Praktični saveti za primenu mašinskog učenja

1. Dijagnostika (debagovanje) obučavajućeg algoritma
2. Analiza grešaka modela
3. Kako započeti rad na problemu mašinskog učenja (izbeći preuranjenu optimizaciju)?

Problem

Radimo na problemu predikcije da li se radi o *ham* ili *spam* emailu

Implementirali smo logističku regresiju

Performanse nisu dovoljno dobre

- Da uvećamo broj opservacija trening skupa?
- Da povećamo skup obeležja (nezavisne varijable ili polinomijalna obeležja)?
- Da smanjimo skup obeležja?
- Da li da promenimo obeležja?
- Optimizacija
  - Da li da pustimo GD da radi veći broj iteracija?
  - Da li da zamenimo GD drugim optimizacionim metodom (npr. Njutnov metod)?
- Da li da uvećamo ili smanjimo λ?
- Da li da promenimo obučavajući algoritam?

Dijagnostika obučavajućeg algoritma

- Slepo isprobavanje može da upali, ali zahteva mnogo vremena i često je pitanje sreće

- Bolji pristup:
  - Izvršiti dijagnostiku da utvrdimo šta je zapravo problem
  - Popraviti taj problem

Dijagnostika obučavajućeg algoritma

- *Machine Learning diagnostics* predstavlja skup tehnika koje možemo primeniti u cilju dobijanja uvida u to šta vezano za obučavajući algoritam radi ili ne radi

- ML dijagnostika zahteva određeno vreme za razumevanje/implementaciju

- Ali ovo je veoma dobro iskorišćeno vreme

Podsetnik: selekcija modela

- Bez regularizacije:

$$J(\theta) = \frac{1}{2N}\left[\sum_{i=1}^{N}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right],$$

