

Regularizacija

Ridge (L2), Lasso (L1), Elastic Net

Ordinary Least Squares (OLS)

- OLS model:

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

- Problem: $X^T X$ može biti singularna matrica
 - Kada je broj obeležja veći od broja instanci ($D > N$)
 - Kada su prediktori multikolinearni

Slučaj $D > N$ nije redak u praksi

Klasifikacija tekstualnih dokumenata

Raw Text

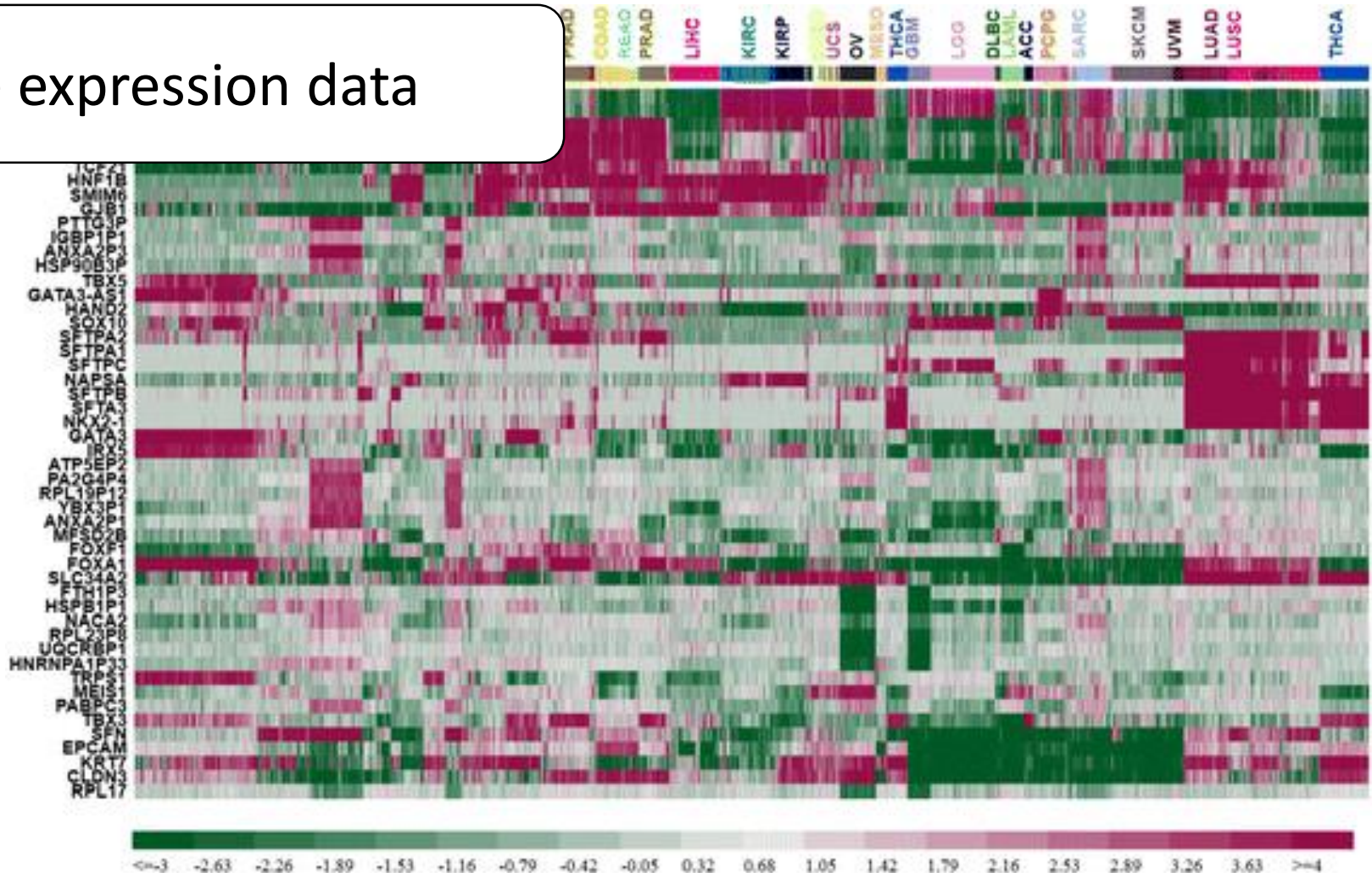
it is a puppy and it
is extremely cute

**Bag-of-words
vector**

| | |
|-----------|-----|
| it | 2 |
| they | 0 |
| puppy | 1 |
| and | 1 |
| cat | 0 |
| aardvark | 0 |
| cute | 1 |
| extremely | 1 |
| ... | ... |

