K najbližih suseda

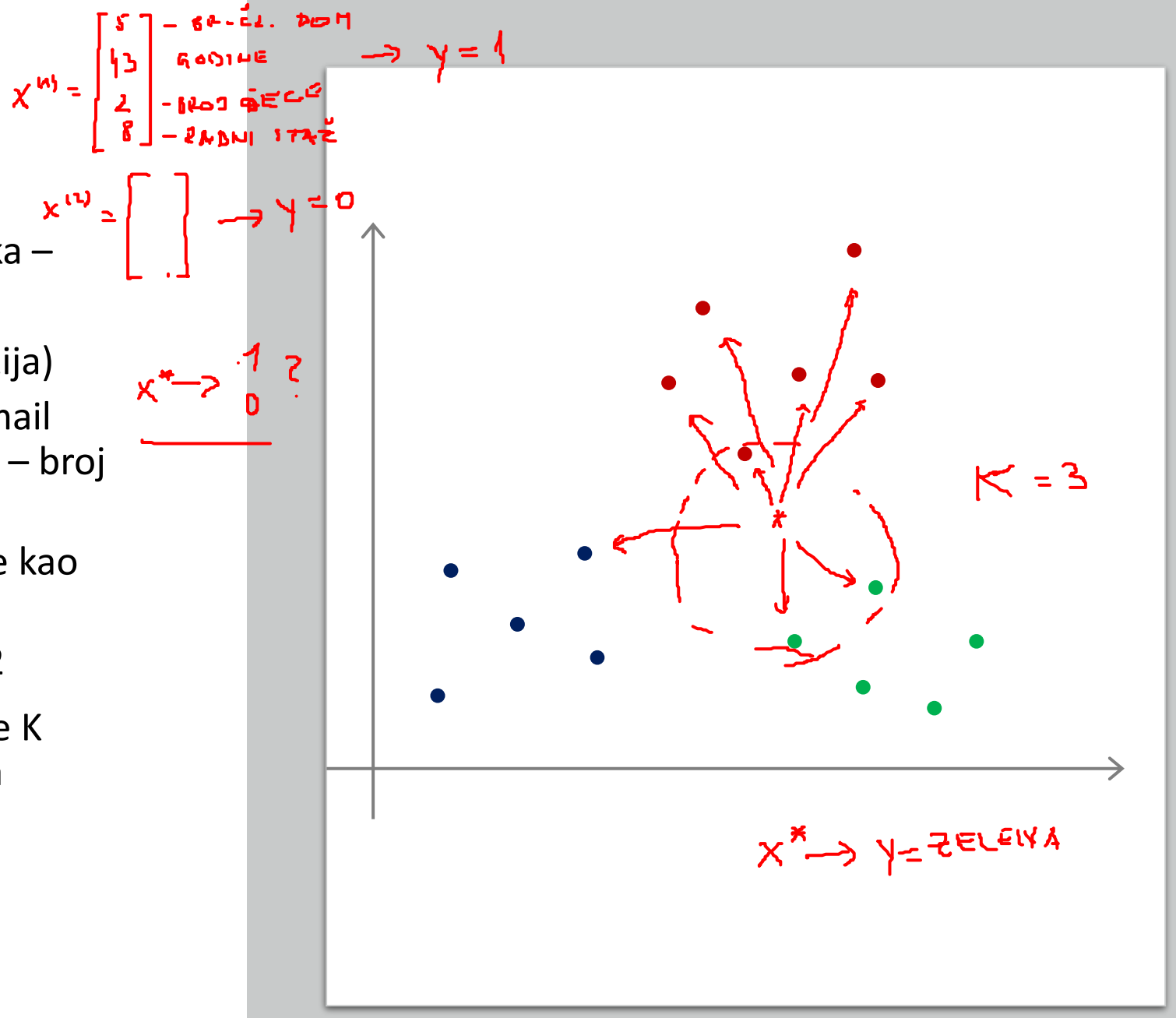
Primenjeni algoritmi

O algoritmu

- Supervizorska tehnika (nadgledano učenje)
- Podaci su predstavljeni vektorom karakteristika X
- Motivacija: obojene tačke u prostoru, svaka dimenzija odgovara jednoj osobini, a boja odgovara kategoriji
  - Cilj: klasifikovati novu tačku (odrediti joj boju) ako joj je zadato mesto u prostoru
- Meri se udaljenost nove tačke i K najbližih već klasifikovanih podataka

Kako radi?

