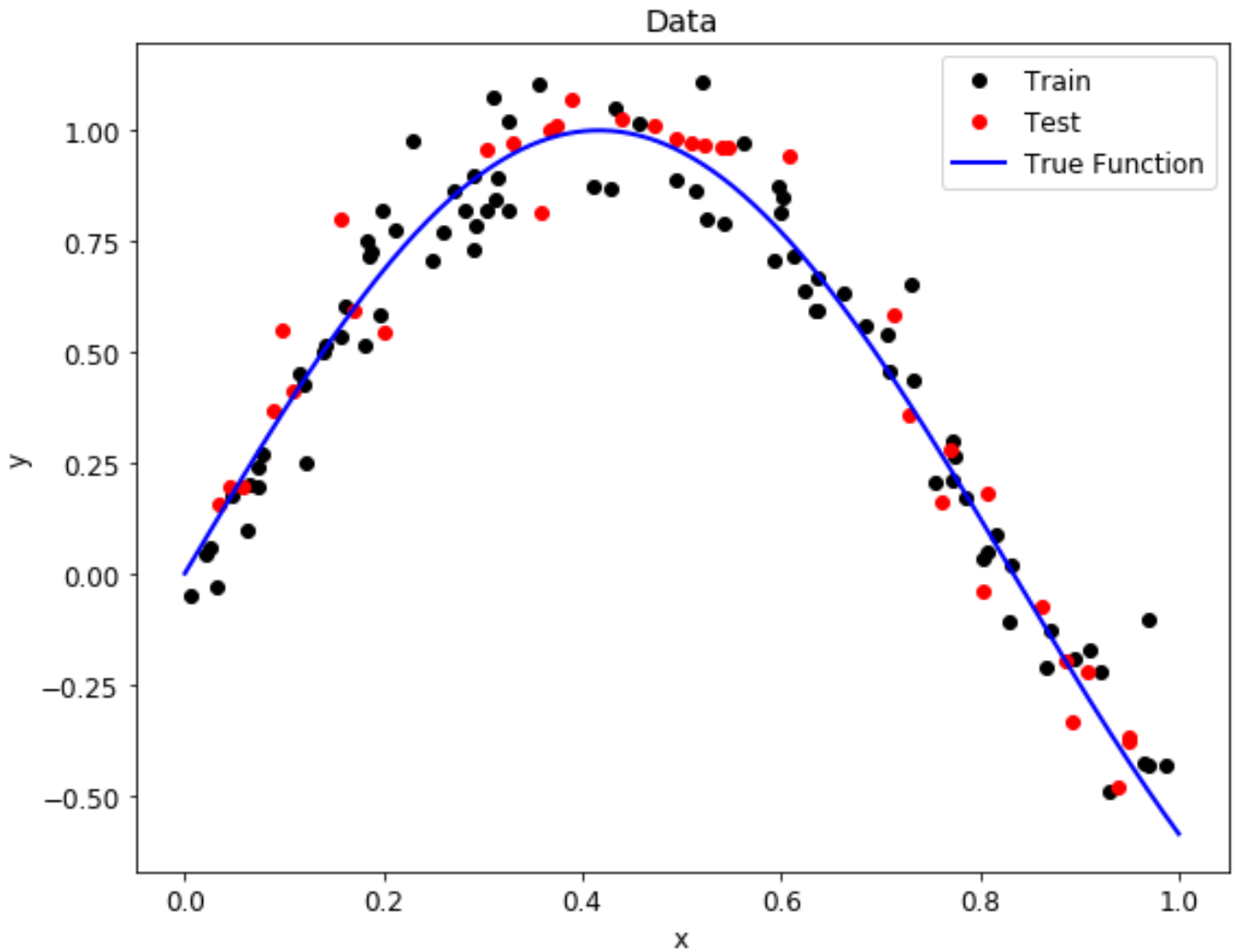
Izbor modela i skupa obeležja

Primenjeni algoritmi

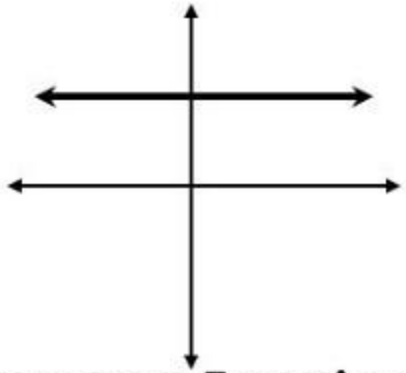
Izbor modela

- Neka je model predstavljen u obliku
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_k x^k$$
- Potrebno je odrediti $k \in \{0,1,2, \dots, 10\}$
- Zadatak je odabrati model iz skupa modela $\mathcal{M} = \{M_1, M_2, \dots, M_d\}$

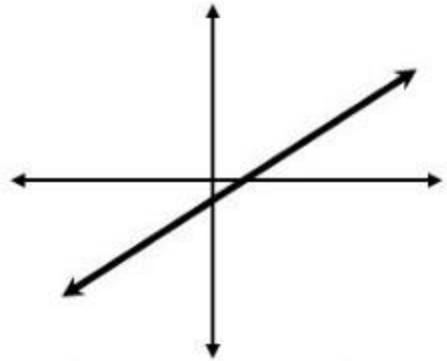
Primer 1


Data
Train
Test
True Function
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
y
0.0
0.2
0.4
0.6
0.8
1.0
x

