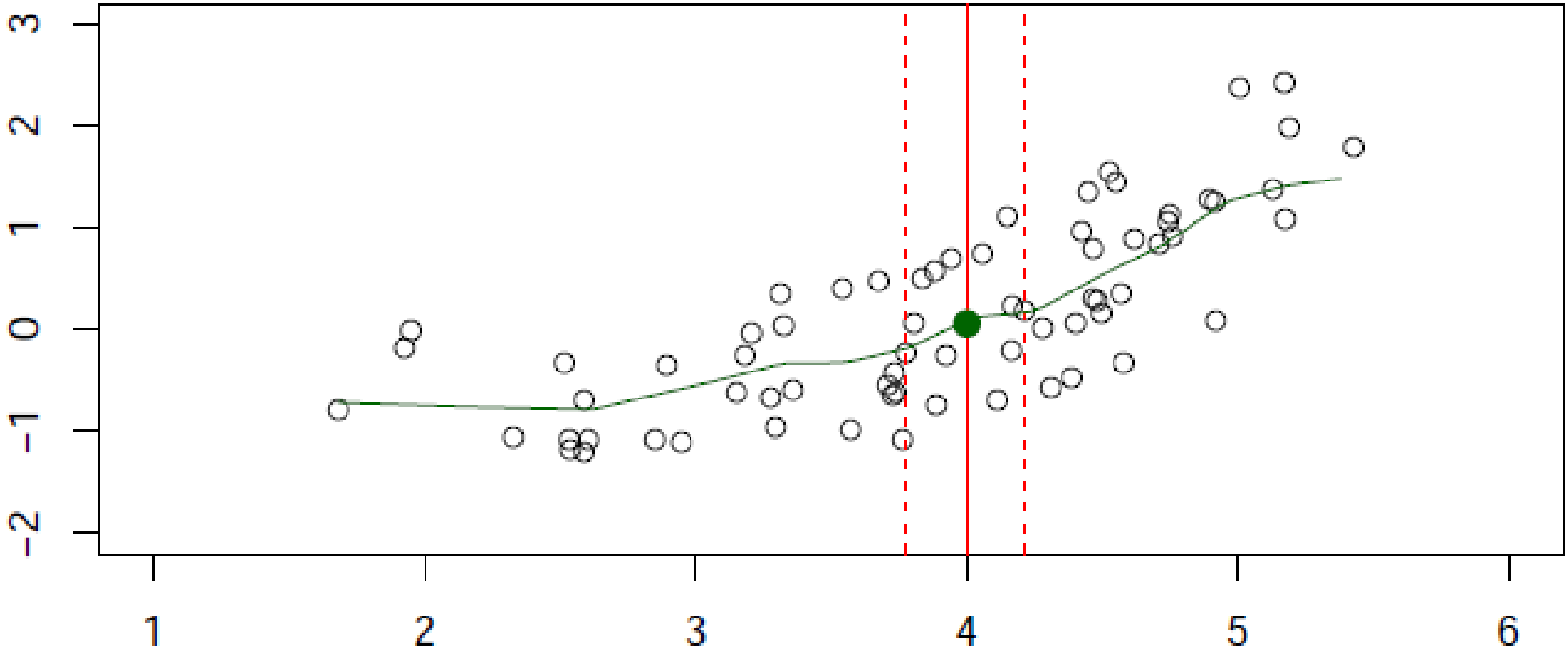
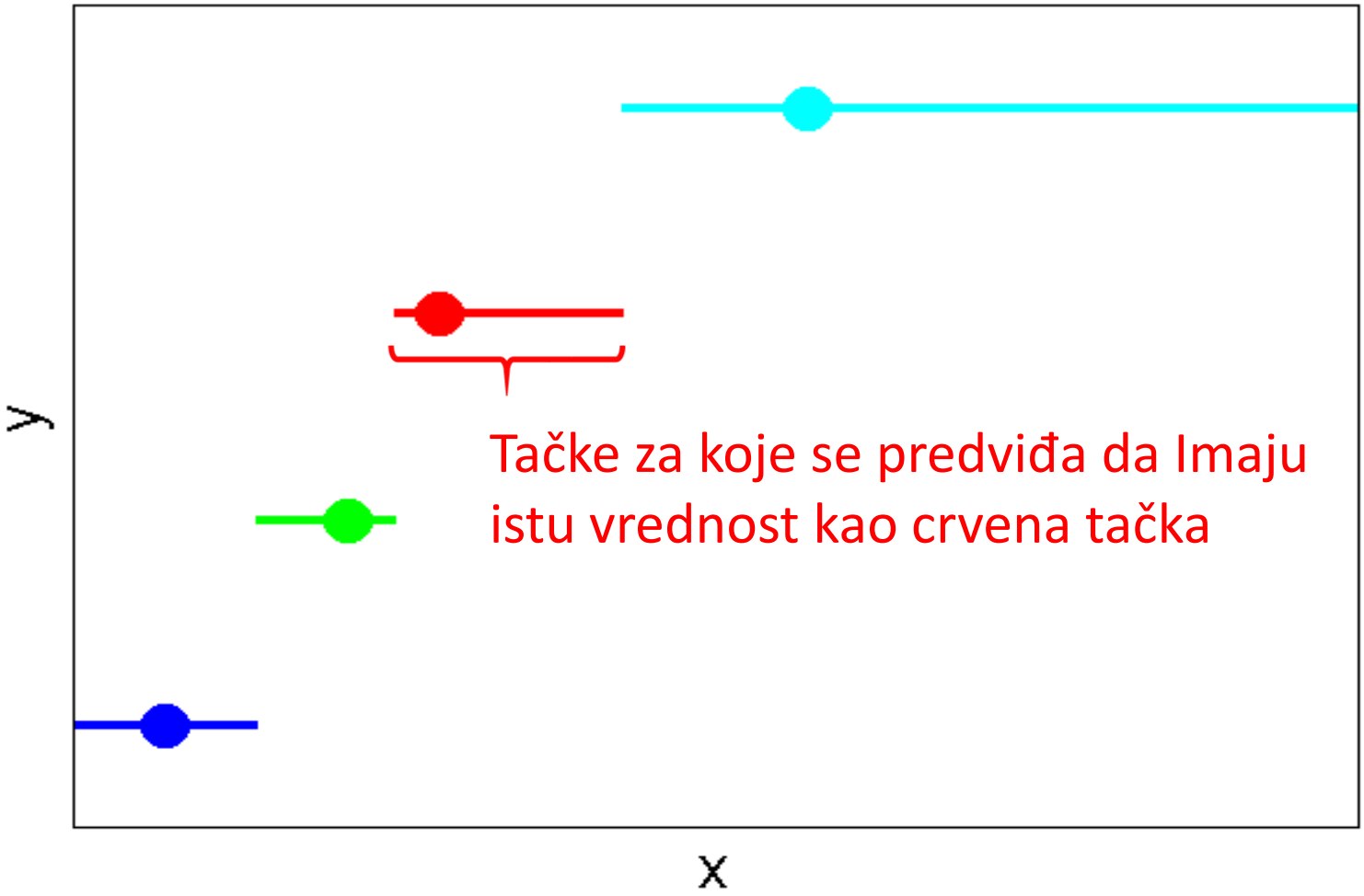
Najbliži susedi (lokalno uprosečavanje)