$$h_\theta(x) = \sum_{j=0}^{D}\theta_j x^j$$

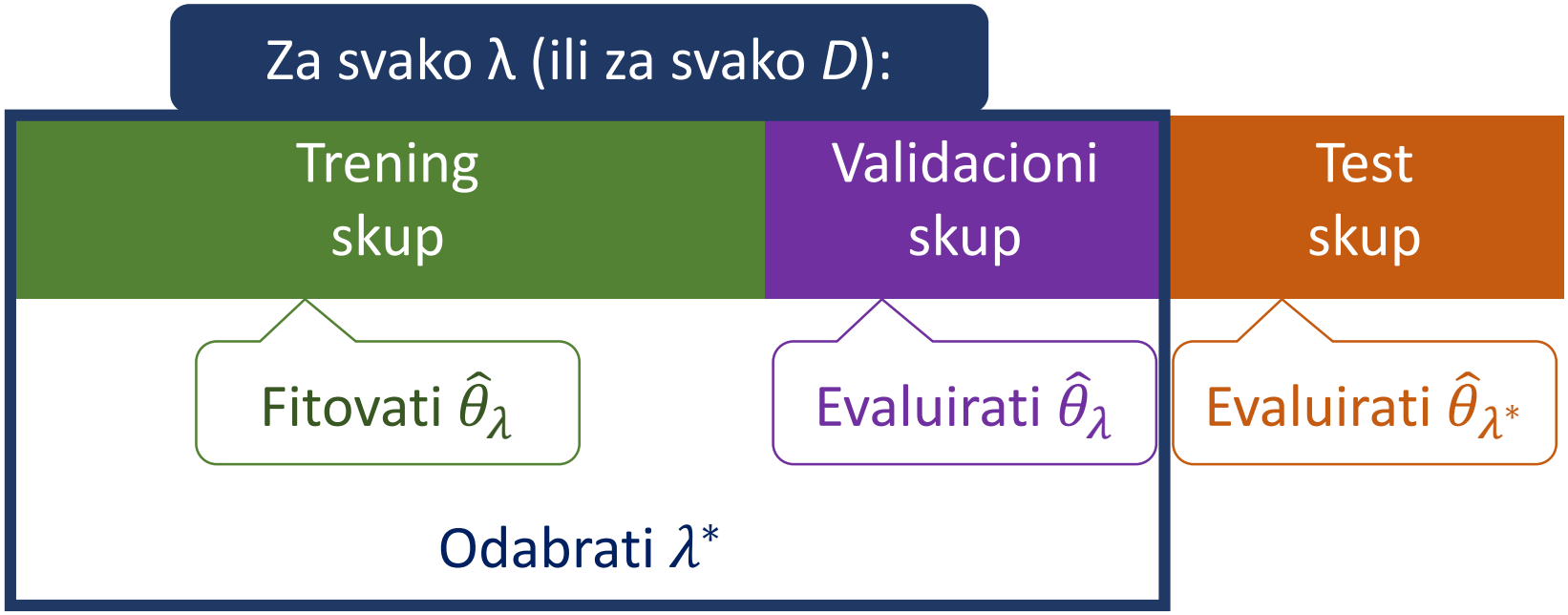
Treba da odaberemo D

- Sa regularizacijom:

Treba da odaberemo λ

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

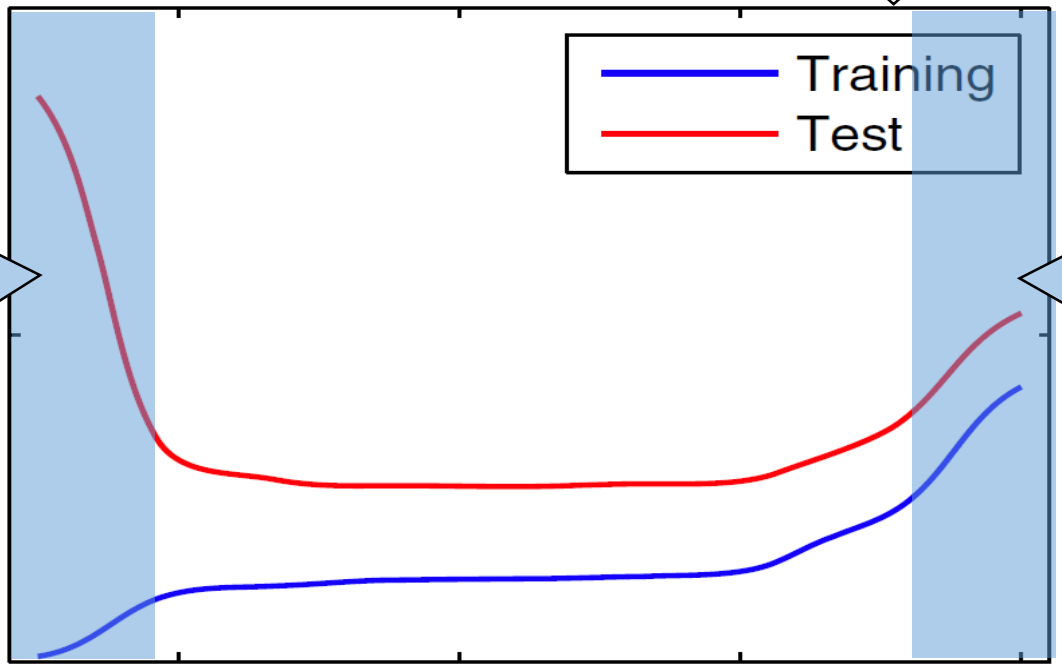
Podsetnik: selekcija modela



Regularizacija – odabir λ

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

Zgodno je iscrtati ovakav grafik za odabir λ

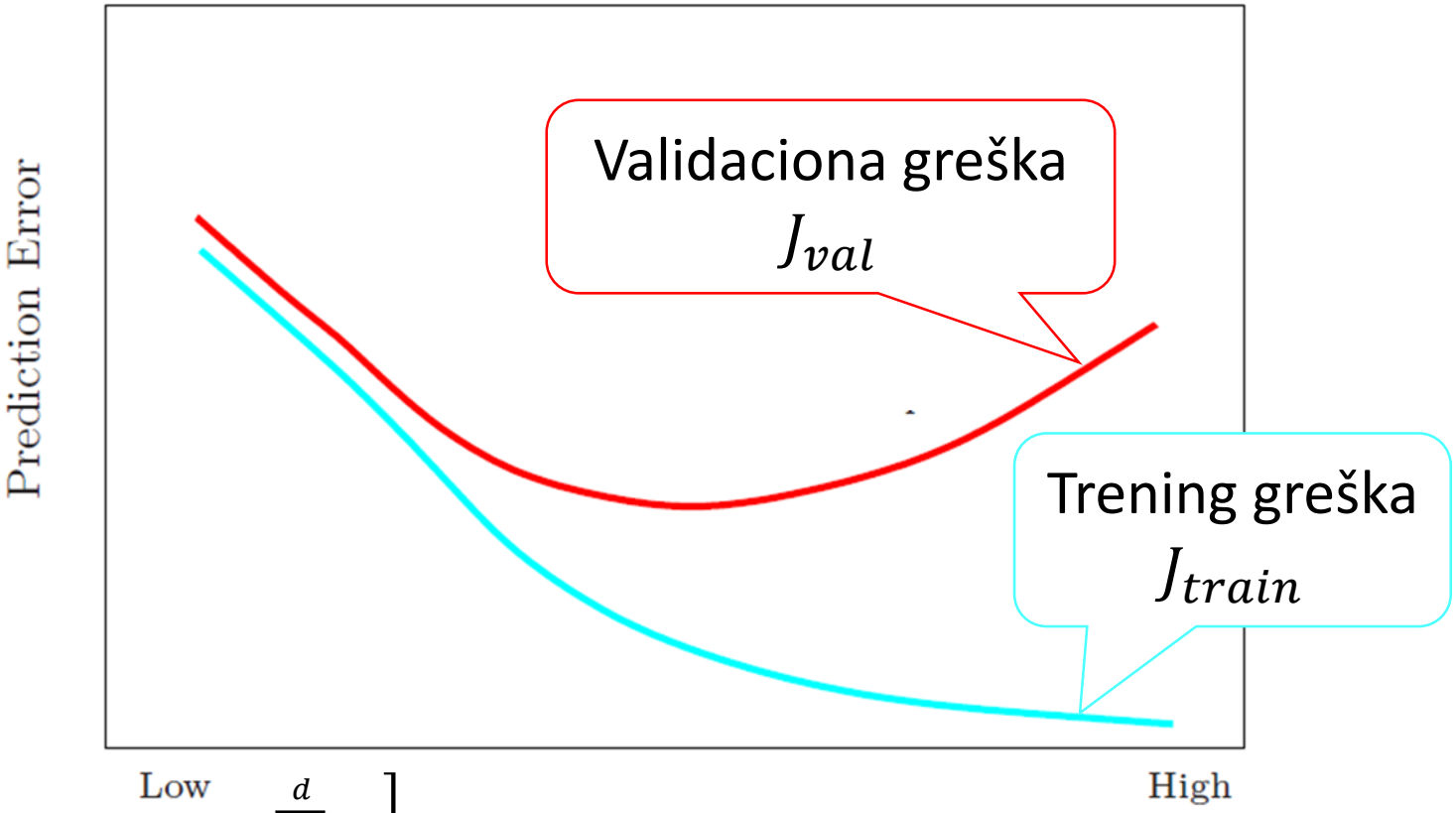