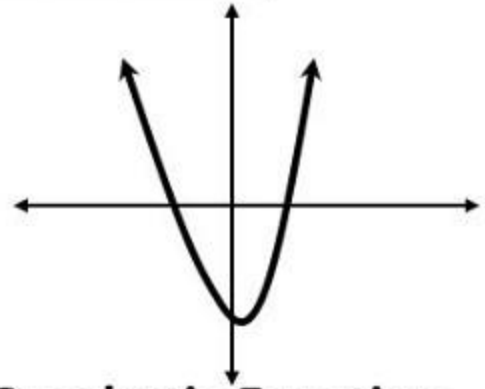
Graphs of Polynomial Functions:



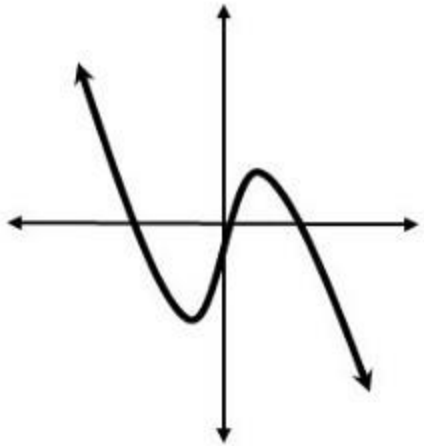
Constant Function
(degree = 0)



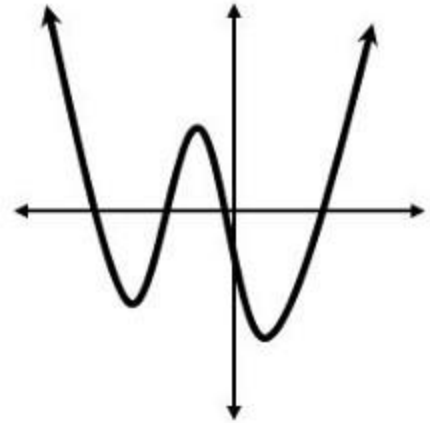
Linear Function
(degree = 1)



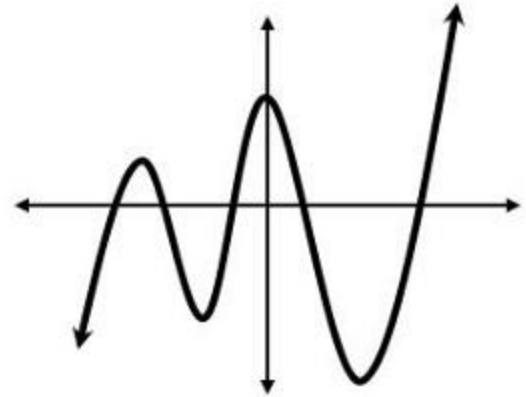
Quadratic Function
(degree = 2)



Cubic Function
(deg. = 3)