Ovo se zove *nearest neighbors* ili local *averaging*



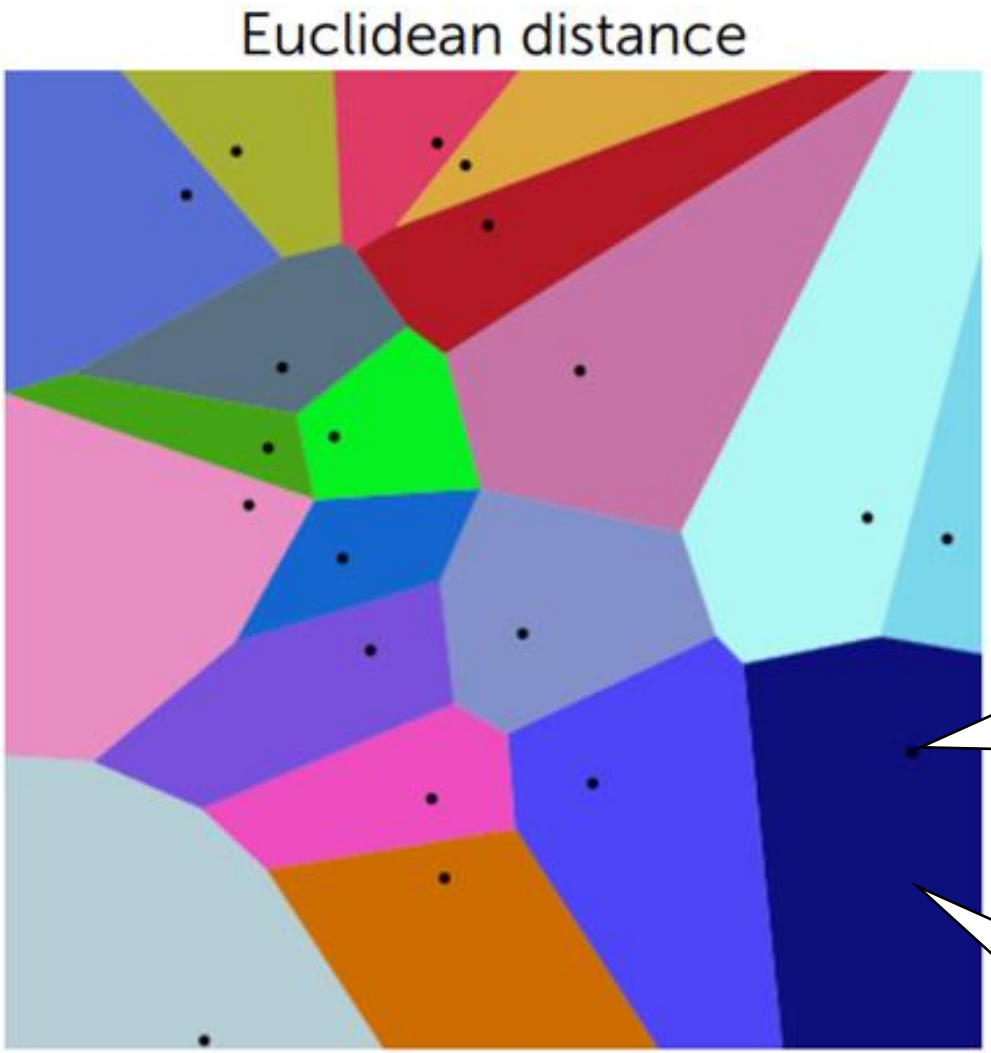
1-NN regresija



1-NN regresija

Ulaz	• $T = \{(x^{(i)}, y^{(i)}), i = 1, \dots, N\}$ • $x^{(q)}$: tačka za koju treba da odredimo $y^{(q)}$ vrednost
Postupak	Inicijalizovati $Dist2NN = \infty$, $\hat{y}^{(q)} = NaN$ for *i* = 1,2,…,*N* $\delta = distance(x^{(i)}, x^{(q)})$ if $\delta < Dist2NN$ $Dist2NN = \delta$ $\hat{y}^{(q)} = y^{(i)}$
Izlaz	$\hat{y}^{(q)}$

1-NN regresija



Voronoi-ev dijagram: vizualizacija 1-NN u više dimenzija

Ne moramo eksplicitno formirati regije, dovoljna nam je definicija udaljenosti

Svaka regija sadrži tačno jednu tačku $x^{(i)}$

svakoj tački regije „najbliža“ tačka $x^{(i)}$

Metrika udaljenosti

- Euklidska udaljenost $distance\left(x^{(j)}, x^{(q)}\right) = \left|x^{(j)} - x^{(q)}\right|$

- U višedimenzionom prostoru možemo pripisati težine dimenzijama (neke varijable su bitnije)
  - Npr. u predikciji životnog veka , vakcinacija i GDP su važnije od stepena kriminala

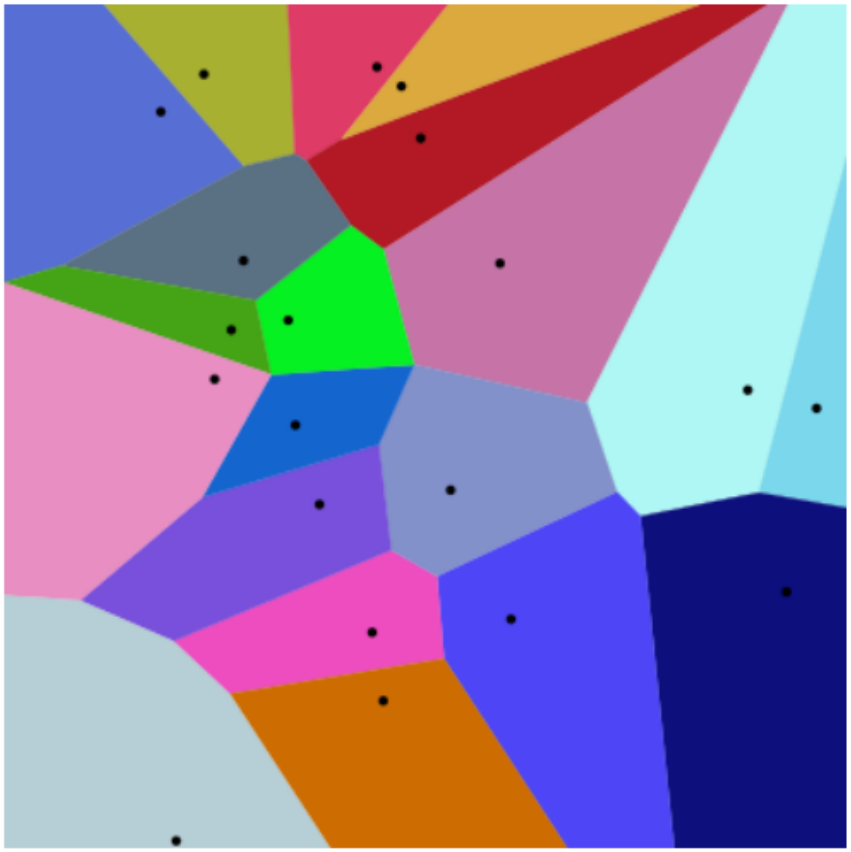
$$distance\left(x^{(j)}, x^{(q)}\right) = \sqrt{w_1({x_1}^{(j)} - {x_1}^{(q)})^2 + \cdots + w_d({x_d}^{(j)} - {x_d}^{(q)})^2}$$