Premalo λ → velika varijansa

Preveliko λ → veliko sistematsko odstupanje

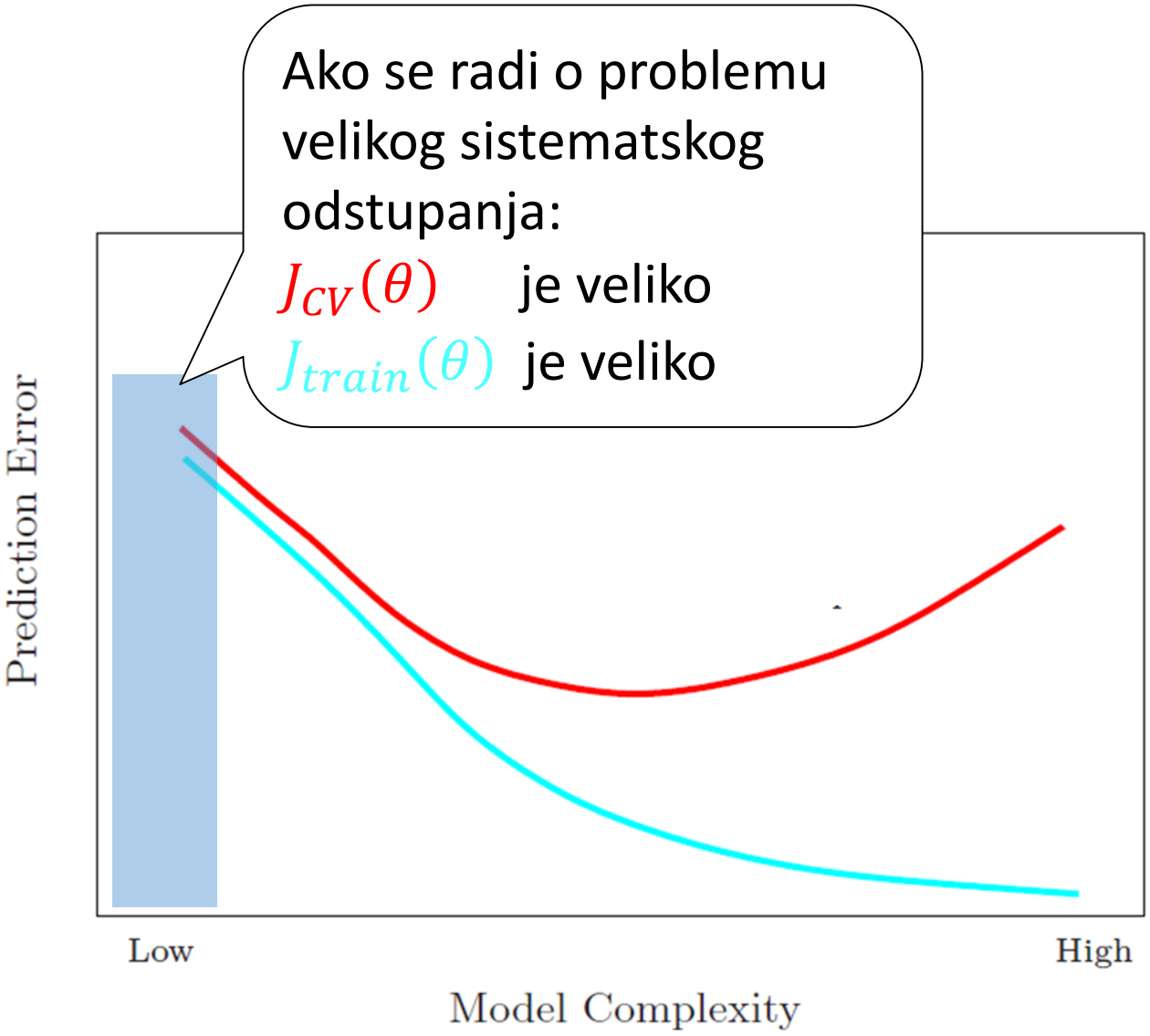
Problem sistematskog odstupanja ili varijanse?

Ako je greška modela na **validacionom** skupu velika, patimo ili od velikog sistematskog odstupanja ili od velike varijanse