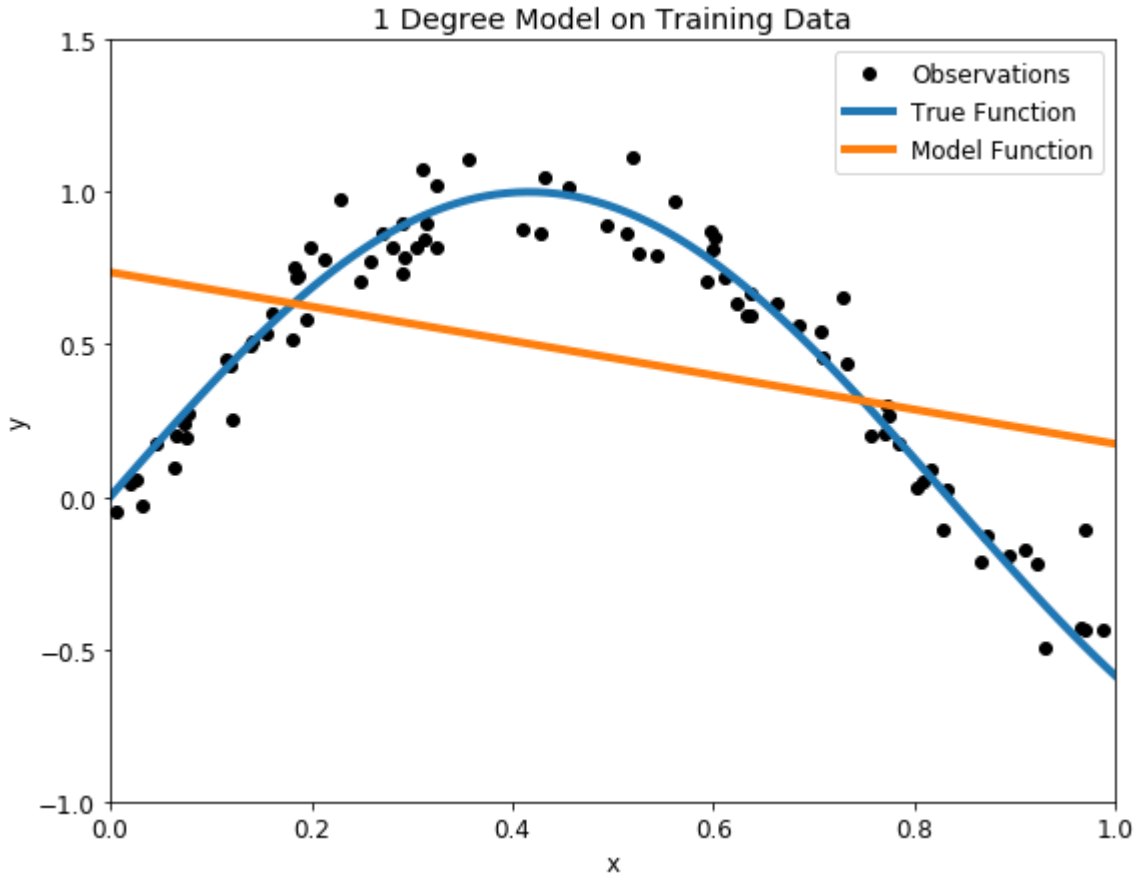


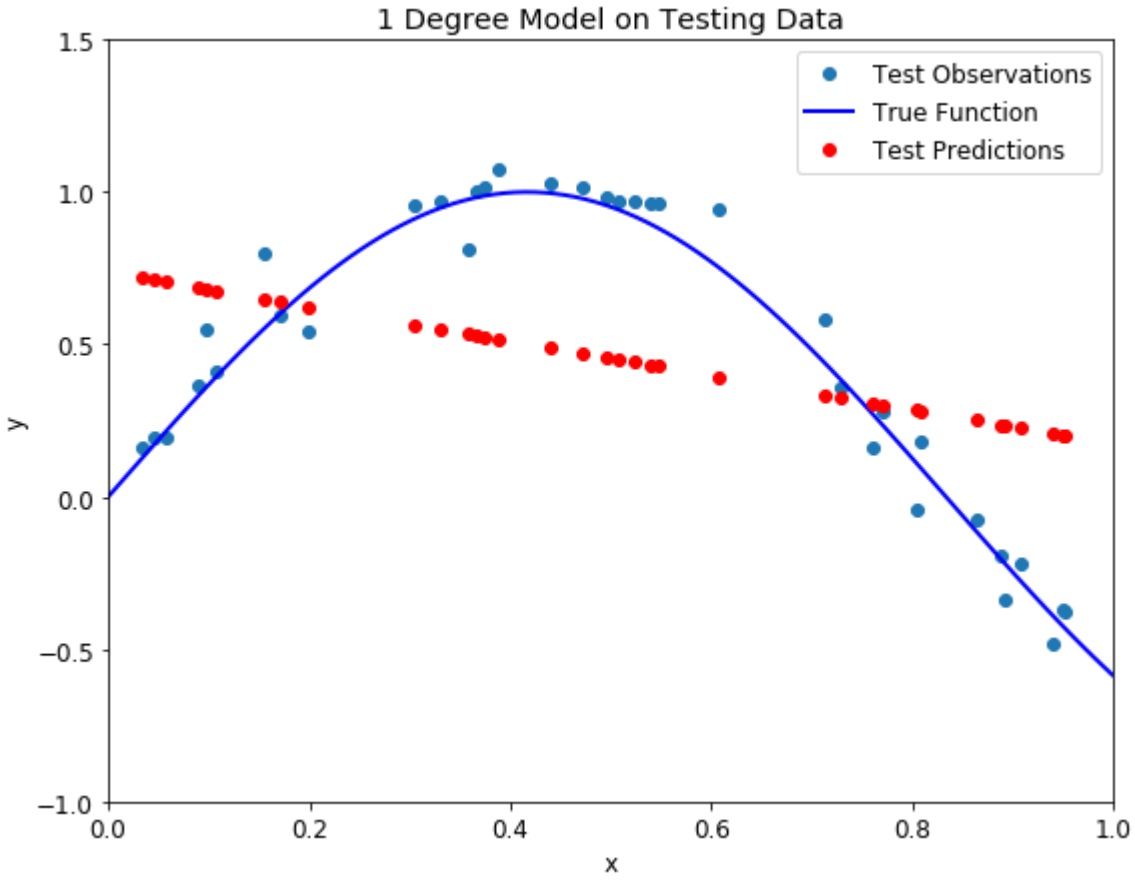
Quartic Function
(deg. = 4)