- Druge metrike udaljenosti: Mahalanobis, rank-based, correlation-based, cosine similarity, Manhattan, Hamming,…

Metrika udaljenosti

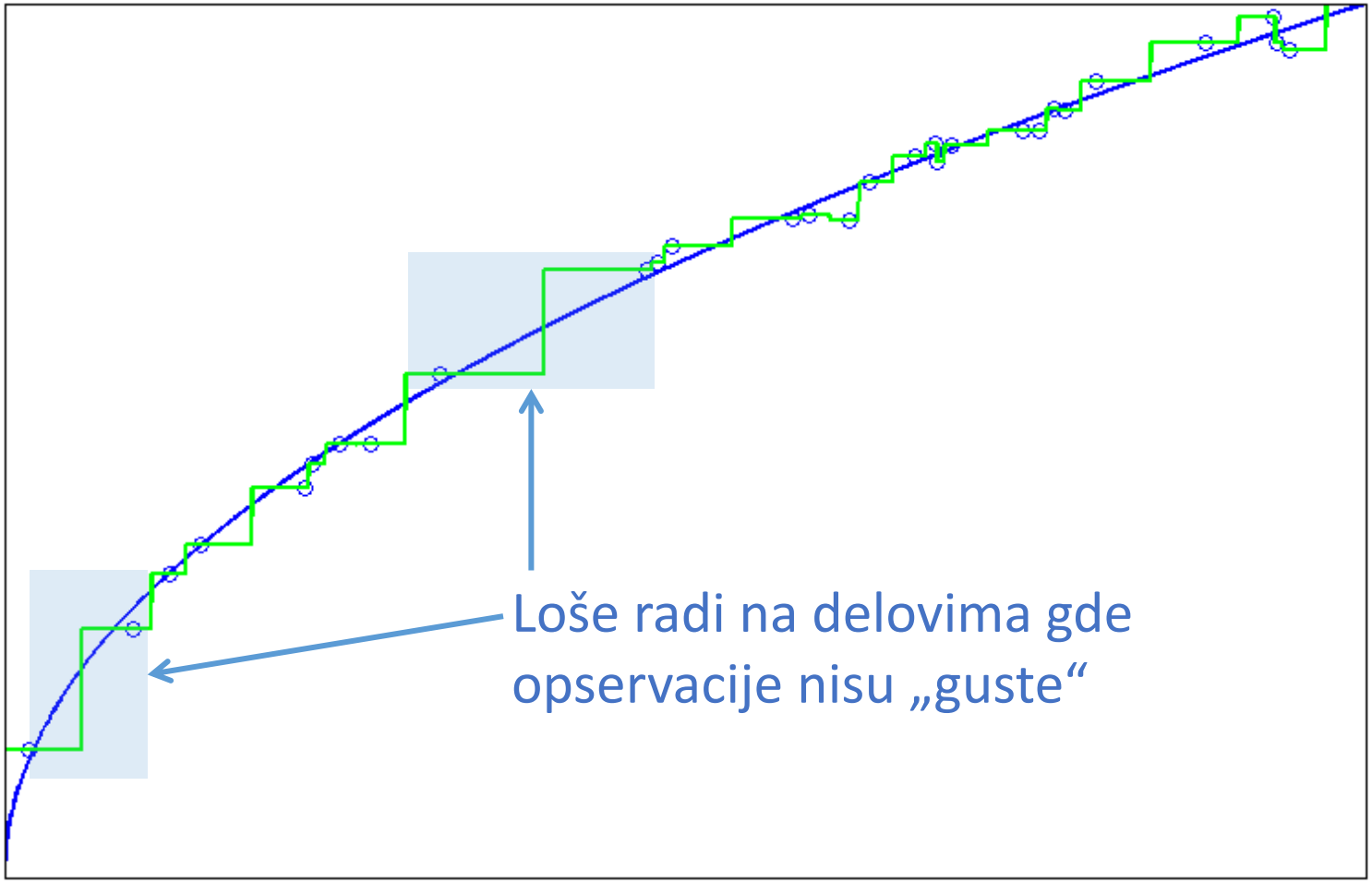
Različite metrike udaljenosti rezultuju različitim prediktivnim površinama

