K Nearest Neighbours (kNN)

Jelena Jovanovic

jelena.jovanovic@fon.bg.ac.rs

kNN: Osnovne karakteristike

Jednostavan i vrlo intuitivan algoritam

Može se koristiti za zadatke klasifikacije i regresije

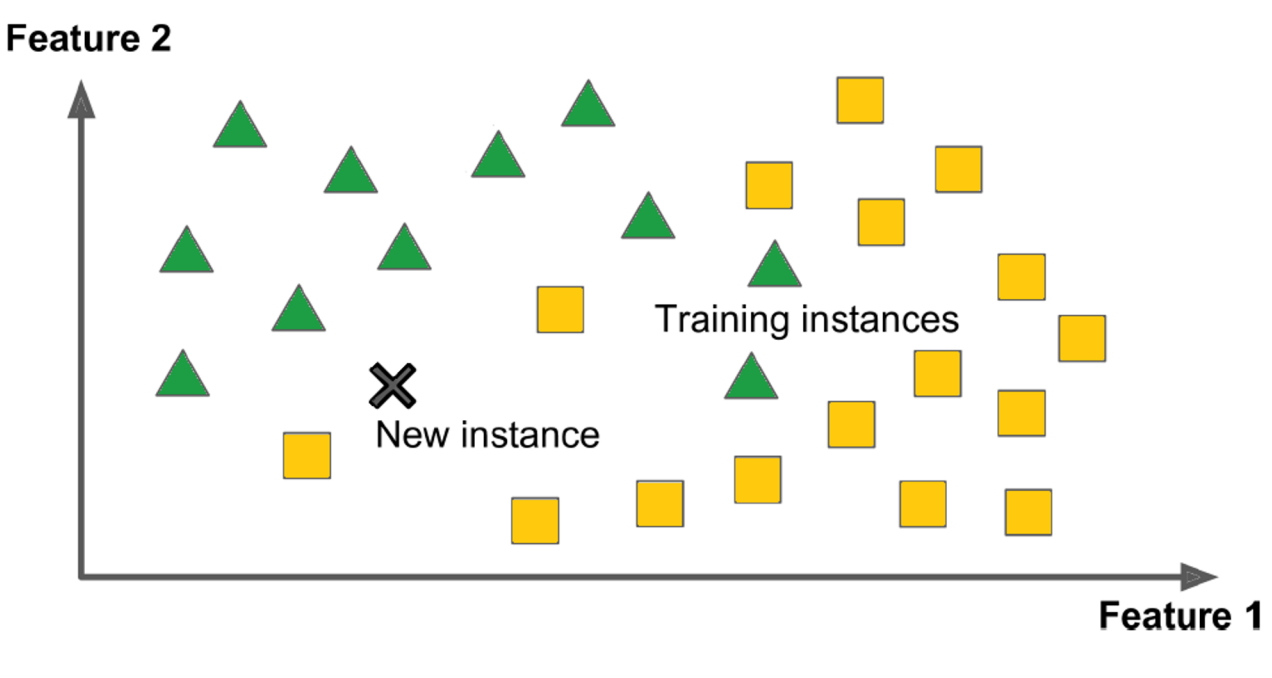
Za razliku od većine ML algoritama, kNN nema trening fazu i ne kreira model

kNN: Osnovne karakteristike

Predviđa izlaznu vrednost za novu instancu u dva koraka:

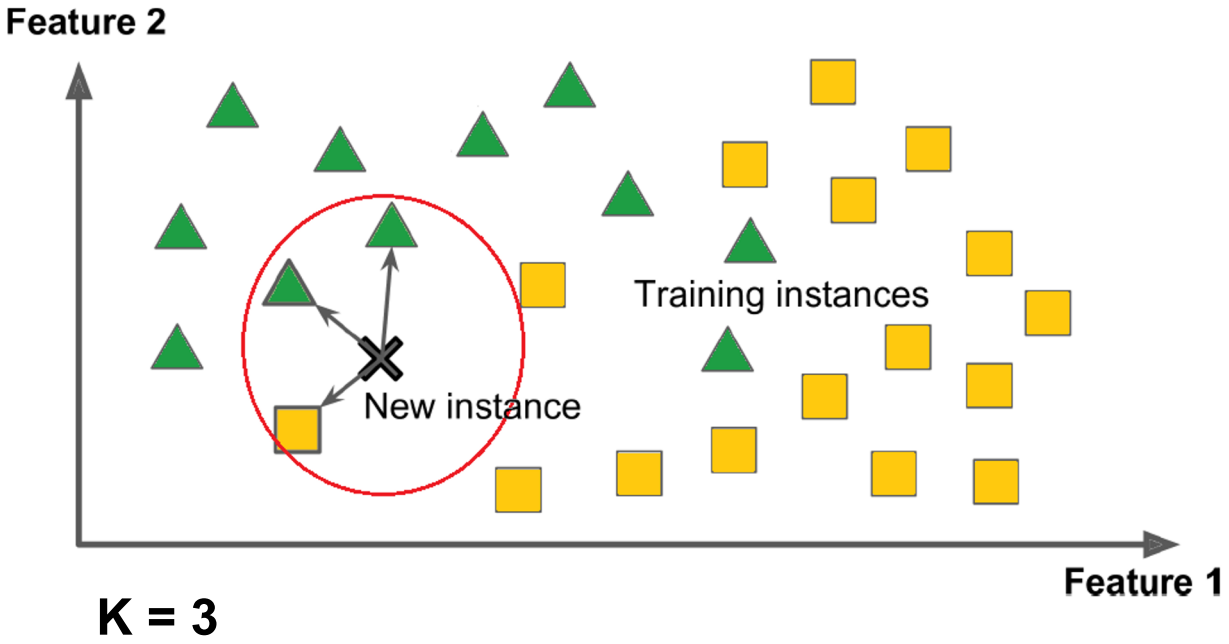
1) Pronađe *k* najsličnijih (tj., najmanje udaljenih) instanci u skupu za trening - tzv. *k* najbližih suseda

Kako kNN radi?



Izbor slike: https://mlarchive.com/machine-learning/k-nearest-neighbor-knn-explained/

Kako kNN radi?