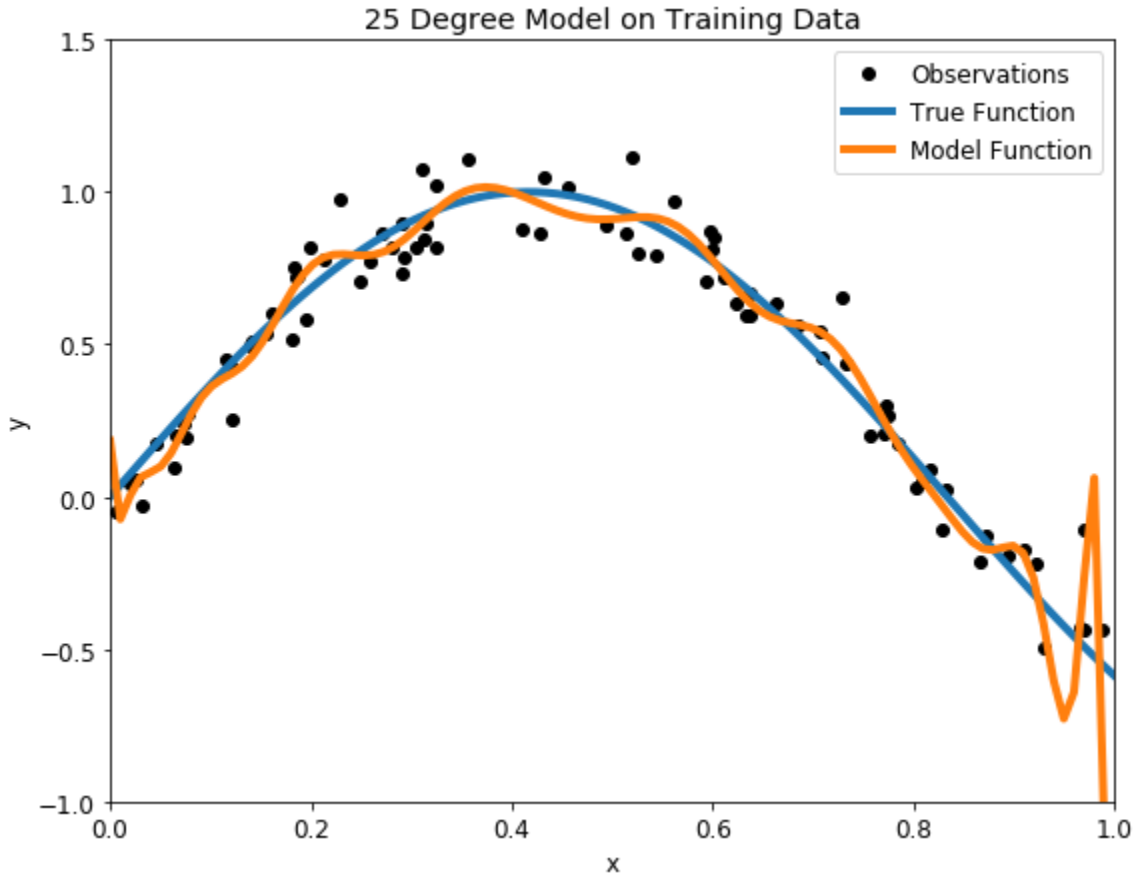
Quintic Function
(deg. = 5)