Euclidean distance



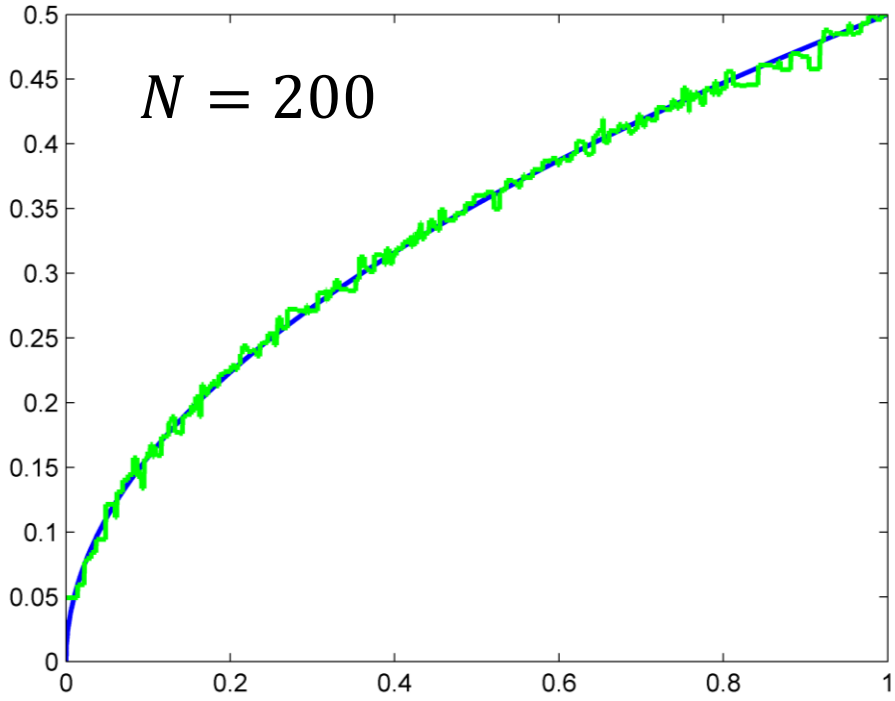
Manhattan distance

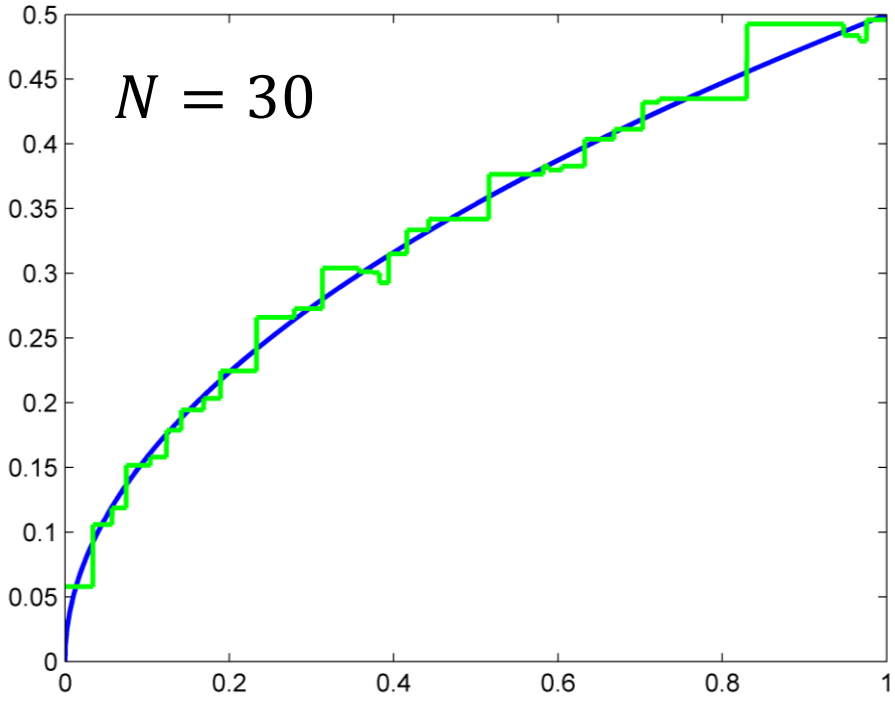
1-NN regresija u praksi



1-NN regresija ima dobre performanse ako je N veliko, a šum mali

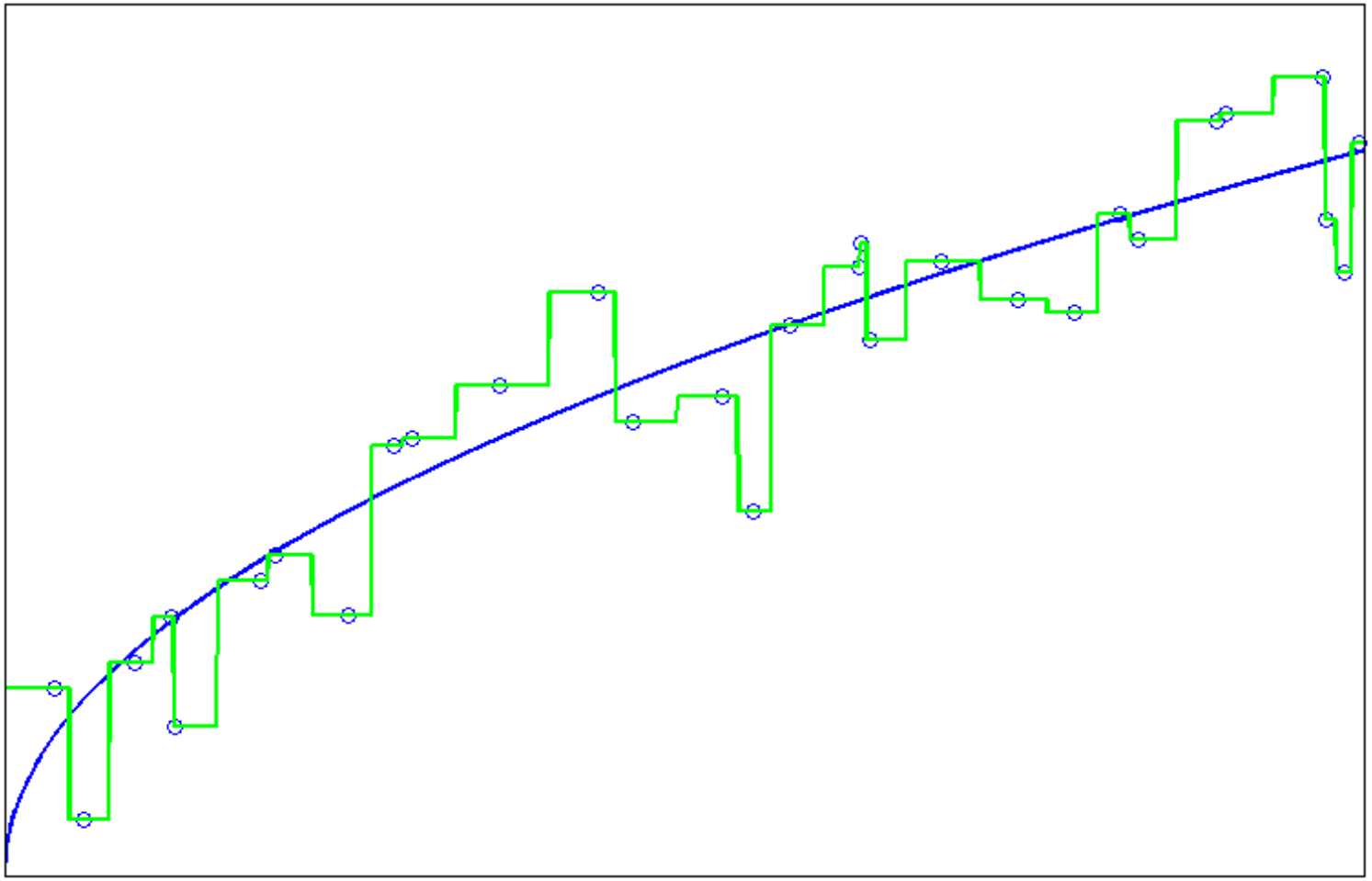
1-NN regresija u praksi





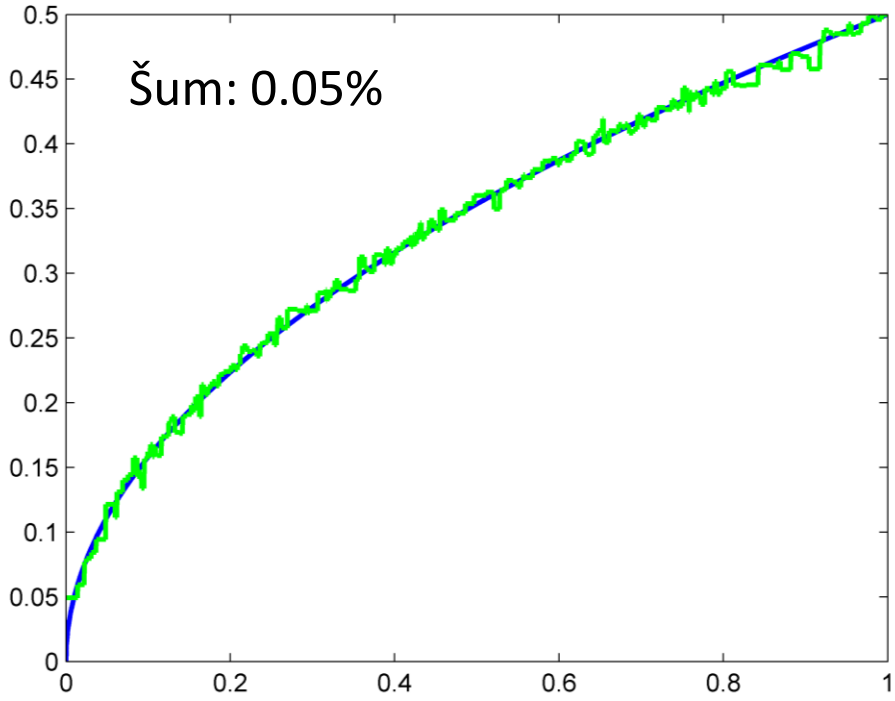