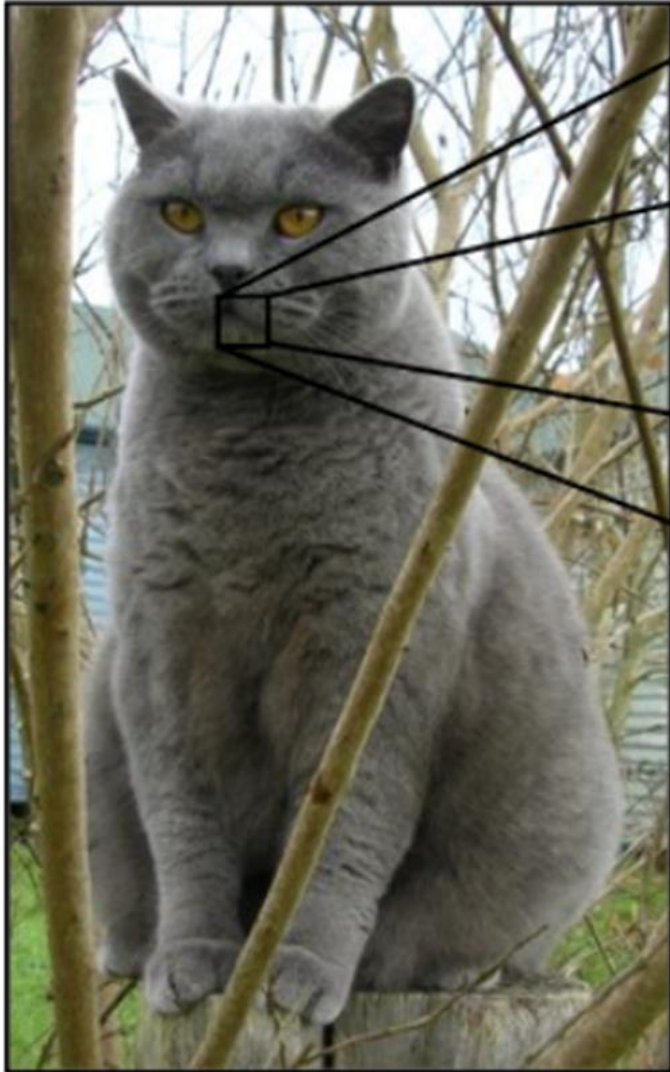
Slučaj $D > N$ nije redak u praksi

Gene expression data



Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L., 2017. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC genomics*, 18(1), p.508.