- Počinje sa m klasifikovanih podataka – supervizorsko učenje
- Svaka tačka ima n obeležja (dimenzija)
  - npr. 1. obeležje – broj reči u email poruci, 2. – broj znakova „!“, 3. – broj pravopisnih grešaka, itd.
  - Svaki takav vektor se klasifikuje kao *spam* poruka ili ne
  - Broj grupa/klasa ne mora biti 2
- Kada se pojavi nova tačka potraži se K najbližih tačaka i odredi njena klasa

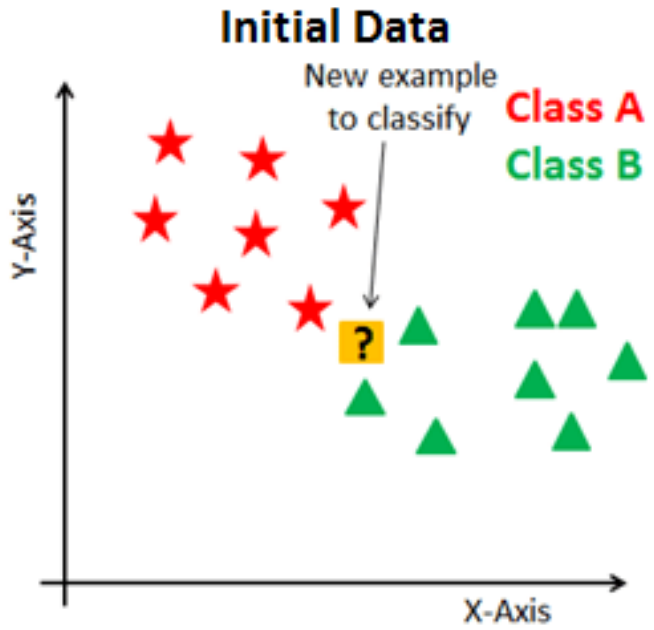


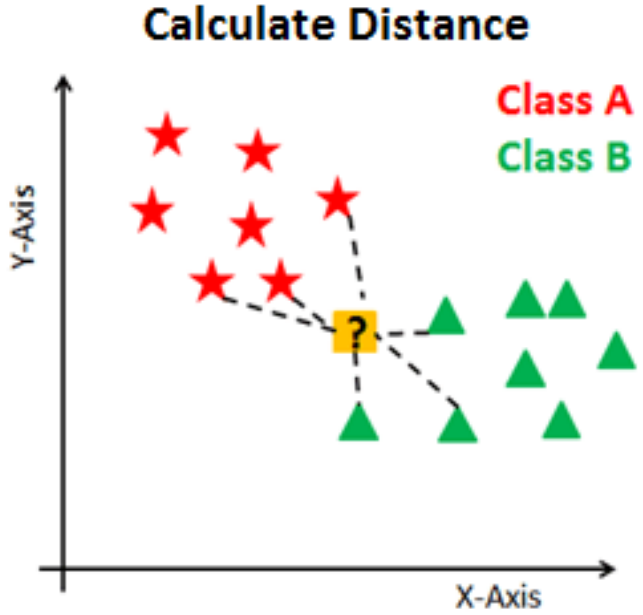
Algoritam