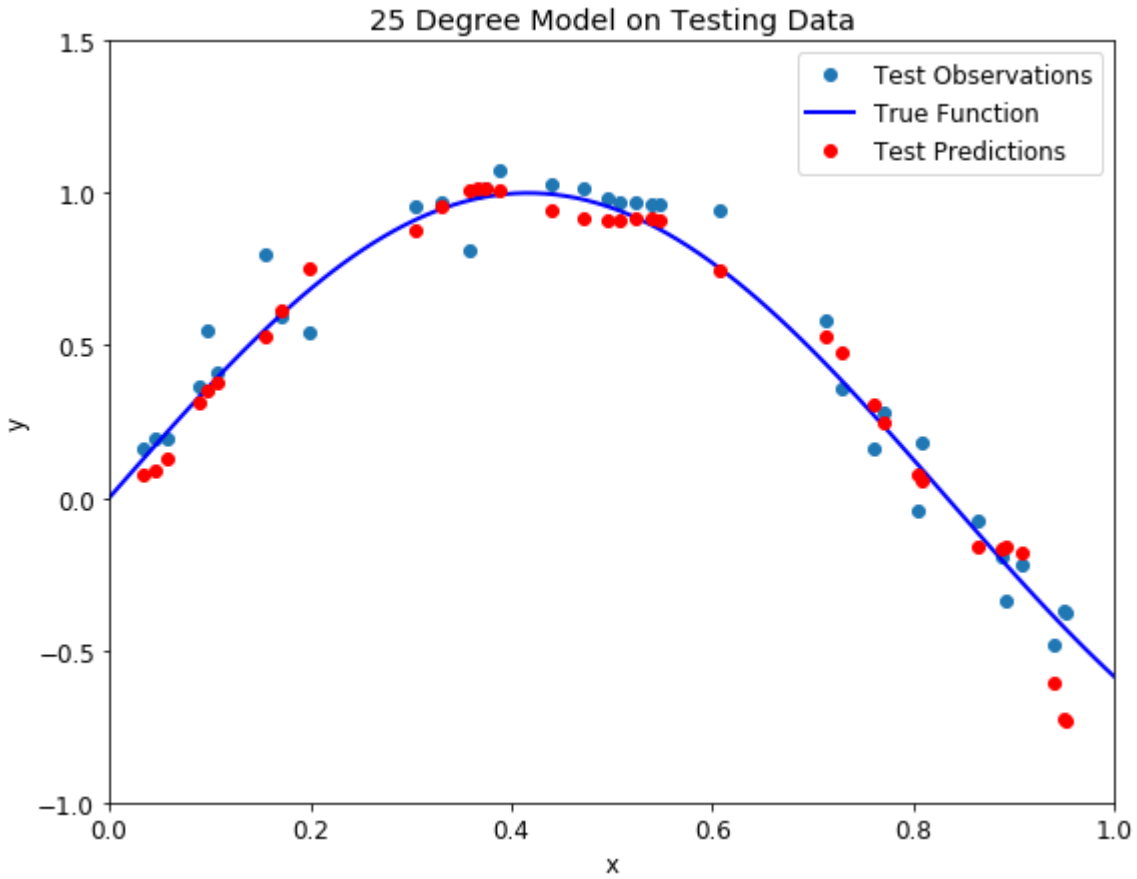
Modelovanje polinomom 1. reda



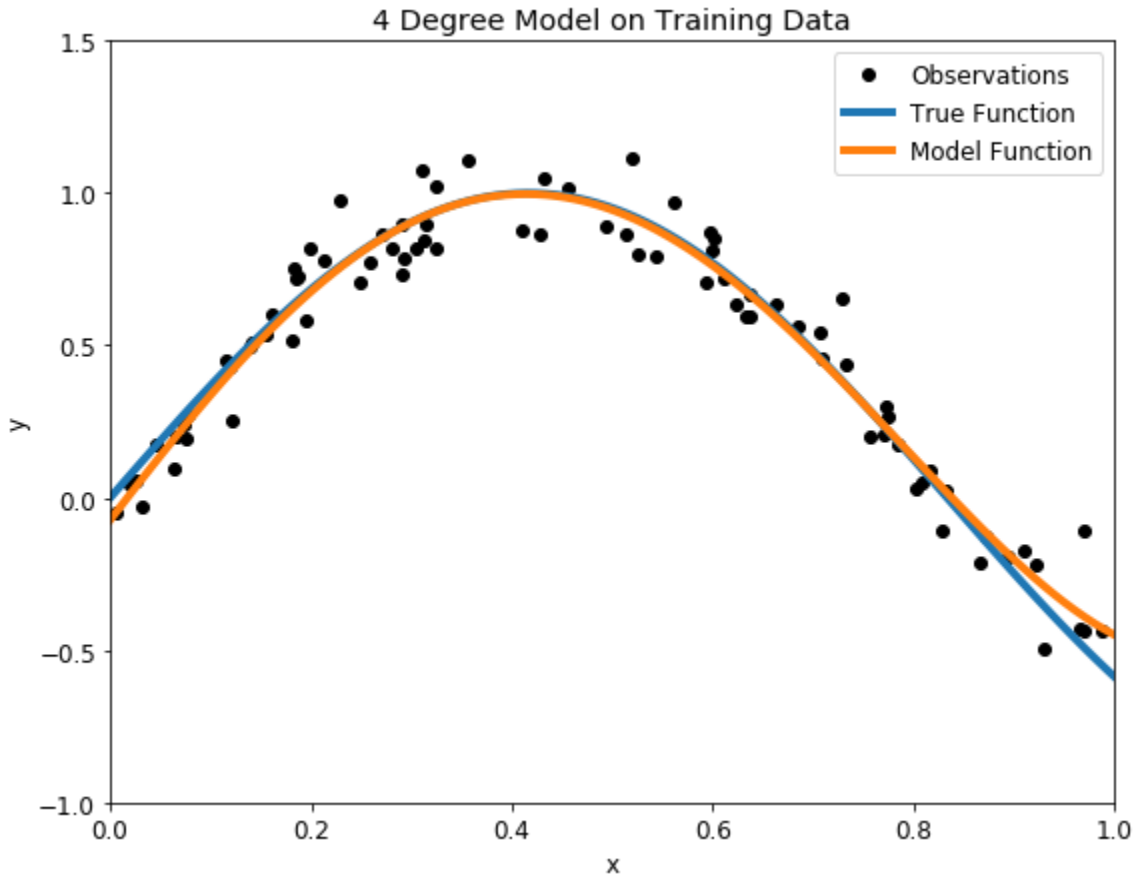


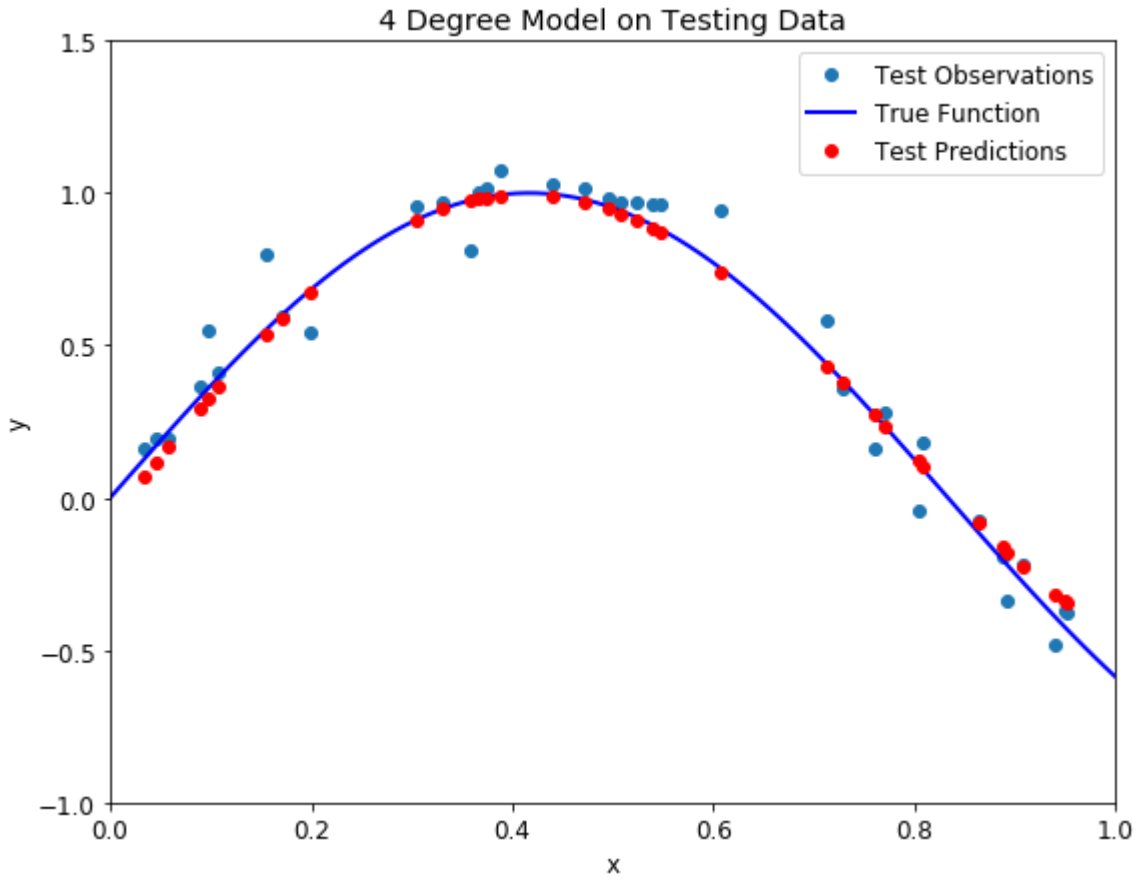