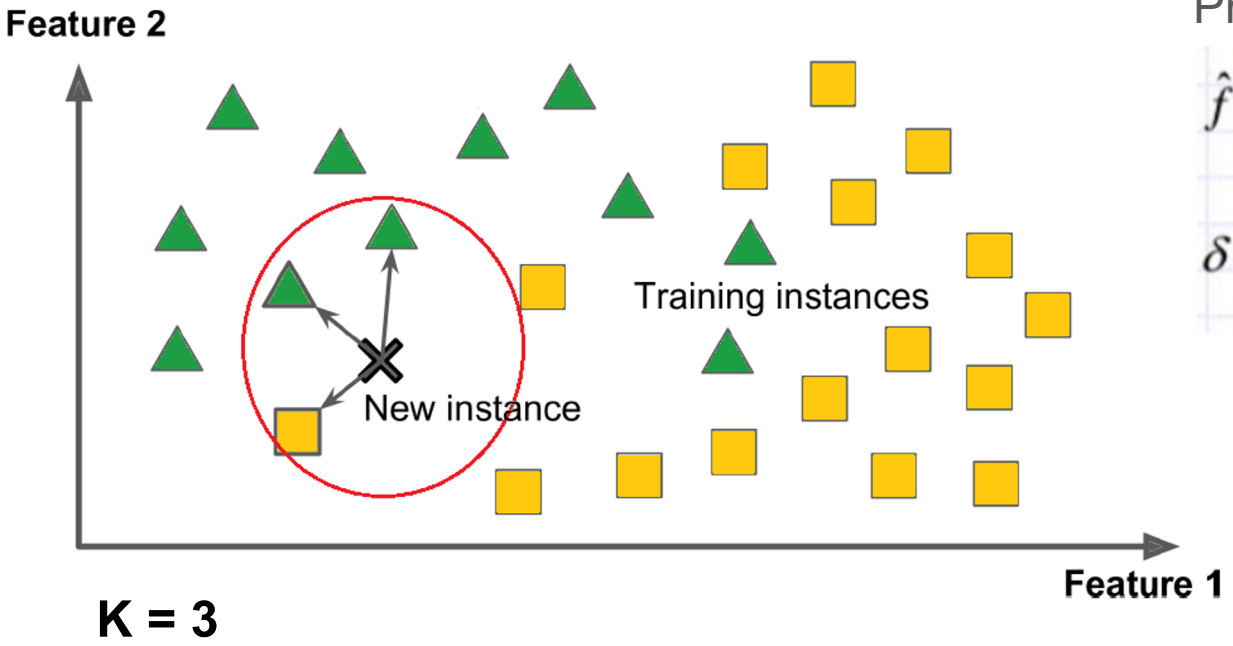


kNN: Osnovne karakteristike

Predviđa izlaznu vrednost za novu instancu u dva koraka:

1) Pronađe *k* najsličnijih (tj., najmanje udaljenih) instanci u skupu za trening - tzv. *k* najbližih suseda
2) Na osnovu izlazne vrednosti *k* najbližih suseda, određuje izlaznu vrednost nove instance i to:
   a) Kao dominantnu klasu među *k* najbližih suseda, u slučaju klasifikacije
   b) Kao prosečnu vrednost izlazne varijable *k* najbližih suseda, u slučaju regresije

Klasifikacija primenom kNN-a

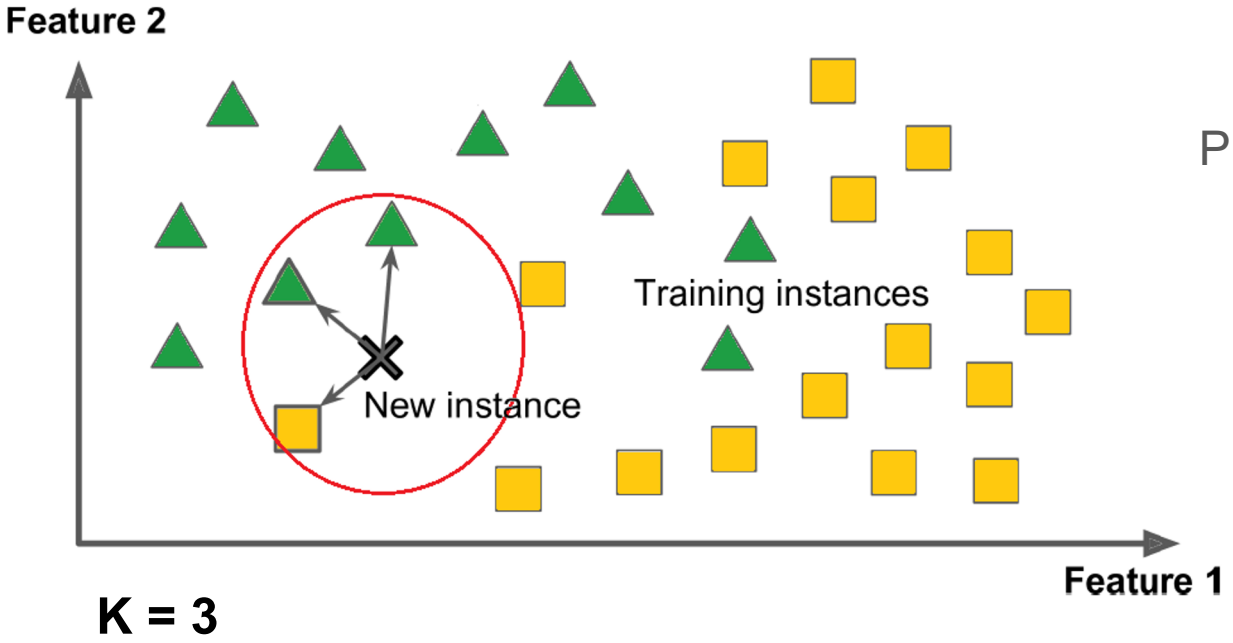


Predviđena izlazna vrednost:

$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$$

$$\delta(a,b) = \begin{cases} 1 & a = b \\ 0 & otherwise \end{cases}$$

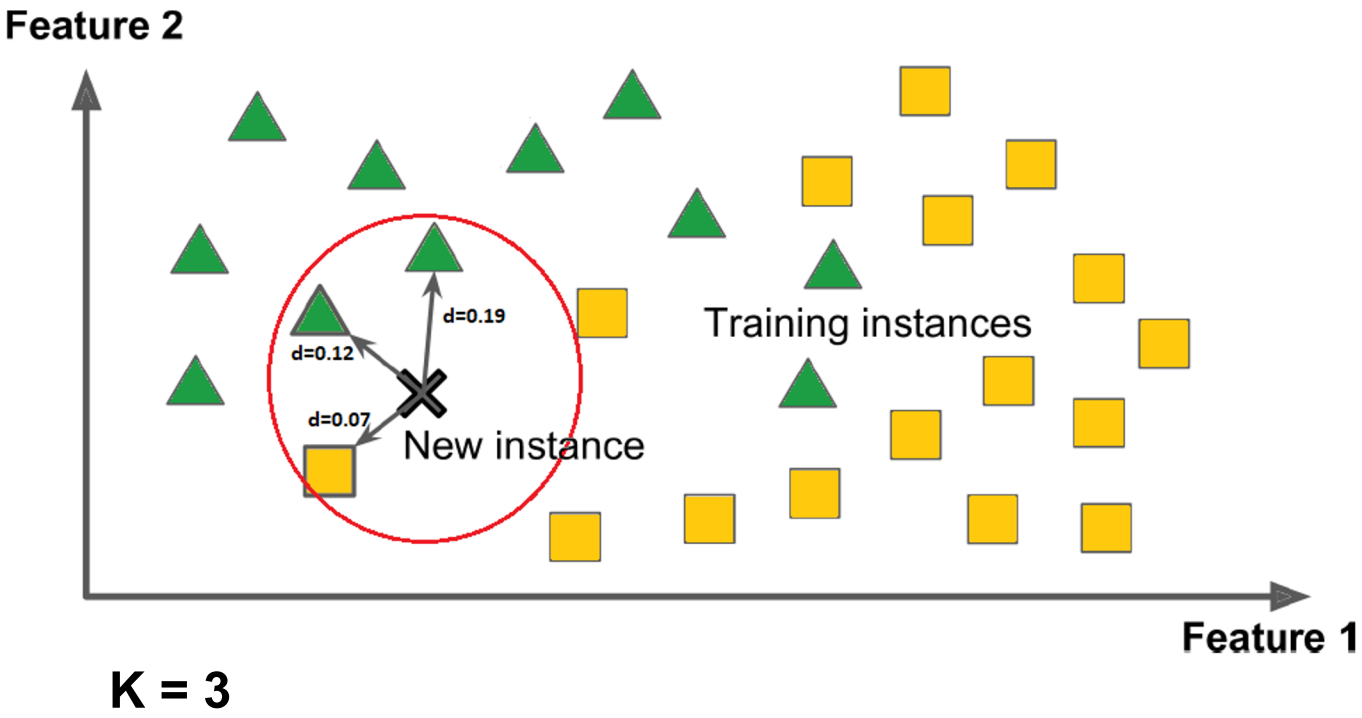
Regresija primenom kNN-a



Predviđena izlazna vrednost:

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} f(x_i)}{k}$$

Težinski kNN



Izvor: https://www.slideserve.com/keena/chapter-eight-instance-based-learning

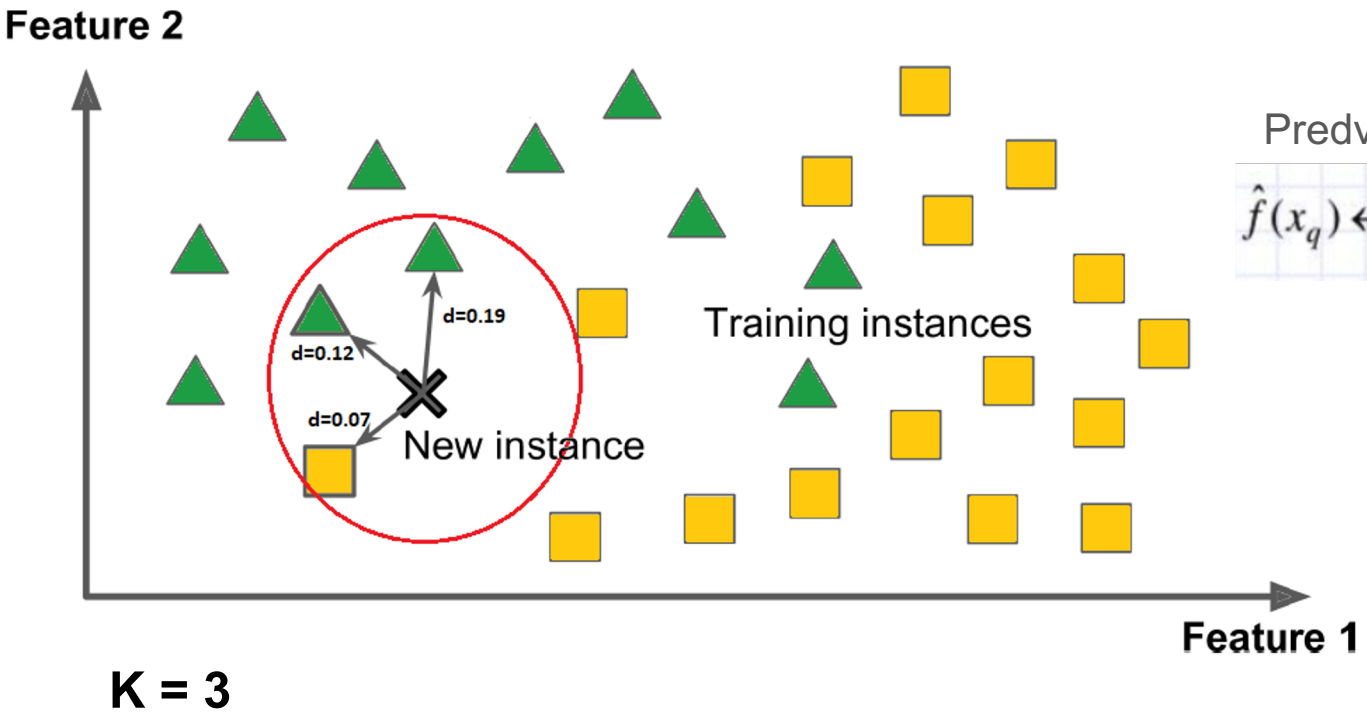
Težinski kNN: klasifikacija

Težina (značaj) “suseda”:

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

Predviđena izlazna vrednost:

$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} w_i \delta(v, f(x_i))$$



Izvor: https://www.slideserve.com/keena/chapter-eight-instance-based-learning

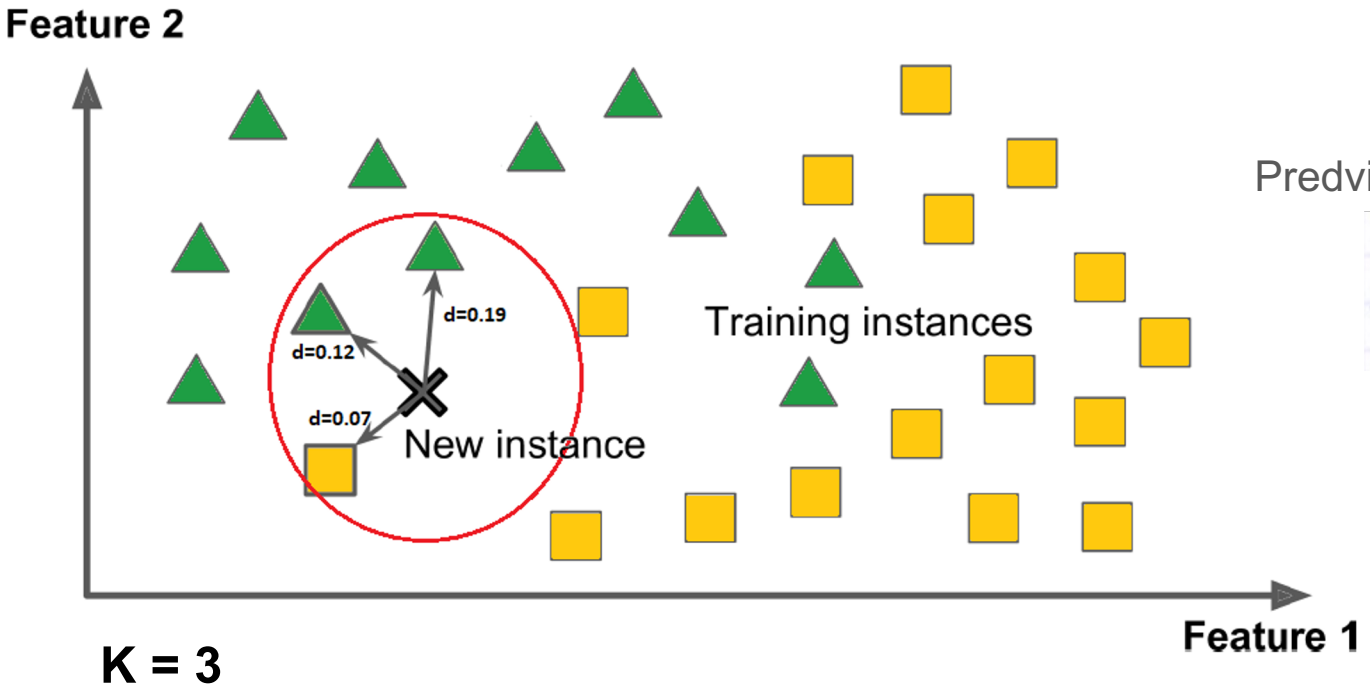
Težinski kNN: regresija

Težina (značaj) “suseda”:

$$w_i \equiv \frac{1}{d(x_q, x_i)^2}$$

Predviđena izlazna vrednost:

$$\hat{f}(x_q) \leftarrow \frac{\sum_{i=1}^{k} w_i f(x_i)}{\sum_{i=1}^{k} w_i}$$

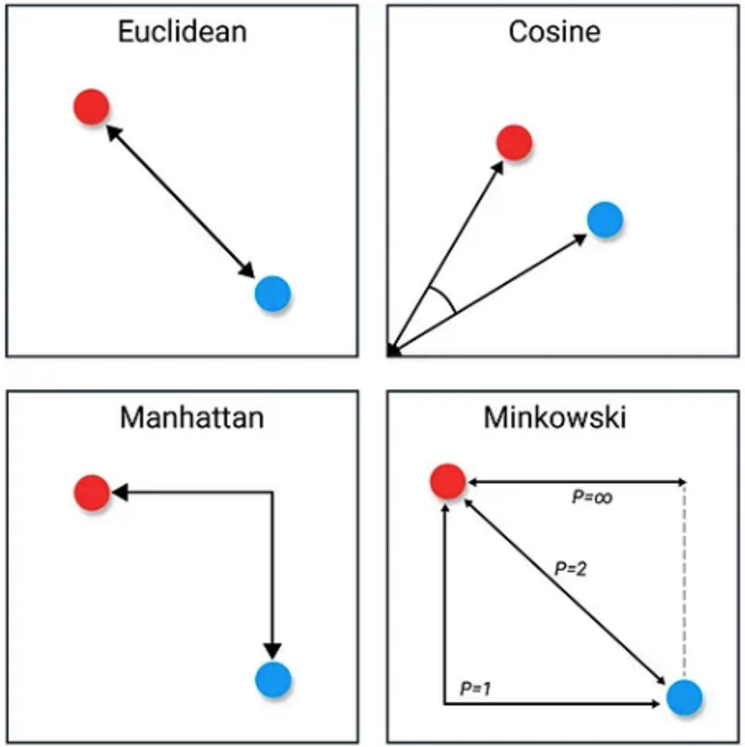


kNN: Hiper-parametri

Rezultati kNN-a zavise od 2 ključne stvari (hiper-parametara modela):

- Broja suseda (k)
- Izabrane mere udaljenosti / sličnosti

Mere udaljenosti / sličnosti



Euclidean: $$d = \sqrt{\sum_{j=1}^{n} (x_{sj} - x_{tj})^2}$$

Manhattan: $$d = \sum_{j=1}^{n} |x_{sj} - x_{tj}|$$

Minkowski: $$d = \left(\sum_{j=1}^{n} |x_{sj} - x_{tj}|^p \right)^{1/p}$$

Cosine: $$d = 1 - \frac{\sum_{j=1}^{n} x_{sj} x_{tj}}{\sum_{j=1}^{n} x_{sj}^2 \sum_{j=1}^{n} x_{tj}^2}$$