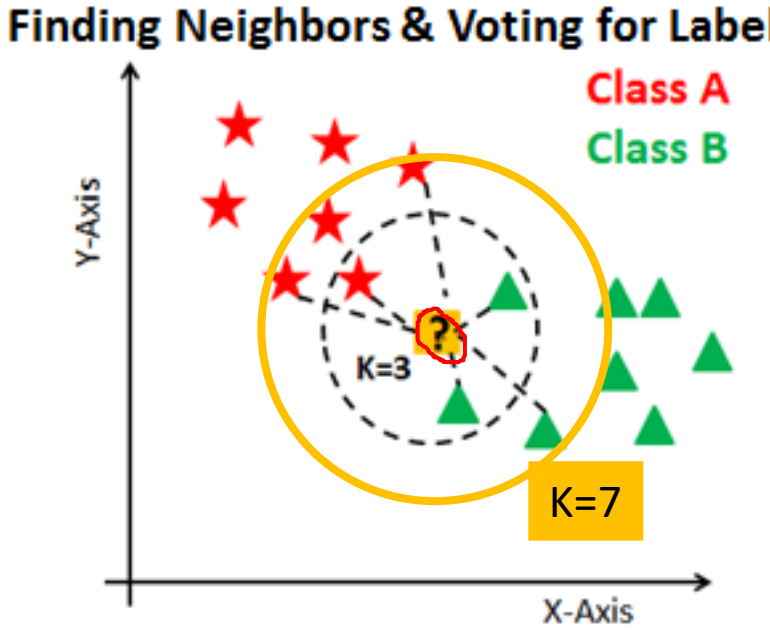
$$\frac{x - \bar{x}}{|x_{max} - x_{min}|} \rightarrow \begin{matrix} [-1, 1] \\ [-0.5, 0.5] \end{matrix}$$

- Korak 1: Skaliranje
  - Prilagođavanje atributa da budu uporedivi. Npr. U email poruci skalirati da prosečan broj reči bude 0 sa standardnom devijacijom 1. Uraditi ovo za sve atribute.

- Korak 2: Nova tačka
  - Meri se udaljenost nove tačke do svih ostalih tačaka. Posmatra se K tačaka sa najmanjim rastojanjem – najzastupljenija klasa među njima je i klasa nove tačke