Modelovanje polinomom 25. reda



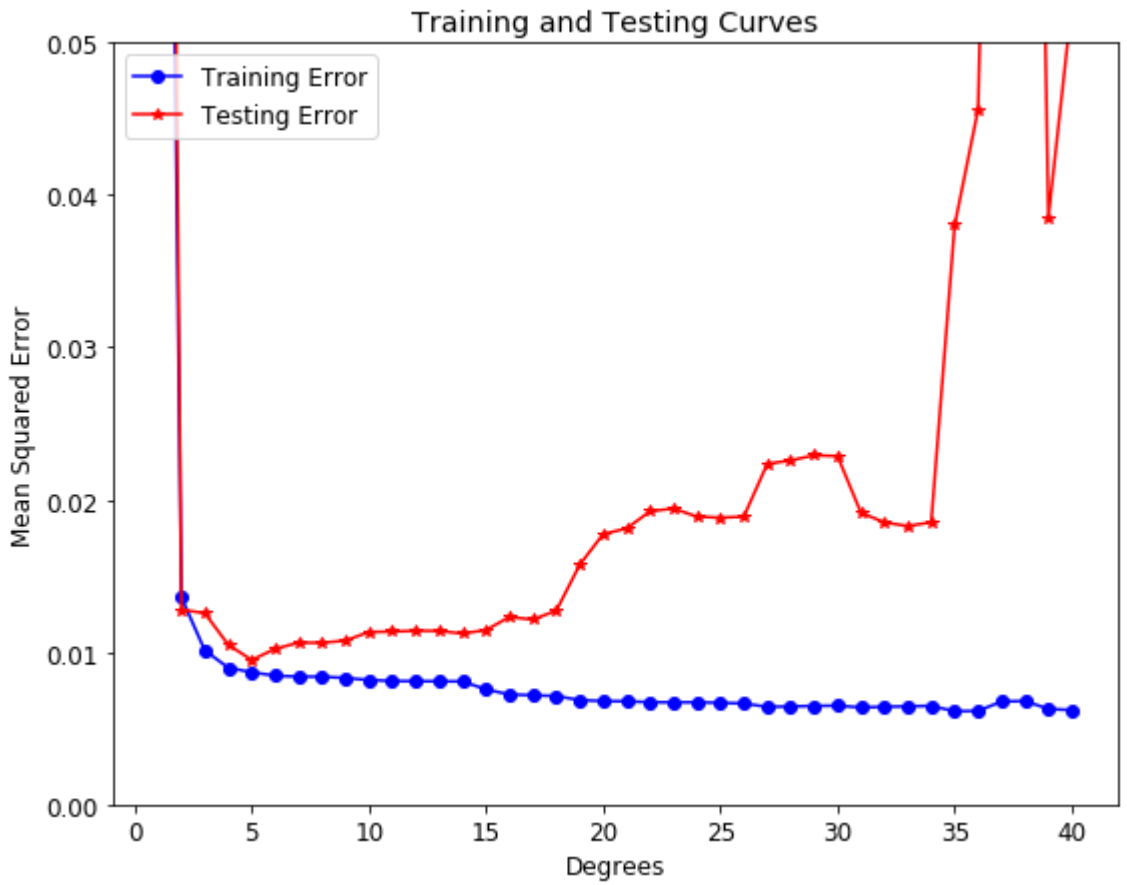


Modelovanje polinomom 4 reda





Analiza greške za različite modele



Unakrsna validacija (*Cross validation*)