Slučaj $D > N$ nije redak u praksi



| | | | | | | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 08 | 02 | 22 | 97 | 38 | 15 | 00 | 40 | 00 | 75 | 04 | 05 | 07 | 78 | 52 | 12 | 50 | 77 | 81 | 22 |
| 49 | 49 | 99 | 40 | 17 | 81 | 18 | 57 | 60 | 87 | 17 | 40 | 98 | 43 | 69 | 48 | 04 | 56 | 62 | 00 |
| 81 | 49 | 31 | 73 | 55 | 79 | 14 | 29 | 93 | 71 | 40 | 67 | 58 | 88 | 30 | 03 | 49 | 13 | 36 | 65 |
| 52 | 70 | 95 | 23 | 04 | 60 | 11 | 42 | 69 | 24 | 68 | 56 | 01 | 32 | 56 | 71 | 37 | 02 | 36 | 91 |
| 22 | 31 | 16 | 71 | 51 | 67 | 02 | 89 | 41 | 92 | 36 | 54 | 22 | 40 | 40 | 28 | 66 | 33 | 13 | 80 |
| 24 | 47 | 32 | 60 | 99 | 03 | 45 | 02 | 44 | 75 | 33 | 53 | 78 | 36 | 84 | 20 | 35 | 17 | 12 | 50 |
| 52 | 98 | 81 | 28 | 64 | 23 | 67 | 10 | 26 | 38 | 40 | 67 | 59 | 54 | 70 | 66 | 18 | 38 | 64 | 70 |
| 67 | 26 | 20 | 68 | 02 | 62 | 12 | 20 | 95 | 63 | 94 | 39 | 63 | 08 | 40 | 91 | 66 | 49 | 94 | 21 |
| 24 | 55 | 58 | 05 | 66 | 73 | 99 | 26 | 97 | 17 | 78 | 78 | 96 | 83 | 14 | 88 | 34 | 89 | 63 | 72 |
| 21 | 36 | 23 | 09 | 75 | 00 | 76 | 44 | 20 | 45 | 35 | 14 | 00 | 61 | 33 | 97 | 34 | 31 | 33 | 95 |
| 78 | 17 | 53 | 28 | 22 | 75 | 31 | 67 | 15 | 94 | 03 | 80 | 04 | 62 | 16 | 14 | 09 | 53 | 56 | 92 |
| 16 | 39 | 05 | 42 | 96 | 35 | 31 | 47 | 55 | 58 | 88 | 24 | 00 | 17 | 54 | 24 | 36 | 29 | 85 | 57 |
| 86 | 56 | 00 | 48 | 35 | 71 | 89 | 07 | 05 | 44 | 44 | 37 | 44 | 60 | 21 | 58 | 51 | 54 | 17 | 58 |
| 19 | 80 | 81 | 68 | 05 | 94 | 47 | 69 | 28 | 73 | 92 | 13 | 86 | 52 | 17 | 77 | 04 | 89 | 55 | 40 |
| 04 | 52 | 08 | 83 | 97 | 35 | 99 | 16 | 07 | 97 | 57 | 32 | 16 | 26 | 26 | 79 | 33 | 27 | 98 | 66 |
| 69 | 46 | 68 | 87 | 57 | 62 | 20 | 72 | 03 | 46 | 33 | 67 | 46 | 55 | 12 | 32 | 63 | 93 | 53 | 69 |
| 04 | 42 | 16 | 73 | 35 | 25 | 39 | 11 | 24 | 94 | 72 | 18 | 08 | 46 | 29 | 32 | 40 | 62 | 76 | 36 |
| 20 | 69 | 36 | 41 | 72 | 30 | 23 | 88 | 54 | 62 | 89 | 69 | 82 | 67 | 59 | 85 | 74 | 04 | 36 | 16 |
| 20 | 73 | 35 | 29 | 78 | 31 | 90 | 01 | 74 | 31 | 49 | 71 | 48 | 56 | 81 | 16 | 23 | 57 | 05 | 54 |
| 01 | 70 | 54 | 71 | 83 | 51 | 54 | 69 | 16 | 92 | 33 | 48 | 61 | 43 | 52 | 01 | 89 | 29 | 67 | 48 |

Šta kompjuter vidi

Slike visoke rezolucije

Multikolinearnost

- Jedno od obeležja predstavlja linearnu kombinaciju drugih obeležja:

$$c_1 f_1 + c_2 f_2 + \dots + c_k f_k = f_j$$

- Perfektna multikolinearnost
 - $X^T X$ nema rešenje
 - Ali se retko dešava. Obično je prisutan šum i varijable su samo snažno korelirane
 - Međutim, i ovo je problem

Snažna korelacija je problematična

- Obeležja

$$x_1 \text{ i } x_2, \text{ gde je } x_1 = x_2$$

- Stvarna (nepoznata) ciljna funkcija

$$y = x_1$$

- Dva alternativna modela:

$$\text{Model 1: } y = 0.5 \cdot x_1 + 0.5 \cdot x_2$$

$$\text{Model 2: } y = 1000 \cdot x_1 - 999 \cdot x_2$$

- Neka je za neku instancu $x_2 = 0.95 \cdot x_1$ (šum)

- Model 1: $y = 0.5 \cdot x_1 + 0.5 \cdot 0.95 \cdot x_1 = 0.975 \cdot x_1$

- Model 2: $y = 1000 \cdot x_1 - 999 \cdot 0.95 \cdot x_1 = 50.95 \cdot x_1$

Multikolinearnost nije retka u praksi

- Reči u tekstu se ne javljaju u potpunosti nezavisno jedna od druge
- Ekspresije mnogih gena su u velikoj korelaciji