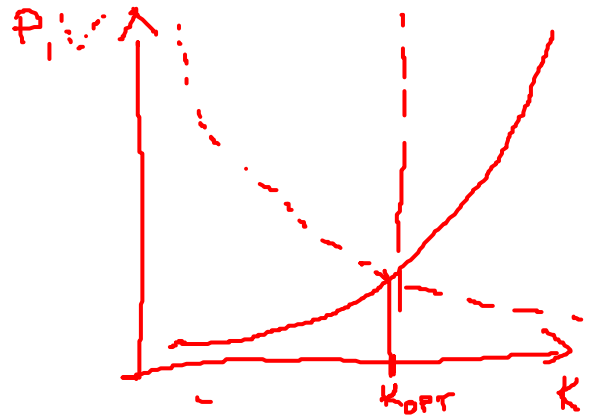

Initial Data
New example
to classify
Class A
Class B
?
Y-Axis
X-Axis

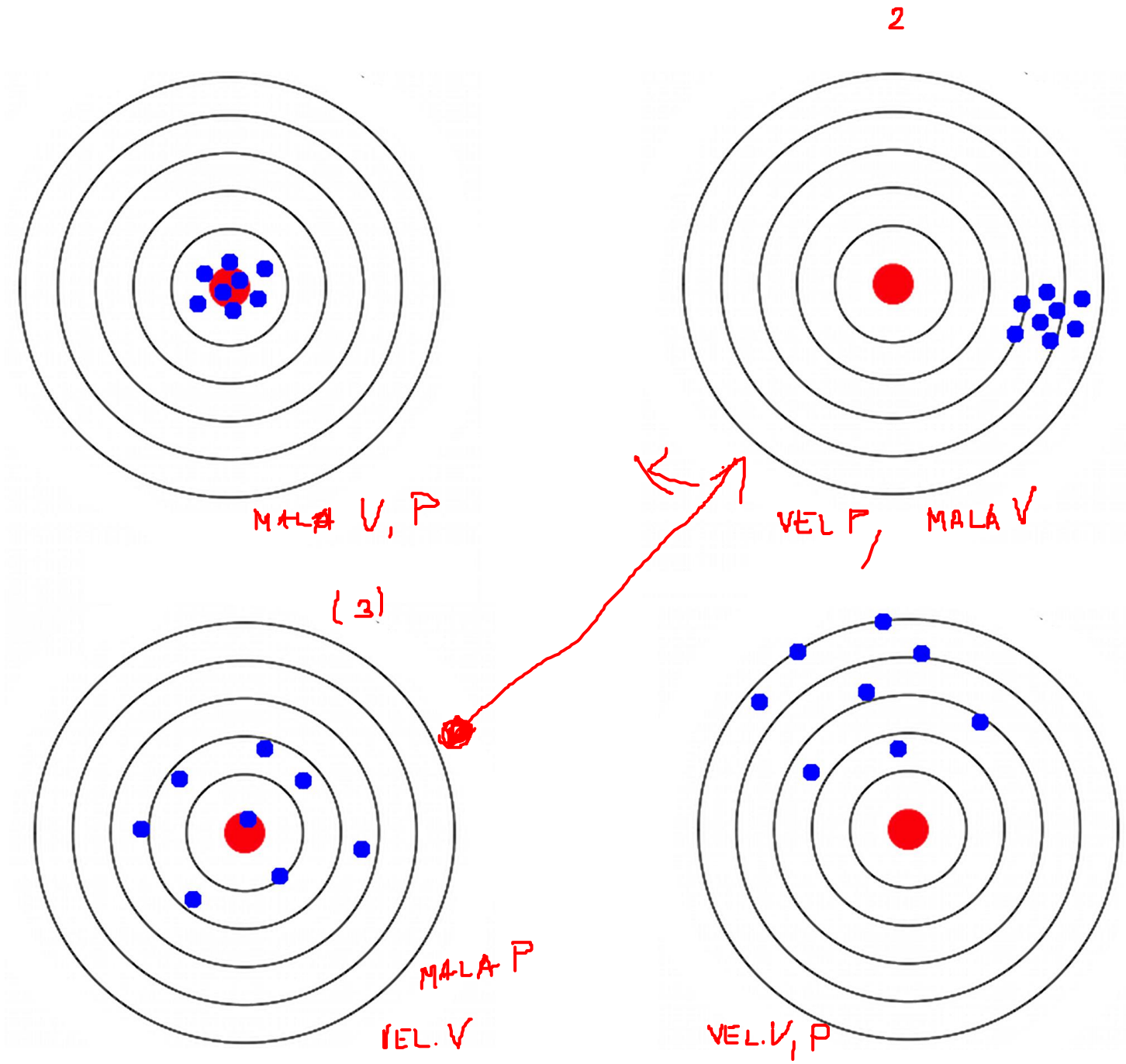

Calculate Distance
Class A
Class B
?
Y-Axis
X-Axis


Finding Neighbors & Voting for Labels
Class A
Class B
?
K=3
K=7
Y-Axis
X-Axis

Kako odabrati K?

