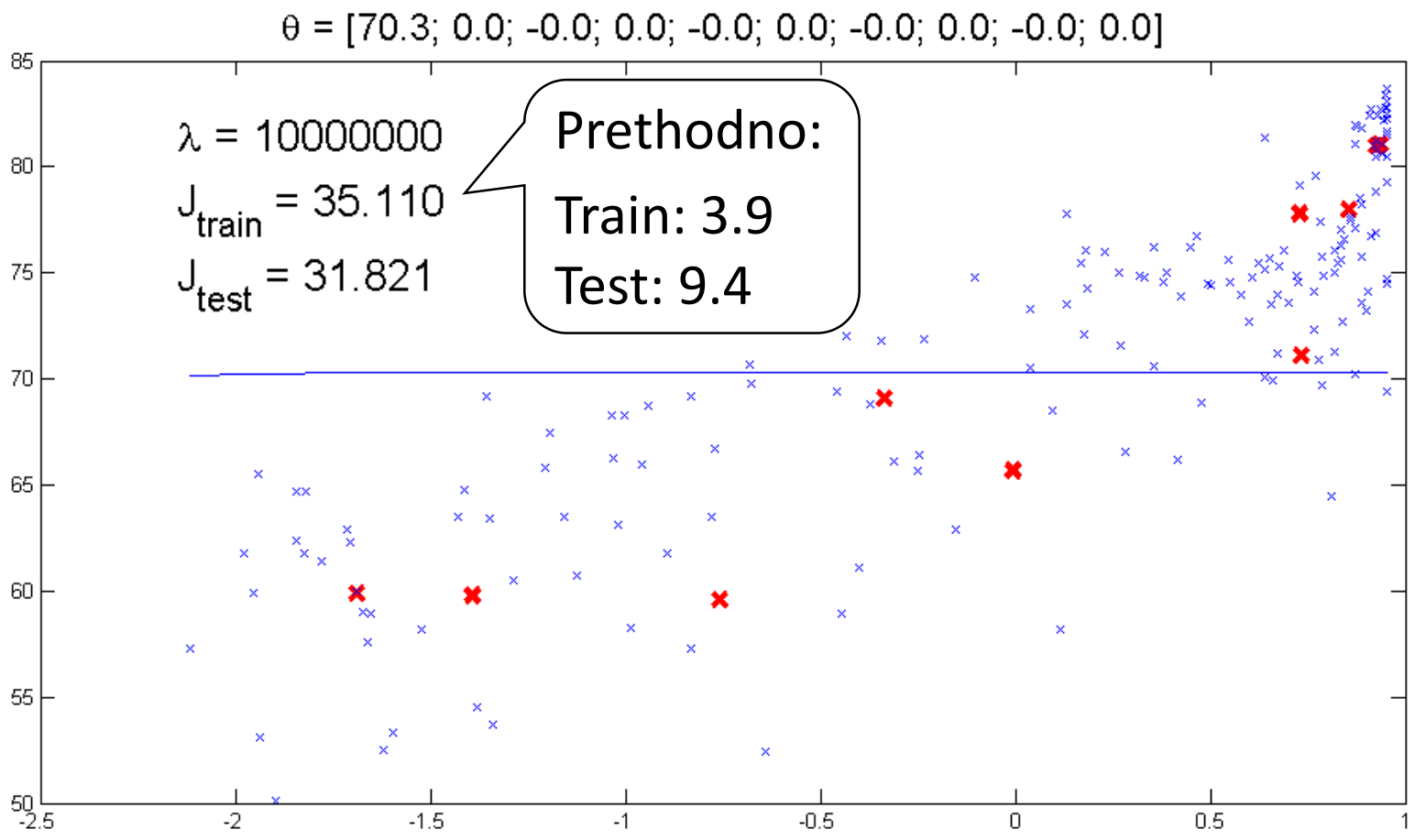
Train: 0

Test: $3 \cdot 10^{11}$

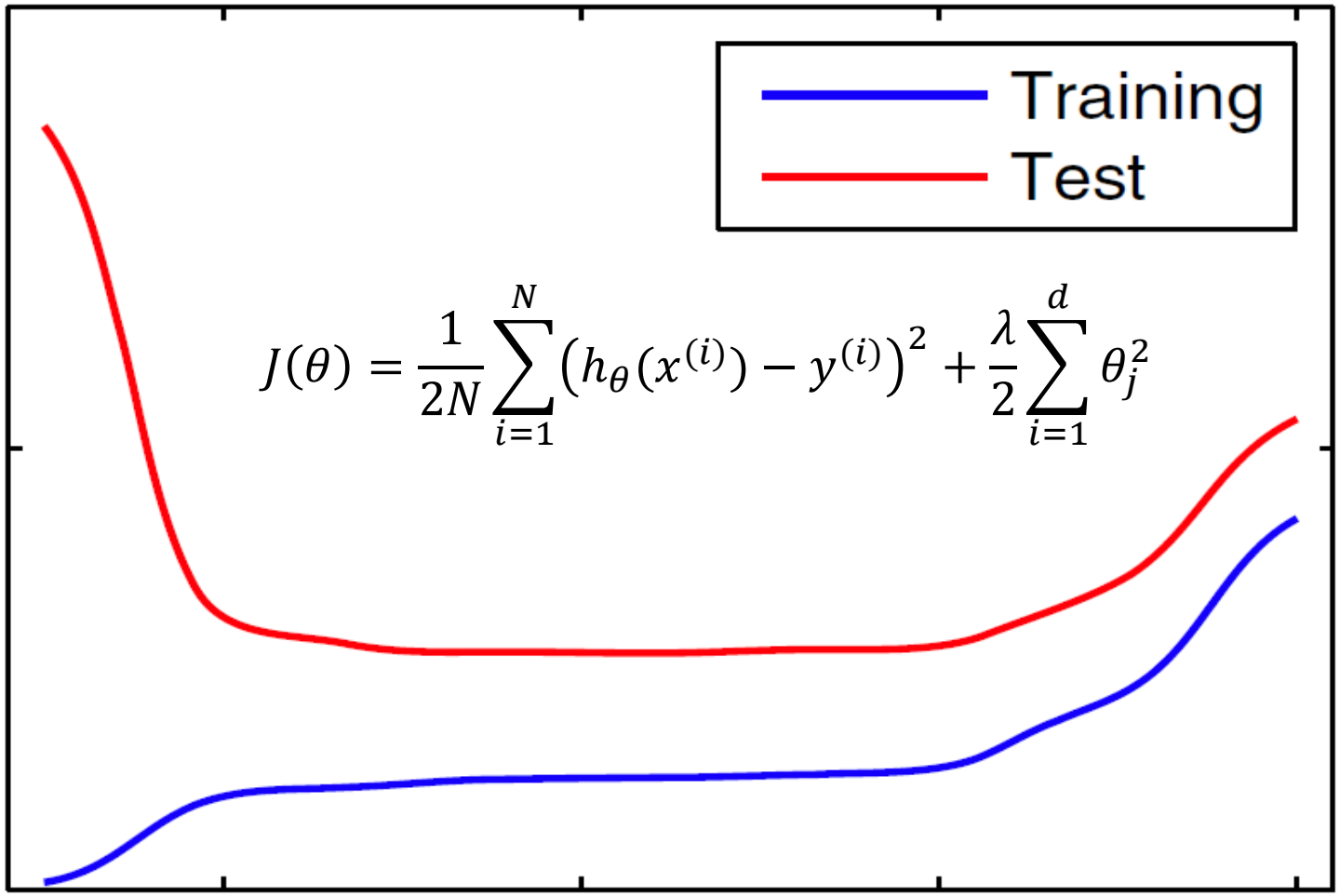
Kako odabrati λ?



Kako odabrati λ?



Kako odabrati λ?



Malo λ: overfit

Veliko λ: underfit

Kako odabrati λ?

- Isprobavanjem:
  1. $\lambda = 0 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(1)}$
  2. $\lambda = 0.01 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(2)}$
  3. $\lambda = 0.02 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(3)}$
  4. $\lambda = 0.04 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(4)}$
  5. $\lambda = 0.08 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(5)}$

  ...