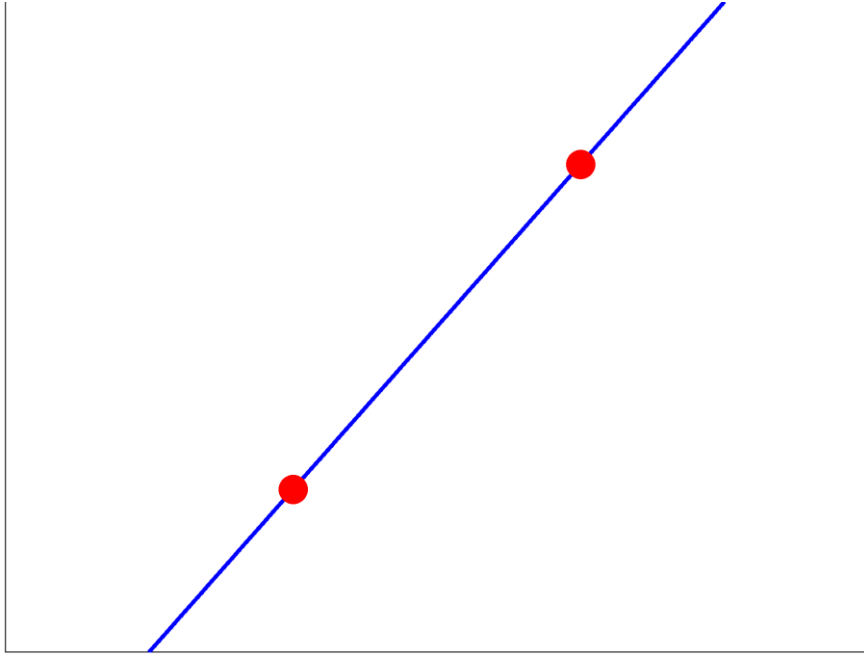
Ordinary Least Squares (OLS)

- OLS model:

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

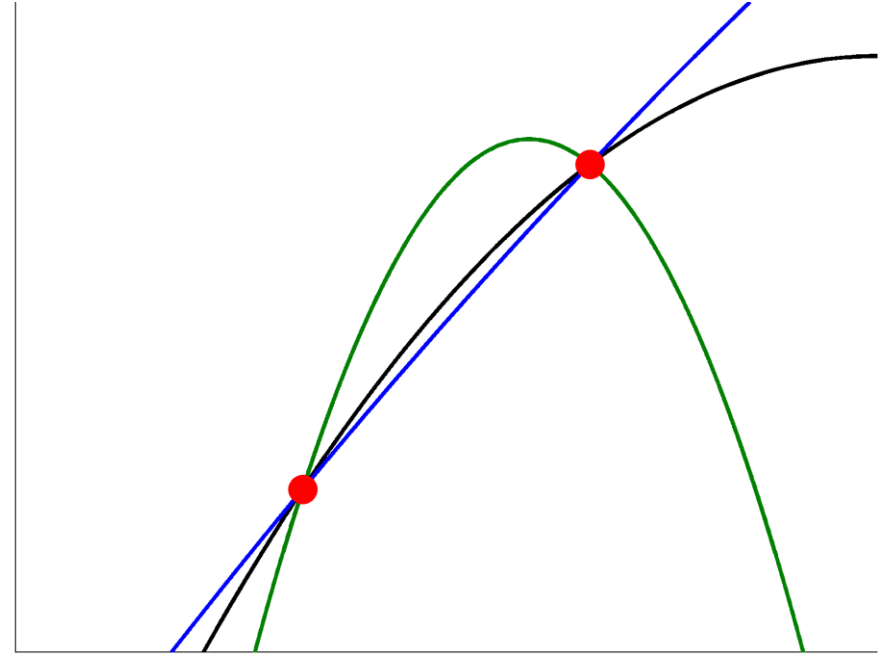
- $X^T X$ je singularna matrica znači da je sistem je neodređen
 - Imamo beskonačno mnogo rešenja i ne znamo koje da odaberemo