Šum: 0.05%

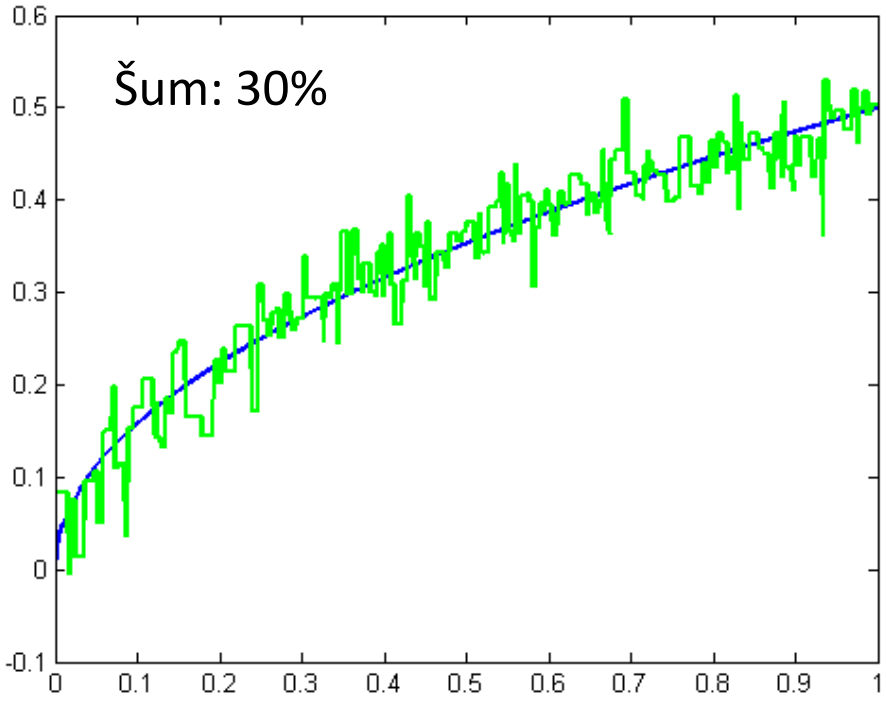
1-NN regresija u praksi



1-NN metod je osetljiv na šum u podacima – *overfitting*

1-NN regresija u praksi



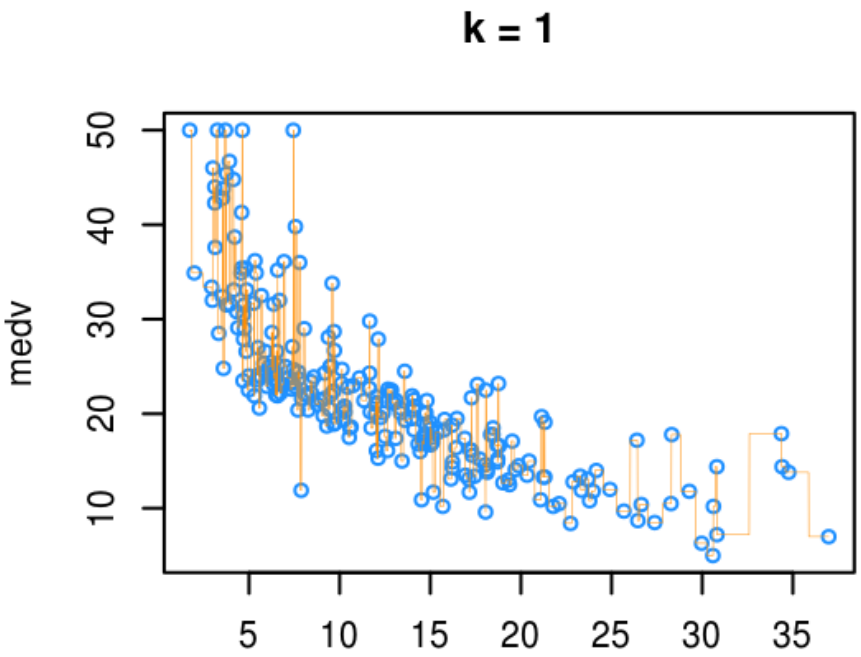


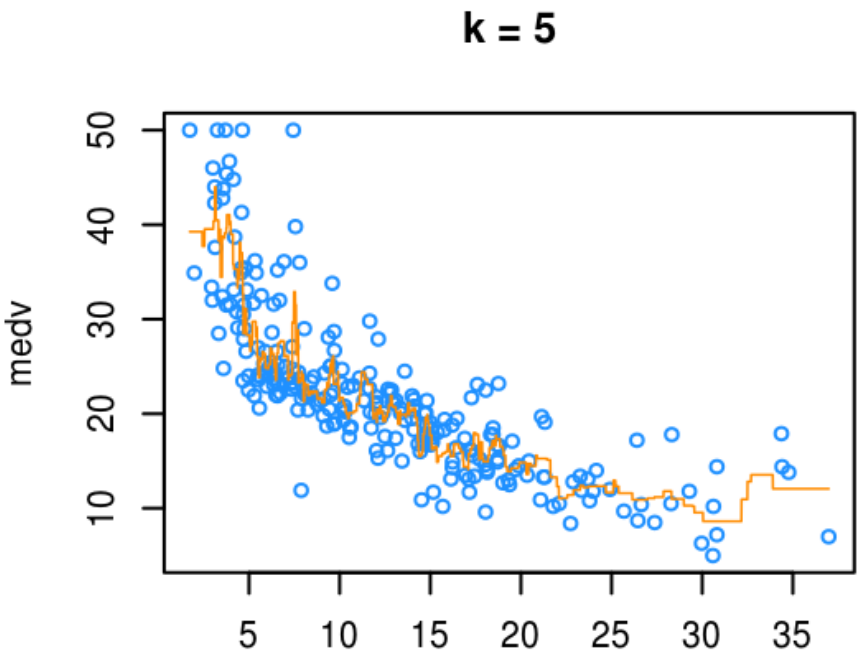
$$N = 200$$

Kako da rešimo preprilagođavanje?