Izvori slike: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Uticaj različitih vrednosti *k* na predviđanje

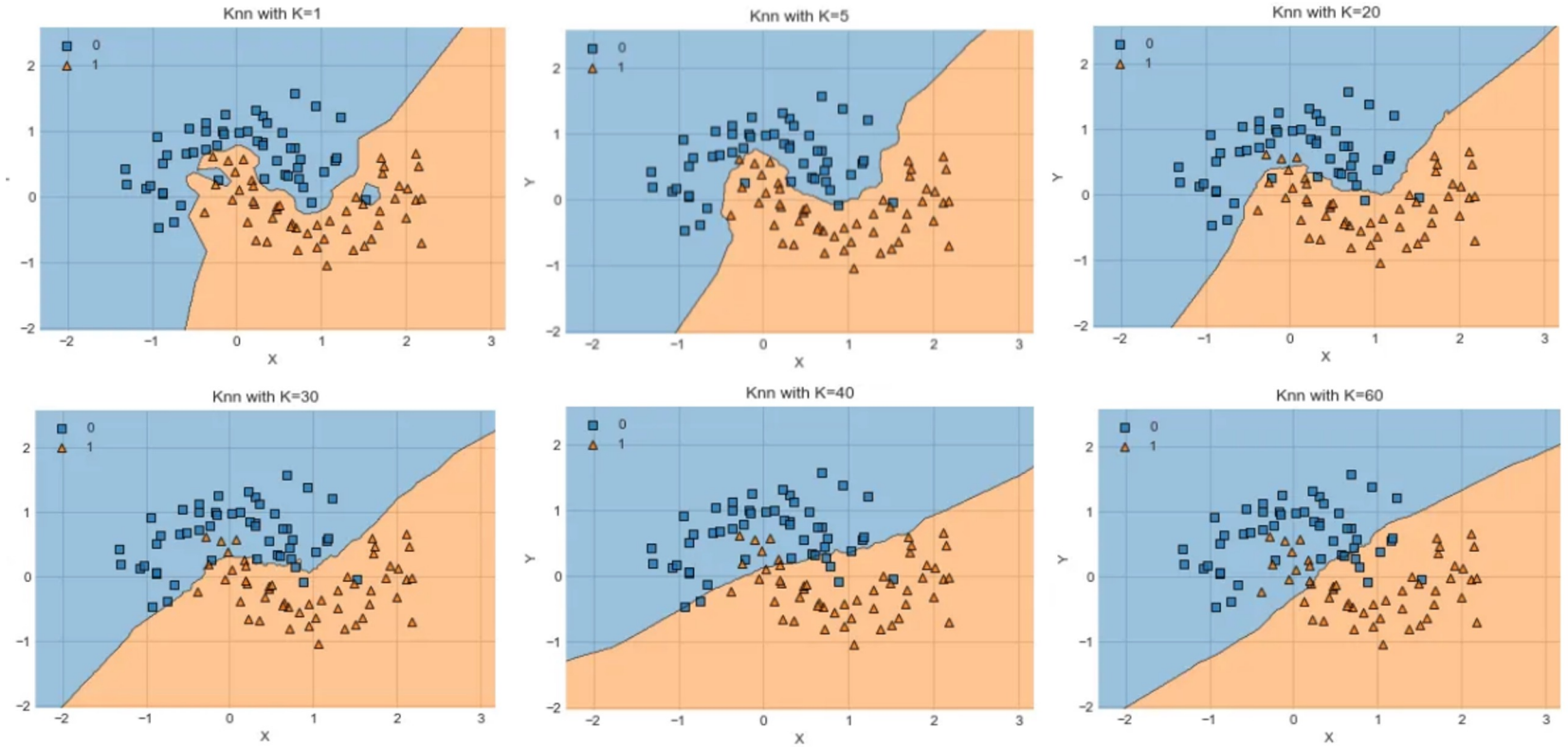
Veoma mala vrednost za K

- *over-fitting*-a i visoke osetljivosti na promene u dataset-u (*high variance*)
- osetljivost na outliers