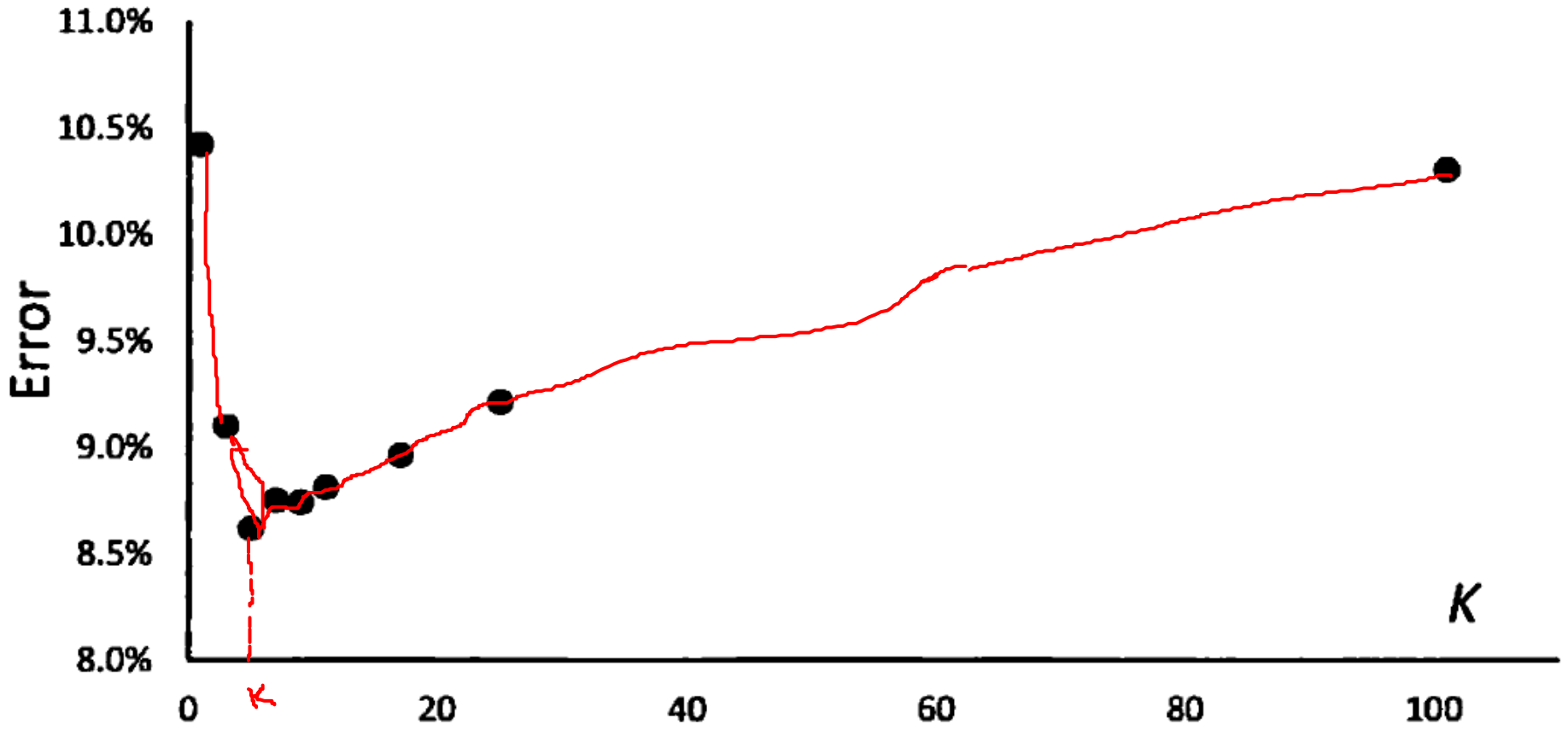
- Mali K, dobićete malu pristrasnost (*Bias*), ali veliku varijansu (*Variance*)
- Veliko K daje veliku pristrasnost, ali malu varijansu.





Odnos K i greške u klasifikaciji

- K=5 je optimalno



Karakteristike KNN

- Ne postoji stvarno učenje
- Algoritmi kod kojih se model generiše samo dodavanjem novih tačaka – lenji algoritam
- Algoritmi koji obučavaju pre dodavanja novih podataka su revnosni, ali nakon nekog vremena (stalnog novi dodavanja podataka) zastarevaju
- Lenji algoritmi – pogodni kod stalne pojave novih podataka

Mane KNN

- Klasifikacija novih tačaka može biti veoma spora
- Neravnomerni podaci (mnogo više podataka jedne klase, npr. trouglova nego krugova)
  - dovode do tendencije dodele novih tačaka toj klasi
  - treba odraditi neku vrstu inverzne udaljenosti ili opadajuće eksponencijalne

Upotreba KNN za Regresiju

- Obežja su brojevi, a ne klase
- Npr. Atributi mogu biti godine, IQ i visina, obeležje je plata. Cilj: odrediti platu za novi uzorak
- Ova regresija se može izvršiti na sledeće načine:
  - Prosečna plata K najbližih suseda
  - Ponderisani prosek u odnosu na udaljenost
  - Eksponencijalno ponderisano rastojanje
  - Gausova funkcija rastojanja ili Gausov kernel
  - Postavljanje hiperravni u K tačaka – lokalna linearna regresija

Linearna regresija sa K=5 najbližih suseda