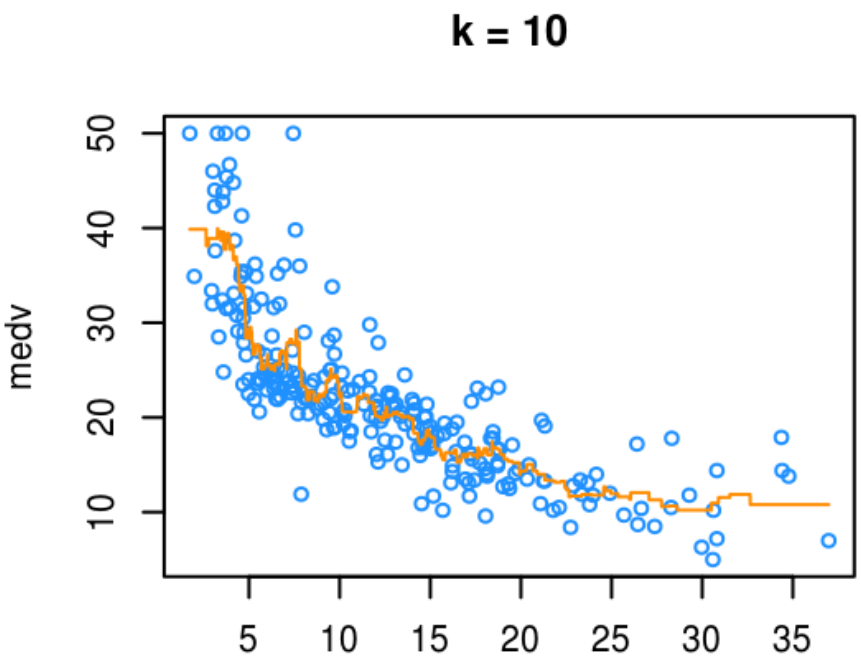
- Kod 1-NN metoda, najbliža tačka koja postoji u trening skupu može imati abnormalno veliku/malu vrednost u odnosu na stvarni trend u podacima
- Umesto da uzmemo samo jednog najbližeg suseda, možemo uzeti više (*k*)
- Ovo može rezultovati boljom aproksimacijom