Neodređen sistem – primer



Određen sistem: jedno obeležje

- Linija koja se savršeno uklapa u podatke

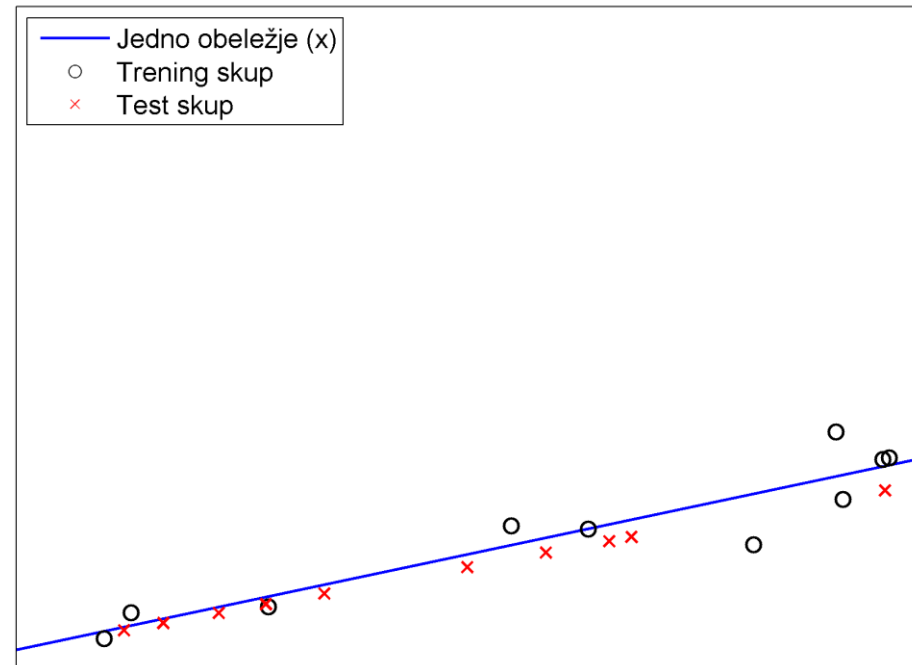
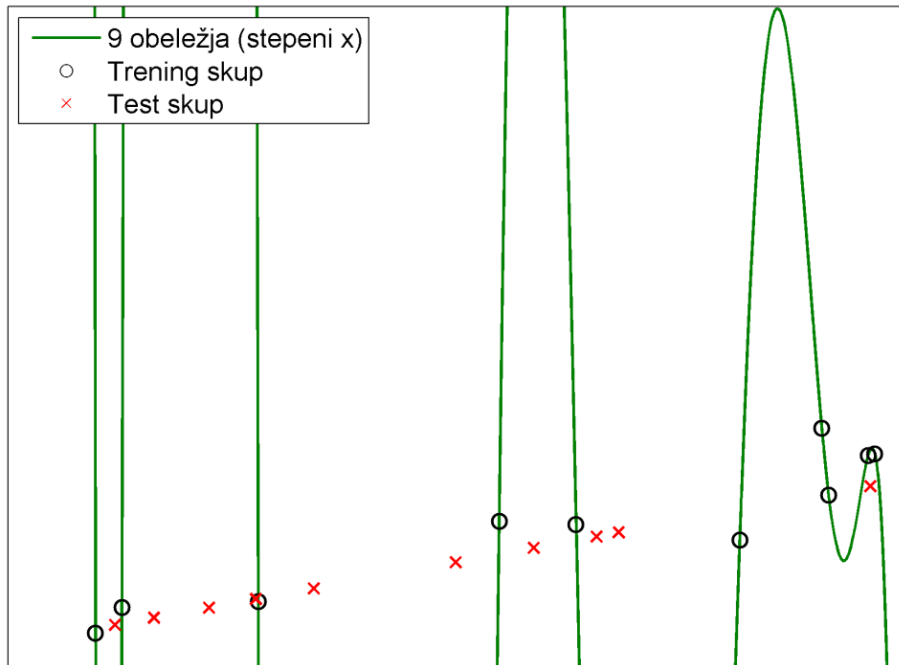