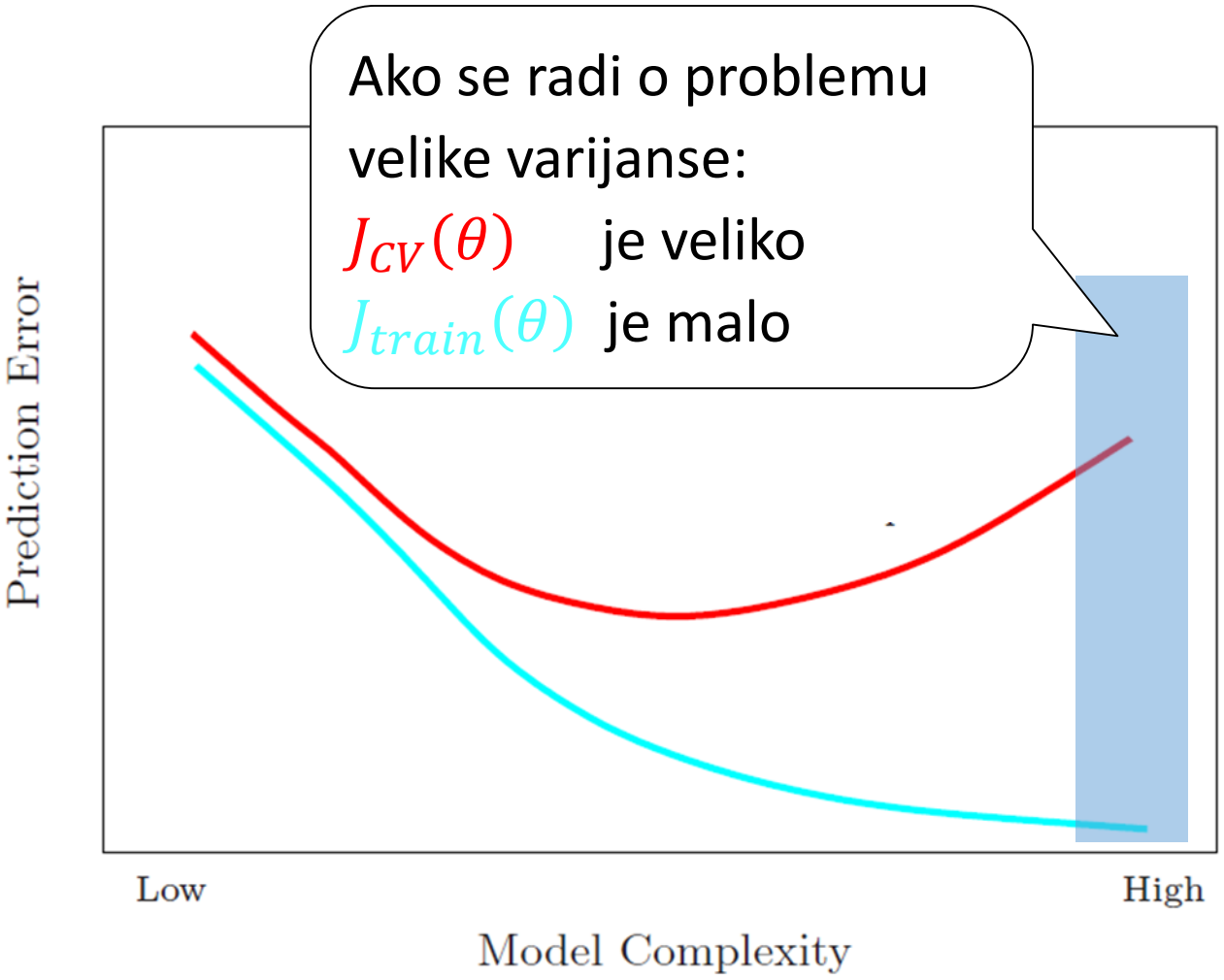


$$J(\theta) = \frac{1}{2N}\left[\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{d}\theta_j^2\right]$$

Problem sistematskog odstupanja ili varijanse?



Problem sistematskog odstupanja ili varijanse?