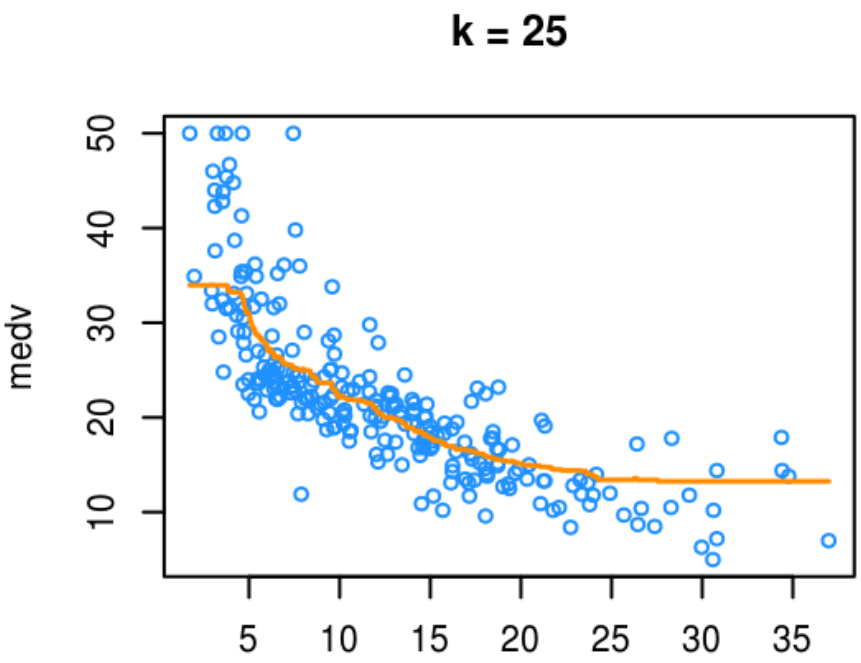
Uticaj *k*


k = 1
medv
10
20
30
40
50
5
10
15
20
25
30
35

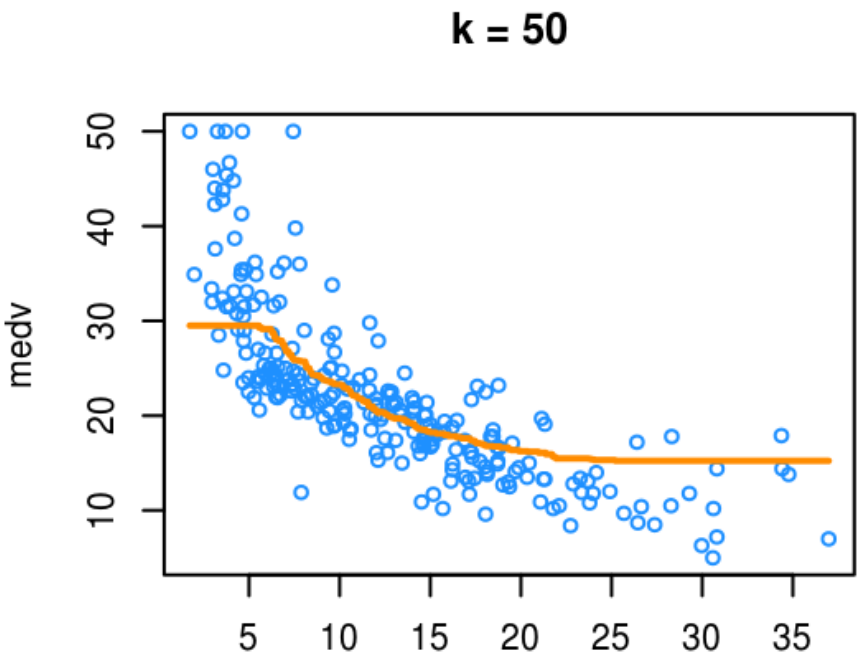

k = 5
medv
10
20
30
40
50
5
10
15
20
25
30
35

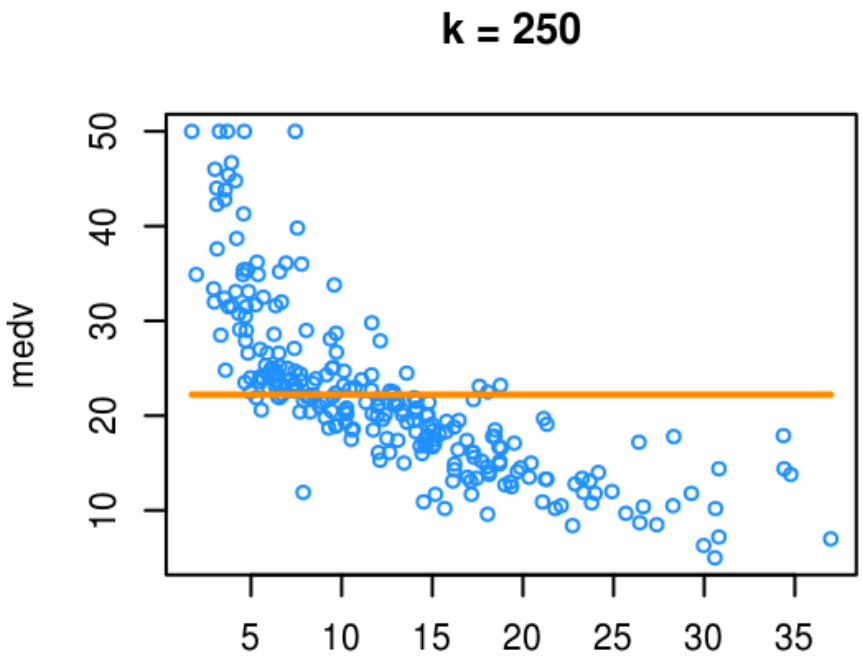
Uticaj *k*


k = 10
medv
10
20
30
40
50
5
10
15
20
25
30
35


k = 25
medv
10
20
30
40
50
5
10
15
20
25
30
35

Uticaj *k*

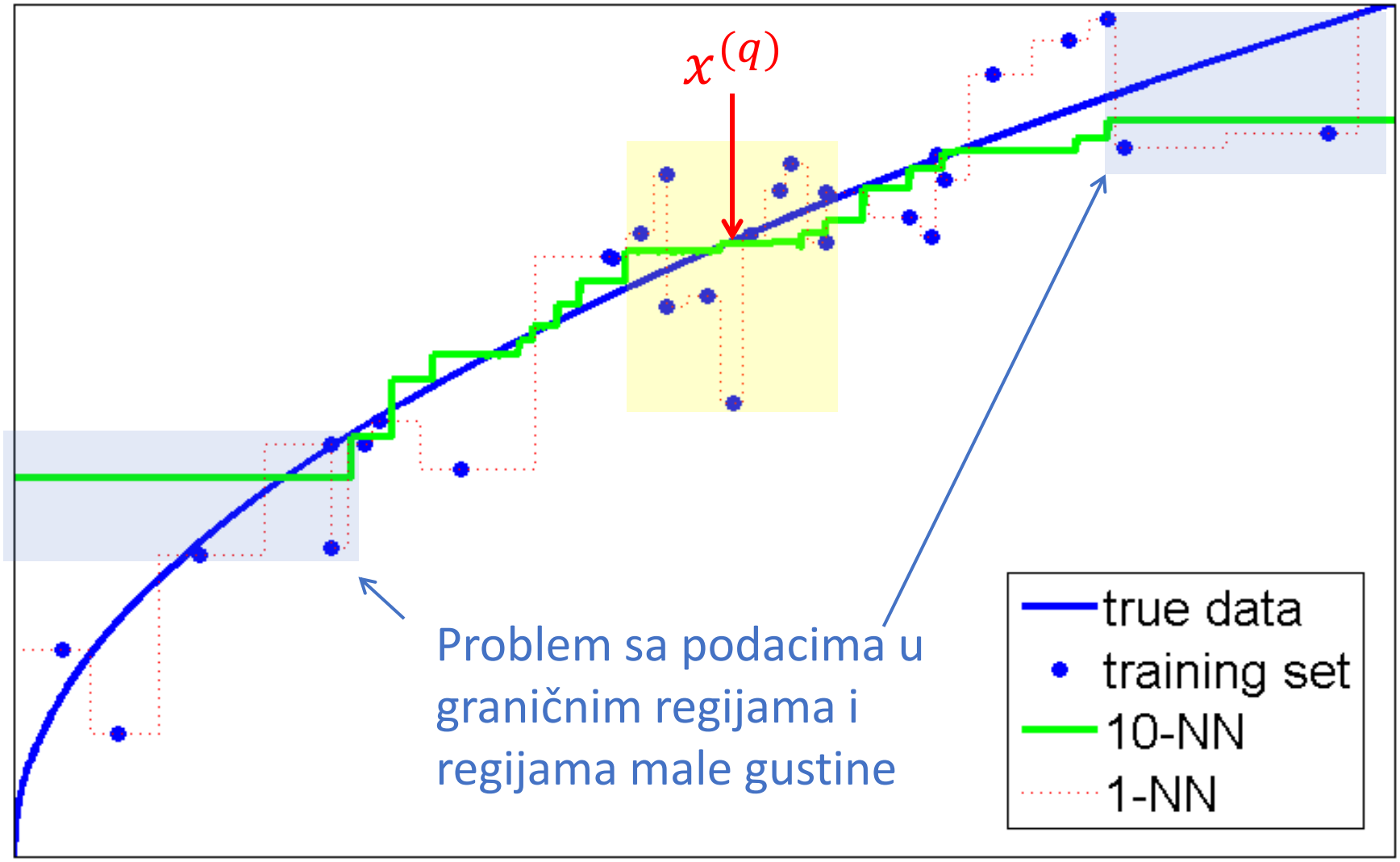

k = 50
medv
10
20
30
40
50
5
10
15
20
25
30
35


k = 250
medv
10
20
30
40
50
5
10
15
20
25
30
35

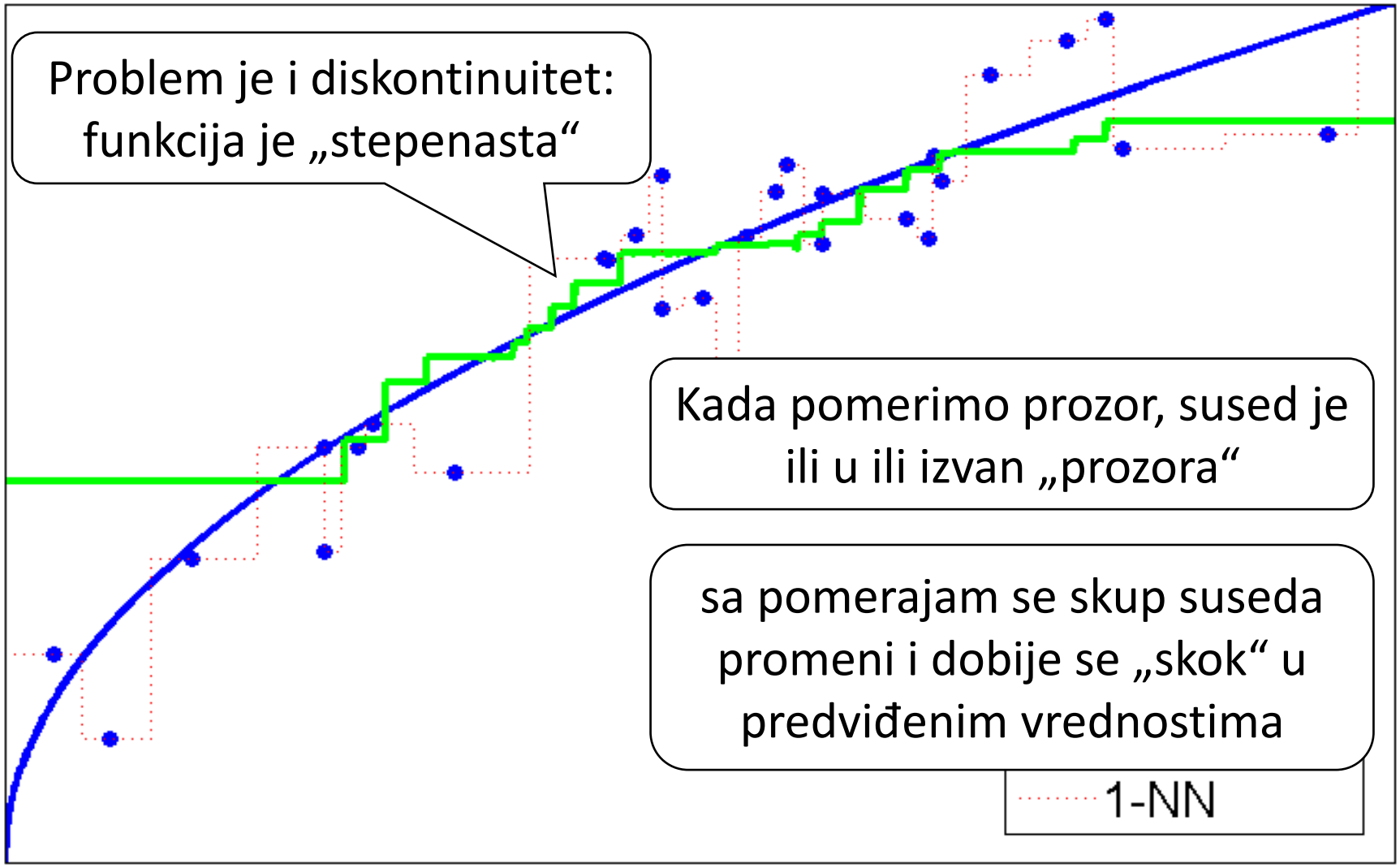
k-NN regresija