Neodređen sistem: više obeležja

- Svaki model se savršeno uklapa u podatke

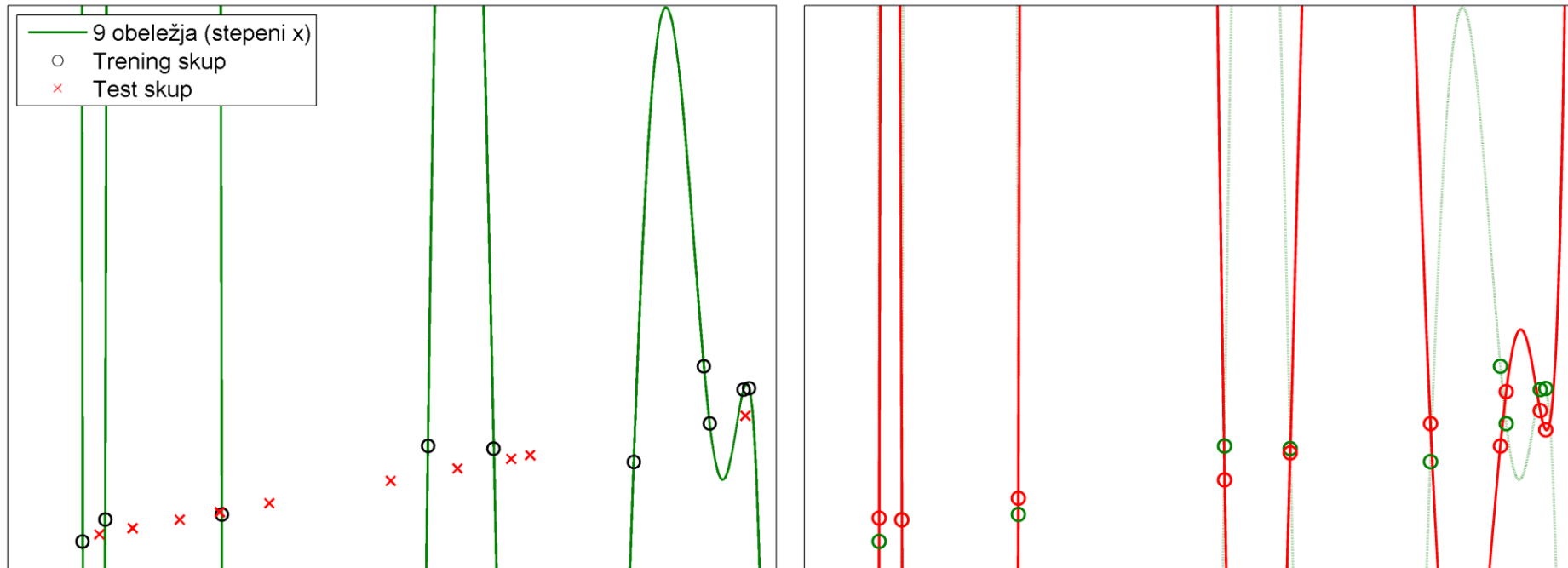
Mnogo varijabli – prilagođavanje

- Za dobru generalizaciju je ključno upravljanje prilagodljivošću modela
- Dobra prilagođenost modela trening podacima ne obezbeđuje dobru generalizaciju