  12. $\lambda = 10 \rightarrow \min_{\theta} J(\theta) \rightarrow \theta^{(10)}$

- Evaluiraćemo svaki od ovih 12 modela i odabraćemo onaj sa najmanjom greškom

Regression workflow

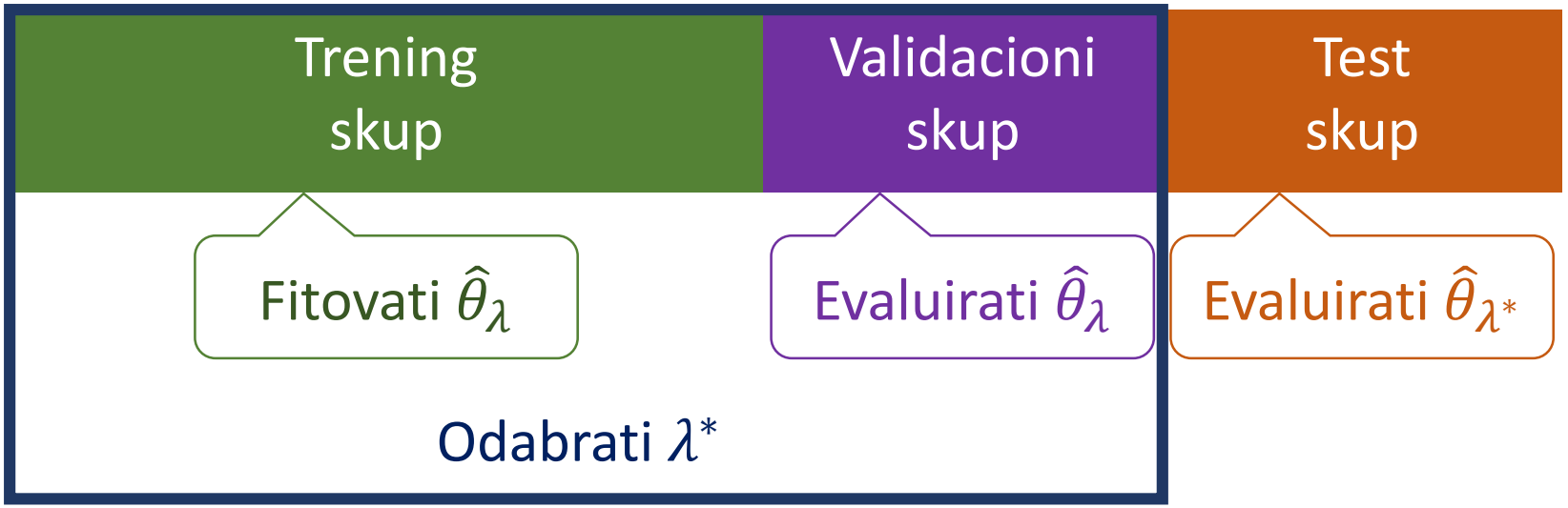
1. Selekcija modela: za svako λ:
   i. Estimirati parametre $\hat{\theta}_\lambda$ na **trening** podacima
   ii. Proceniti performanse modela $\hat{\theta}_\lambda$ na **test** podacima
   iii. Odabrati λ^* kao λ sa najmanjom greškom na **test** skupu

2. Evaluacija modela
   - Generalizacionu grešku modela proceniti računanjem greške modela $\hat{\theta}_{\lambda^*}$ na **test** skupu

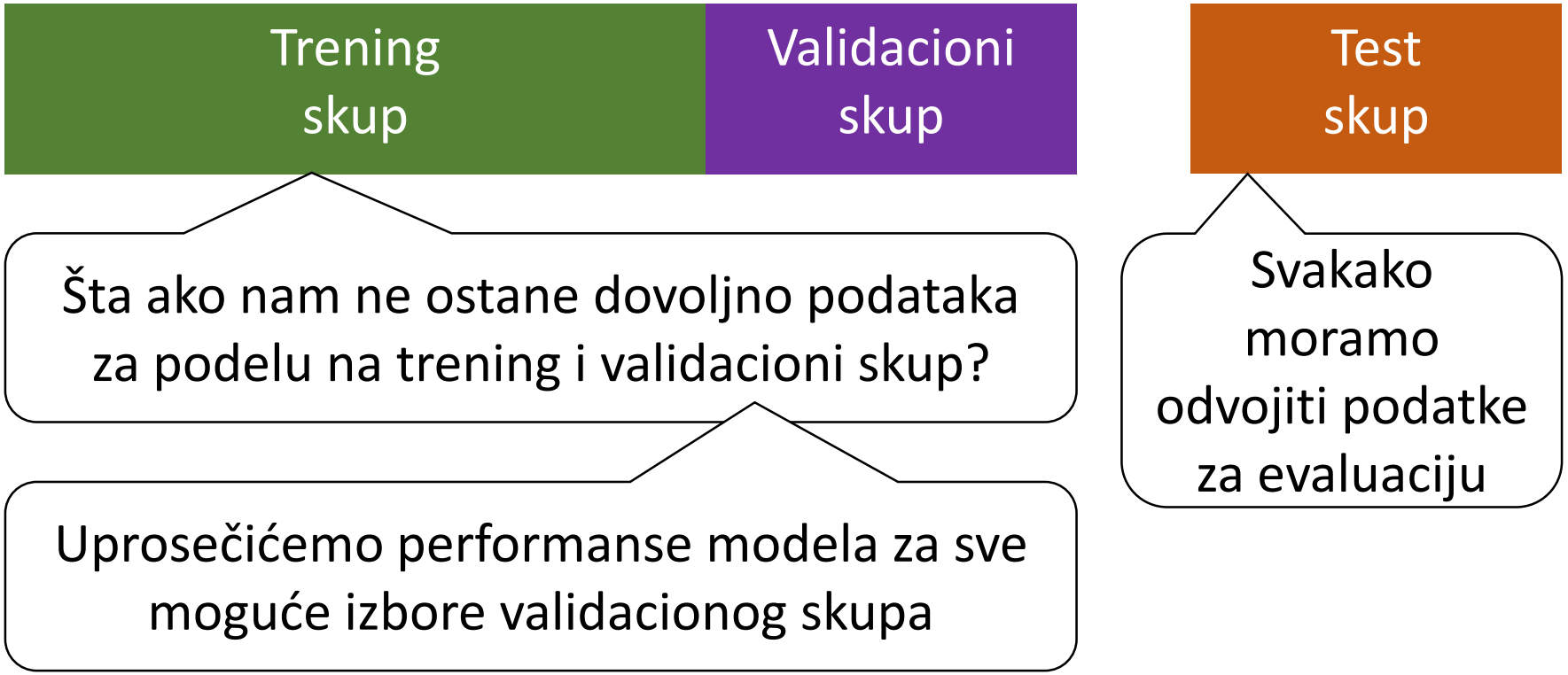
λ je još jedan parameter modela:
ne smemo ga optimizovati na test skupu!