Ulaz	• $T = \{(x^{(i)}, y^{(i)}), i = 1, \dots, N\}$ • $x^{(q)}$ – tačka za koju treba da odredimo $y^{(q)}$ vrednost
Postupak	1. U trening skupu pronaći *k* tačaka najbližih tački $x^{(q)}$: $\{(x^{(NN1)}, y^{(NN1)}), \dots, (x^{(NNk)}, y^{(NNk)})\}$ 2. Predvideti $\hat{y}^{(q)} = \frac{1}{k}(y^{(NN1)} + \dots + y^{(NNk)})$
Izlaz	$\hat{y}^{(q)}$

k-NN regresija

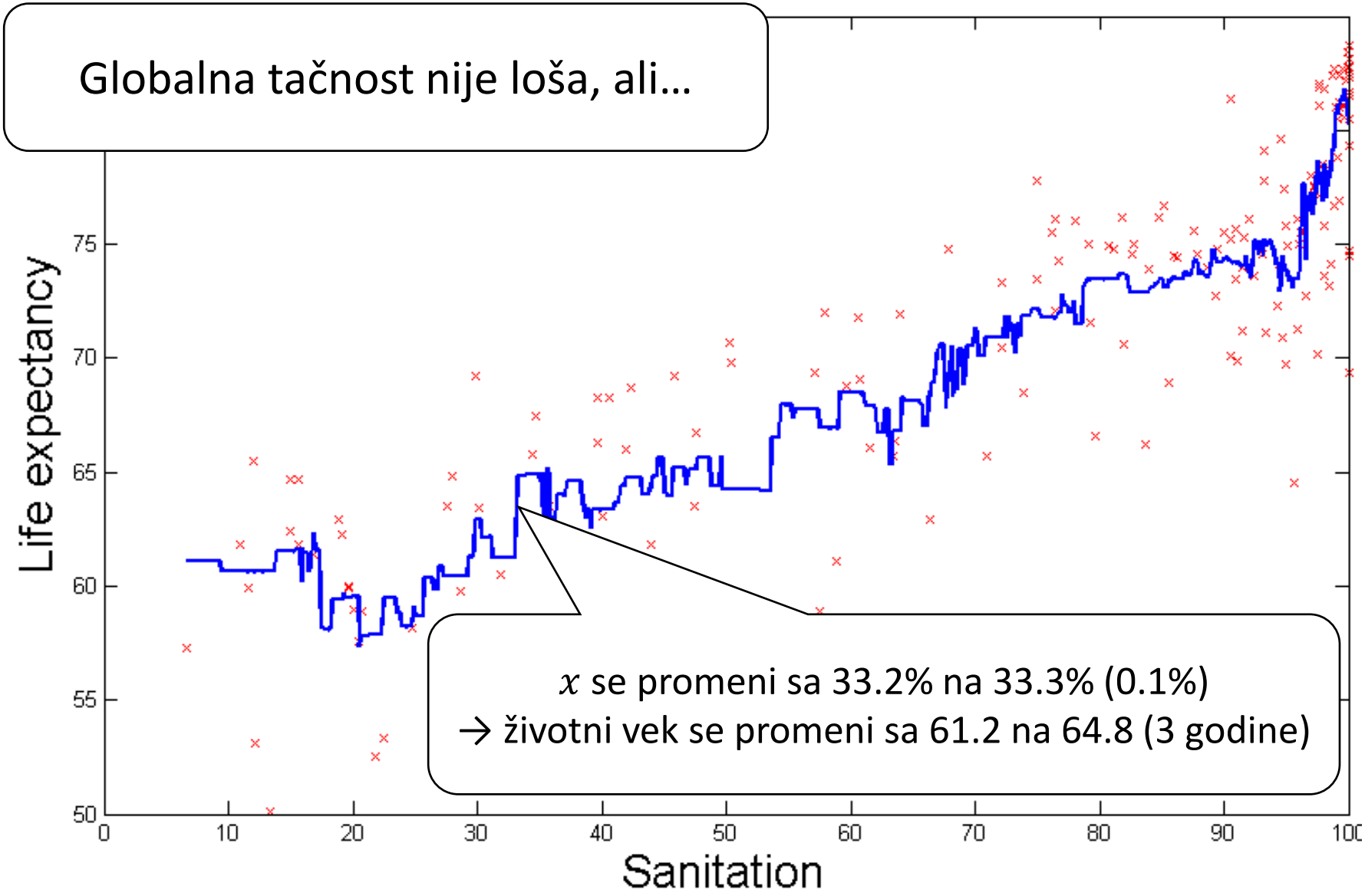


k-NN regresija



Problem sa diskontinuitetom

Globalna tačnost nije loša, ali…



x se promeni sa 33.2% na 33.3% (0.1%)
→ životni vek se promeni sa 61.2 na 64.8 (3 godine)