Suviše velika vrednost za K

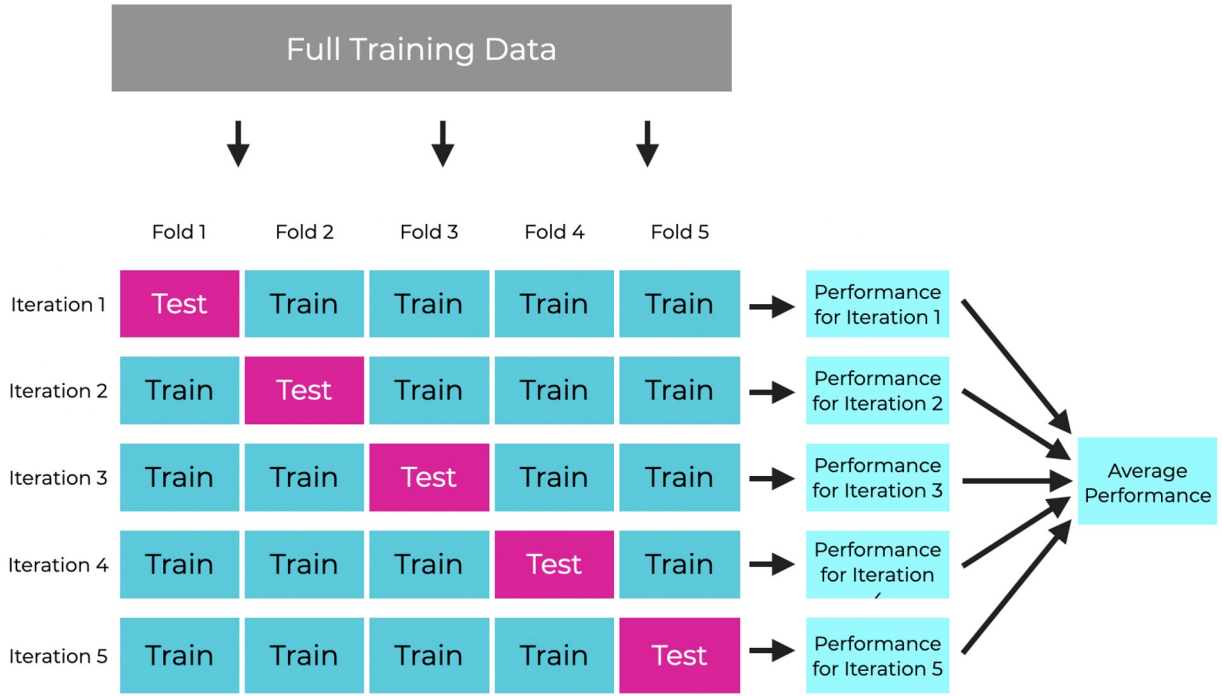
- *under-fitting*
- pristrasnost prema većinskoj klasi

Uticaj različitih vrednosti *k* na predviđanje



Izvor slike: https://towardsdatascience.com/knn-visualization-in-just-13-lines-of-code-32820d72c6b6

Određivanje *k* primenom kros-validacije



Izvor slike: https://www.sharpsightlabs.com/blog/cross-validation-explained/

kNN: Prednosti

***Jednostavnost**: kNN ima malo pretpostavki i zahteva minimalno podešavanje hiper-parametara (svega dva)*

***Neparametrijski**: ne zahteva neku posebnu distribuciju podataka*

***Fleksibilnost**: može se koristiti za zadatke klasifikacije i regresije*

***Interpretabilnost**: predviđanja se mogu lako protumačiti*

***Bez faze treninga**: skup podataka za obuku se koristi direktno tokom faze predviđanja*

kNN: Mane / Izazovi

Slabo skaliranje: kako veličina skupa podataka raste, vreme predviđanja se povećava

Osetljivost na opseg vrednosti atributa: atributi sa većim opsegom vrednosti mogu da dominiraju u proračunu udaljenosti

Neadekvatan u slučaju nebalansiranih podataka: većinska klasa ima tendenciju da dominira predviđanjima, što čini izazovom identifikovanje slučajeva manjinske klase

Osetljivost na vrednost k