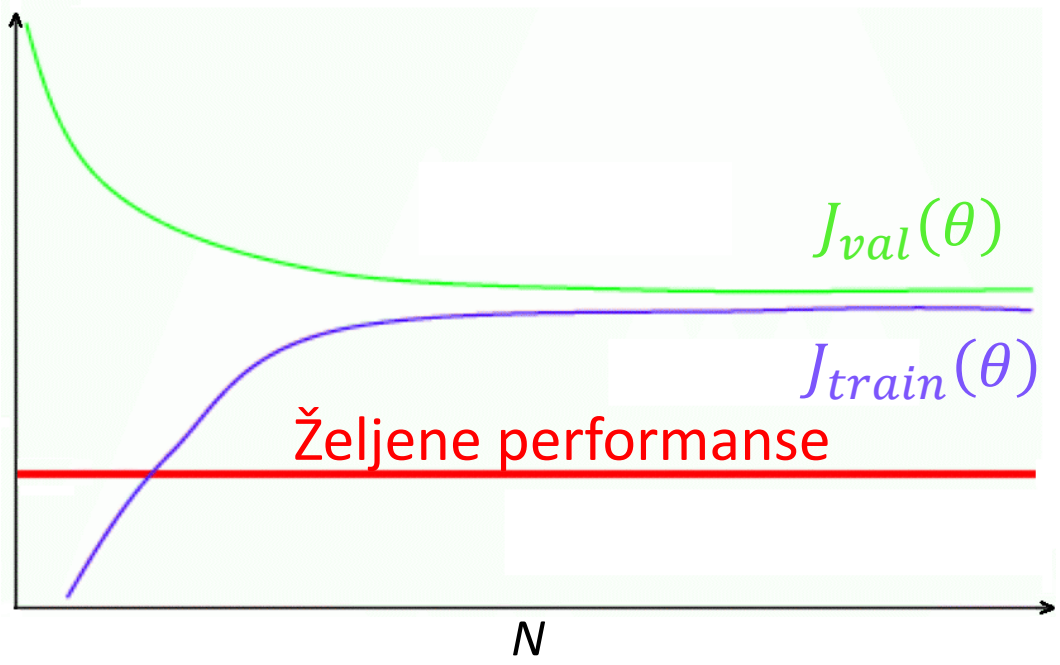


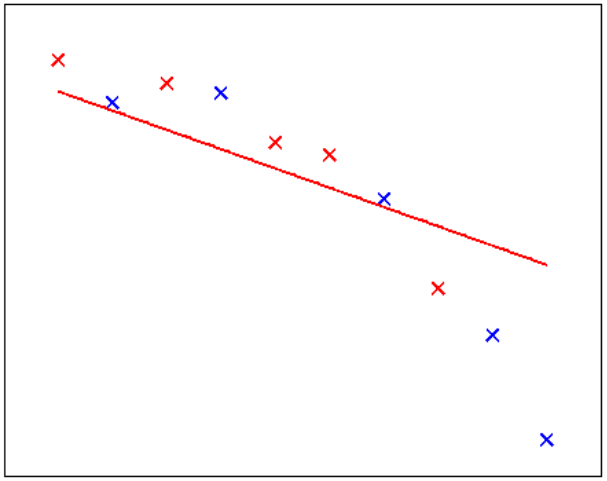
ML dijagnostika: *learning curves*

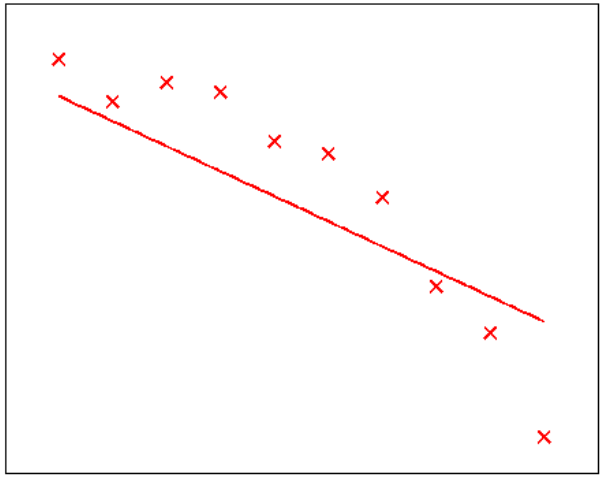
- Ovaj metod nam pomaže da utvrdimo da li naš model pati od velikog sistematskog odstupanja ili velike varijanse
- Kompleksnost modela je fiksirana
- Iscrtavaćemo $J_{train}(\theta)$ i $J_{val}(\theta)$ kao funkciju veličine trening skupa *N*
- Veštački ćemo smanjivati količinu dostupnih podataka za obučavanje i iscrtavati ove dve greške da vidimo šta bi se dešavalo da imamo manji trening skup