Regression workflow

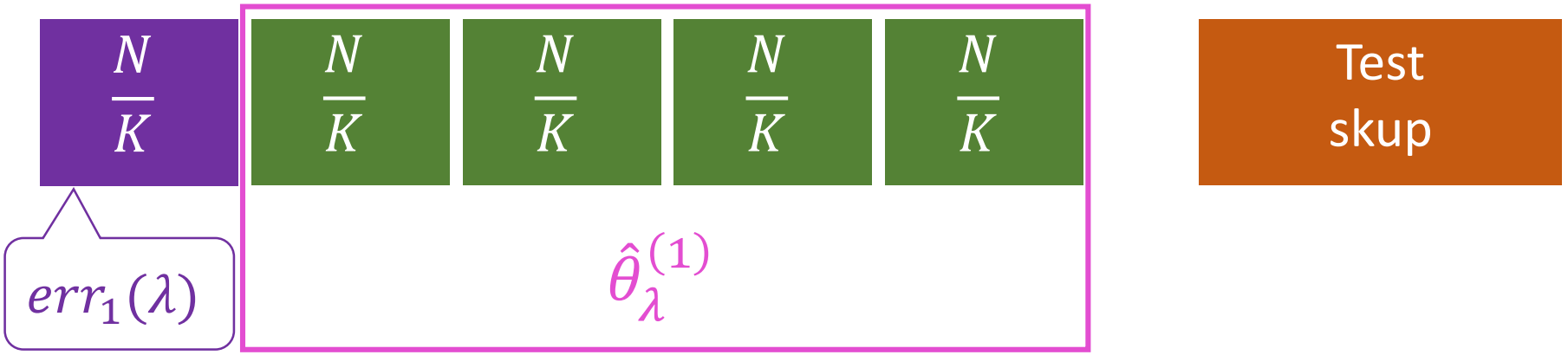
- Podeliti dostupne podatke na:
  - *trening* skup – treniranje modela
  - *validation* (*hold-out*) skup – optimizacija kompleksnosti modela
  - *test* skup – evaluacija performansi modela
- Za svako λ:



Unakrsna validacija



Unakrsna validacija



- Izmešati trening podatke i podeliti ih na *K* jednakih delova
- For *k*=1,…,*K*
  - Izdvojiti *k*-ti deo skupa kao validacioni skup
  - Na ostatku trening skupa proceniti parametre $\hat{\theta}_\lambda^{(k)}$
  - Odrediti grešku modela na validacionom skupu: $err_k(\lambda)$
- Izračunati prosečnu grešku $CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} err_k(\lambda)$

Unakrsna validacija

- Postupak unakrsne validacije bismo ponovili za svaku vrednost λ koju razmatramo
- Odabiramo λ^* za koje je $CV(\lambda^*)$ minimalno

- Koliko *K* odabrati?
  - Najbolja aproksimacija se dobija za *K*=*N* (validacioni skup od jednog primera) – *leave-one-out cross validation*
  - Ova procedura je računarski zahtevna – zahteva fitovanje *N* modela za svaku razmatranu vrednost λ
  - Tipično se koristi *K*=5 (5-*fold CV*) ili *K*=10 (10-*fold CV*)

Sumarizacija

- Regularizacija:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \lambda\Omega(\theta)$$

- L_2 regularizacija/ridge regression:

$$\Omega(\theta) = \sum_{i=1}^{d} \theta_j^2$$

- Pričali smo o tome kako odabrati λ