- Trenirati svaki model M_i na obučavajućem skupu S – dobiju se hipoteze h_i
- Izabrati hipotezu sa najmanjom greškom
  - Ovo ne radi. Ako se izabere polinom velikog reda on će bolje fitovati podatke iz obučavajućeg skupa S i dati manju obučavajuću grešku. Ali nije dobra zato što daje veliku varijansu – kod pojave novih podataka – veća greška.
- Algoritmi validacije:
  - Jednostavna unakrsna validacija
  - K-tostruka validacija
  - Validacija jednostruke eliminacije

Jednostavna unakrsna validacija

1. Na slučajan način se podeli skup *S* na S_{train} (npr. 70% podataka) – obučavajući skup i S_{cv} (preostalih 30%) – validacioni skup
2. Trenira se svaki model na skupu S_{train} i dobijaju hipoteze h_i
3. Bira se hipoteza h_i sa najmanjom greškom $\varepsilon_{S_{CV}}(h_i)$ na validacionom skupu S_{cv}

- Nedostatak: „Gubitak“ oko 30% podataka

K-tostruka unakrsna validacija

1. Slučajno se podeli skup S na k disjunktnih podskupova sa m/k primera u svakom: $S_1, S_2, \dots, S_k$

2. Svaki model M_i se određuje iz:

   For j = 1,...,k

   trenirati model M_i na $S_1 \cup \dots \cup S_{j-1} \cup S_{j+1} \cup \dots S_k$ (sve sem S_j)

   i dobiti hipotezu $\hat{h}_{ij}$

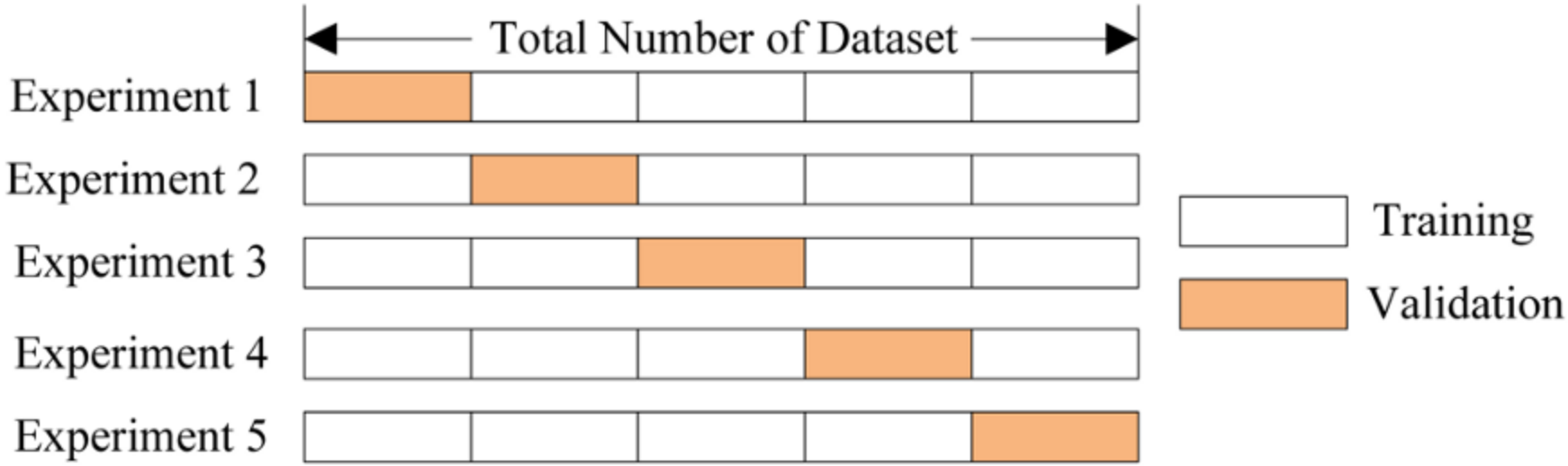
   testirati hipotezu $\hat{h}_{ij}$ na $S_j \longrightarrow \hat{\varepsilon}_{S_j}(\hat{h}_{ij})$

   Odrediti grešku modela kao $e_i = \frac{1}{k}\sum_{j=1}^{k} \hat{\varepsilon}_{S_j}(\hat{h}_{ij})$