Learning curves

Veliko sistematsko odstupanje



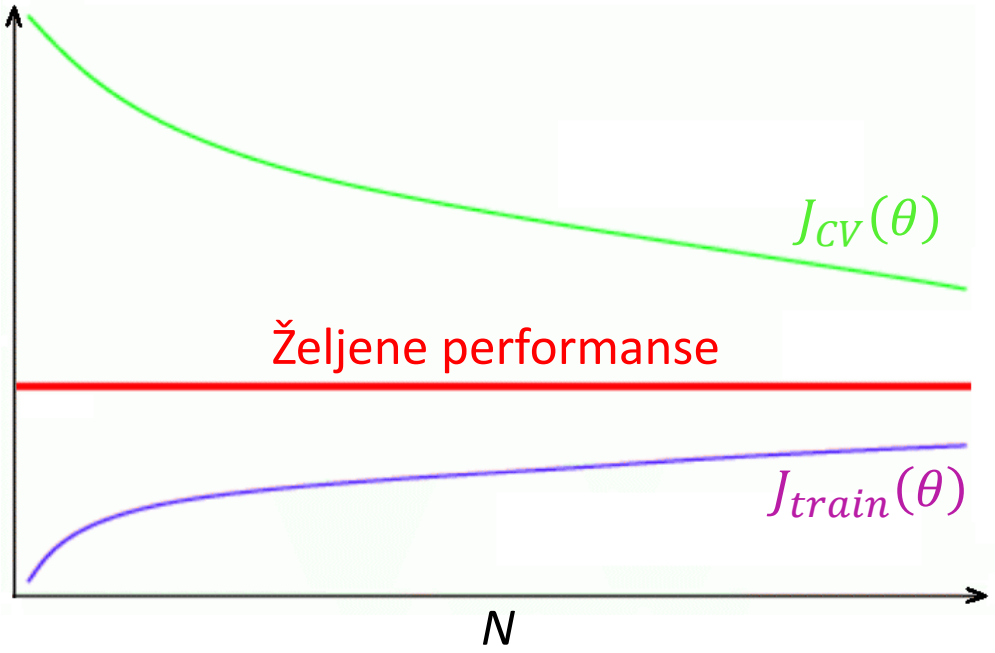


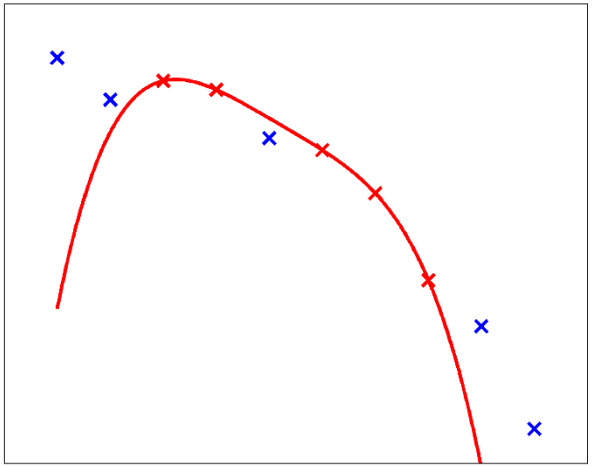


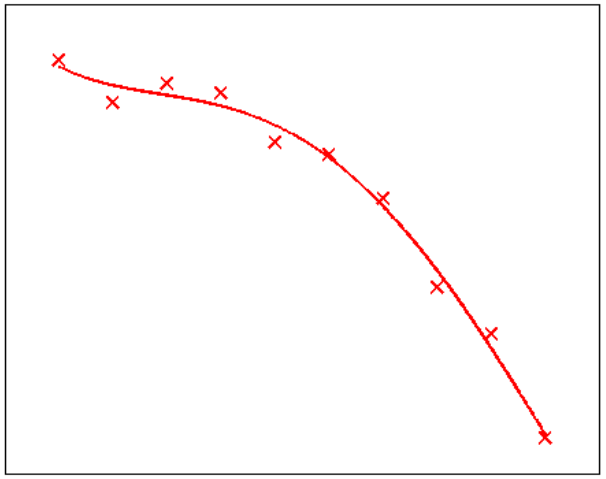
- Uočavamo da se trening greška brzo izravnava
- Greška na validacionom skupu bliska je grešci na trening skupu i obe su velike
- Kod modela koji ima veliko sistematsko odstupanje, dodavanje novih primera neće pomoći

Learning curves

Velika varijansa







- Uočavamo da trening greška još raste, a greška na validacionom skupu još opada
- Veliki raskorak greške na validacionom i trening skupu
- Kod modela koji ima veliku varijansu, dodavanje novih primera <u>će verovatno pomoći</u>

Learning curves – zaključak

- Ispravka velike varijanse:
  - Uvećati trening skup
  - Smanjiti skup obeležja
  - Regularizacija

- Ispravka velikog sistematskog odstupanja:
  - Povećati skup obeležja (dodati nove nezavisne varijable ili polinomijalna obeležja)
  - Odabrati fleksibilniji model