Simptom peprilagođavanja: velika varijansa

Velika varijansa: male promene trening skupa rezultuju veoma različitim modelima

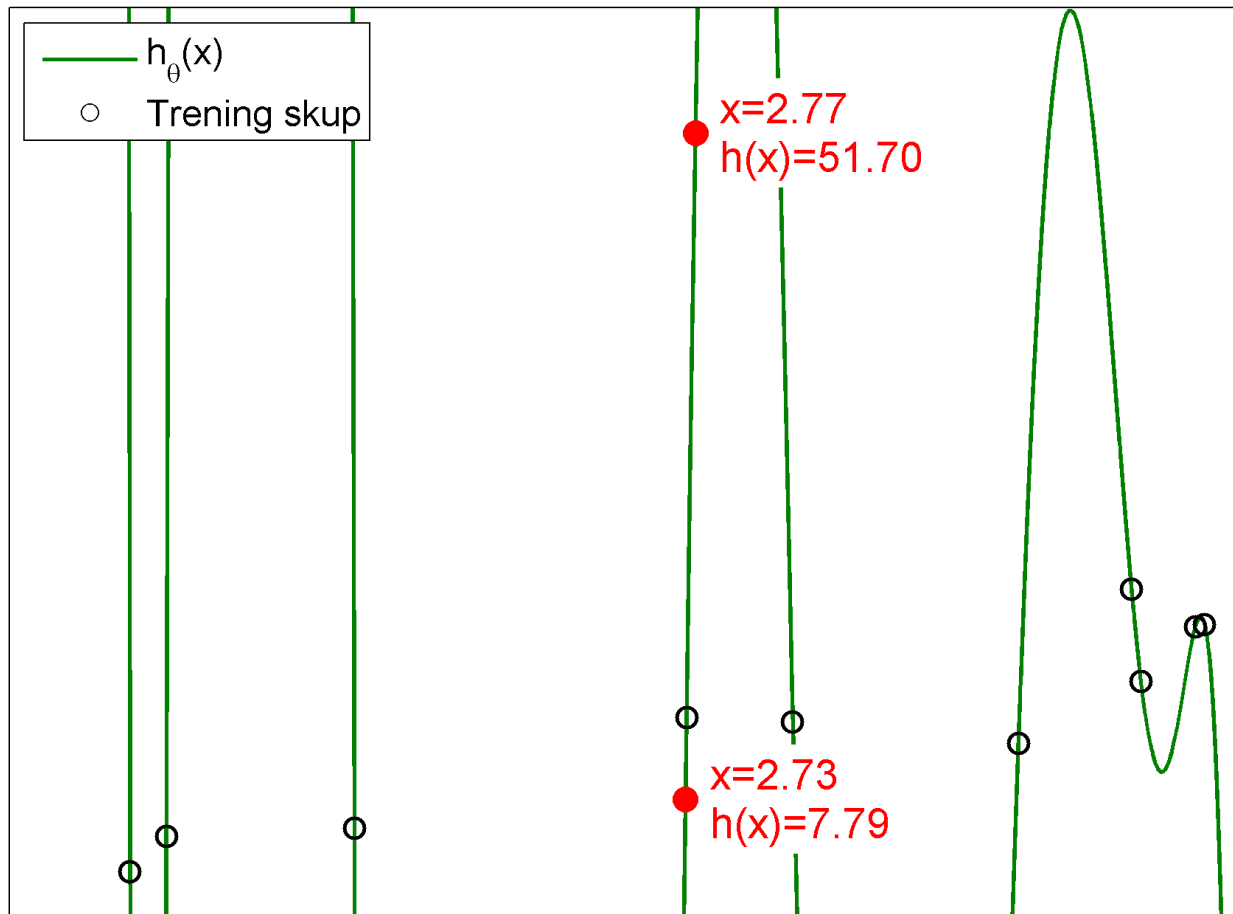


U oba slučaja, y vrednosti su dobijene tako što je dodat šum na stvarnu vrednost y .

U oba slučaja je šum jednake jačine

Simptom peprilagođavanja: velika varijansa

U slučaju velikih vrednosti θ , male promene vrednosti x dovode do velikih promena vrednosti $h_{\theta}(x)$



Kako rešiti preprilagođavanje?

Uvećati N

Nije uvek
moguće

Smanjiti D

- Možda ćemo izbaciti obeležja bitna za rešenje problema
- Kompleksnost modela treba da odgovara kompleksnosti problema, a ne N

Regularizacija

Regularizacija

- Sačuvati sva obeležja, ali smanjiti magnitude vrednosti θ
 - *shrinkage methods*
- Ovo će rezultovati „jednostavnijom“ hipotezom koja je manje podložna prilagođavanju
- Radi dobro u slučaju kada imamo mnogo obeležja takvih da svako doprinosi (u manjoj meri) predikciji y

Regularizacija

- U slučaju linearnih modela

$$h_{\theta}(x) = \sum_{d=1}^D \theta_d x_d$$

važi:

$$\theta = \nabla_x h_{\theta}(x)$$

Efekat regularizacije:

ograničavamo gradijent funkcije $h_{\theta}(x)$ (brzinu promene funkcije $h_{\theta}(x)$ sa promenom x)

Regularizacija

- U opštijem smislu, regularizacijom se naziva bilo koja modifikacija optimizacionog problema koja ograničava prilagodljivost modela i čini ga manje podložnim preprilagođavanju
- U još opštijem smislu, regularizacija je bilo kakva modifikacija matematičkog problema koja ga čini bolje uslovljenim