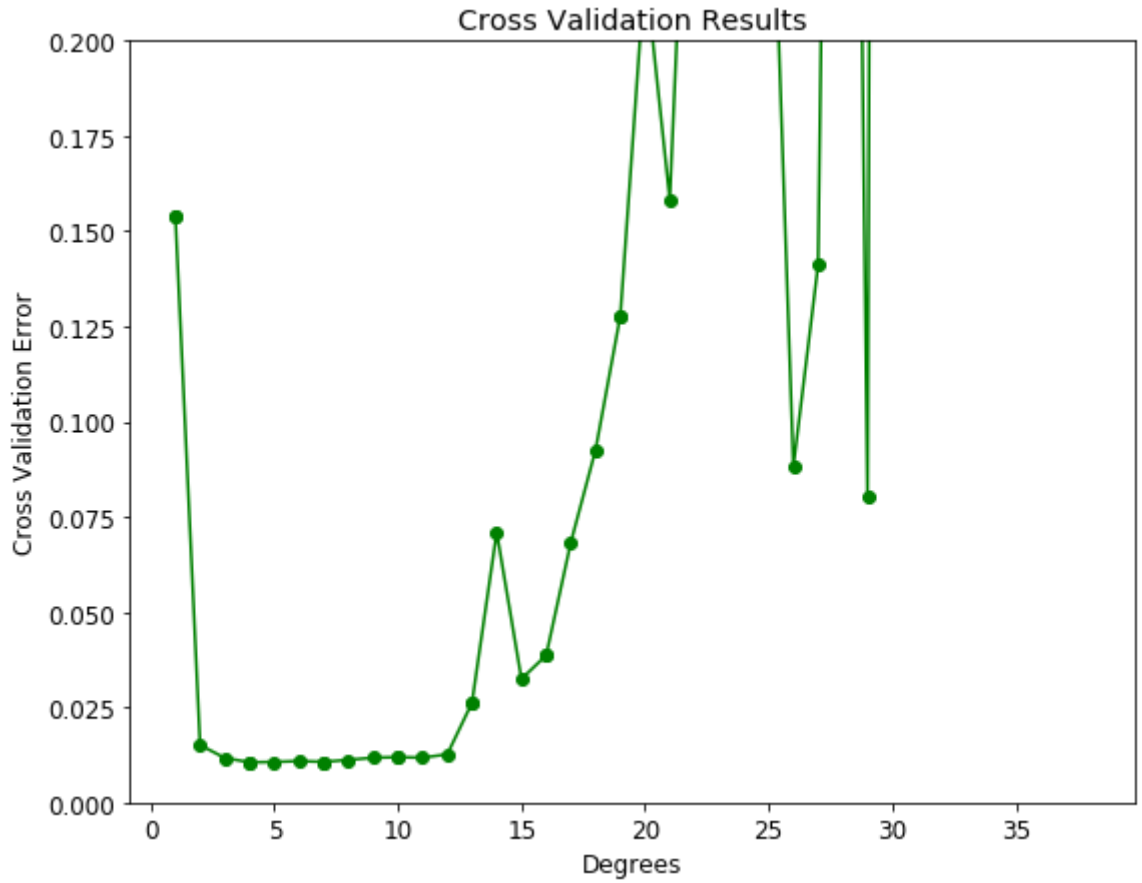
3. Bira se model M_i sa najmanjom greškom

- Tipičan izbor za k=10

5 –tostruka unakrsna validacija



	degrees	cross_valid
0	4	0.010549
1	5	0.010637
2	7	0.010665
3	6	0.010887
4	8	0.011182
5	3	0.011695
6	9	0.011757
7	11	0.011769
8	10	0.011902
9	12	0.012642



Validacija jednostruke eliminacije

- Ako je broj primera jako mali uzima se k=m
  - Svaki model se obučava na svakom podskupu m-1
  - Testira se na jednom (izostavljenom) primeru
  - Uzima se prosek

Izbor svojstava

- Ako je broj svojstava *d* veoma velik $n \gg m$ samo je mali skup svojstava relevantan
- Zadatak: izabrati poskup „značajnih“ svojstava
- Postoji ukupno 2^m poskupova – obimna pretraga
- Algoritmi:
  - Algoritmi omotača
    - Pretraga unapred – dodavanje svojstava
    - Pretraga unazad – uklanjanje svojstava
  - Rangirajući algoritmi
    - Uzajamna informacija svojstva i izlaza - korelacija

Algortimi omotača – pretraga unapred

1. Polazi se od praznog skupa svojstava $\mathcal{X} = \emptyset$
2. Odlika $x_j \notin \mathcal{X}$ sa najmanjom greškom se dodaje u skup $\mathcal{X} = \mathcal{X} \cup x_j$
3. Ako postoji još neizabranih odlika – > korak 2
4. U suprotnom vrati podskup $\mathcal{X}$, kraj

Algortimi omotača – pretraga unazad

1. Polazi se od skupa svih svojstava $\mathcal{X}$
2. Eliminiše se u svakoj iteraciji jedno svojstvo x_j sa najvećom greškom $\mathcal{X} = \mathcal{X} \setminus x_j$
3. Ako postoji još neizabranih odlika – > korak 2
4. U suprotnom vrati podskup $\mathcal{X}$, kraj