K NEAREST NEIGHBORS (KNN) KLASIFIKATOR

Jelena Jovanovic

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

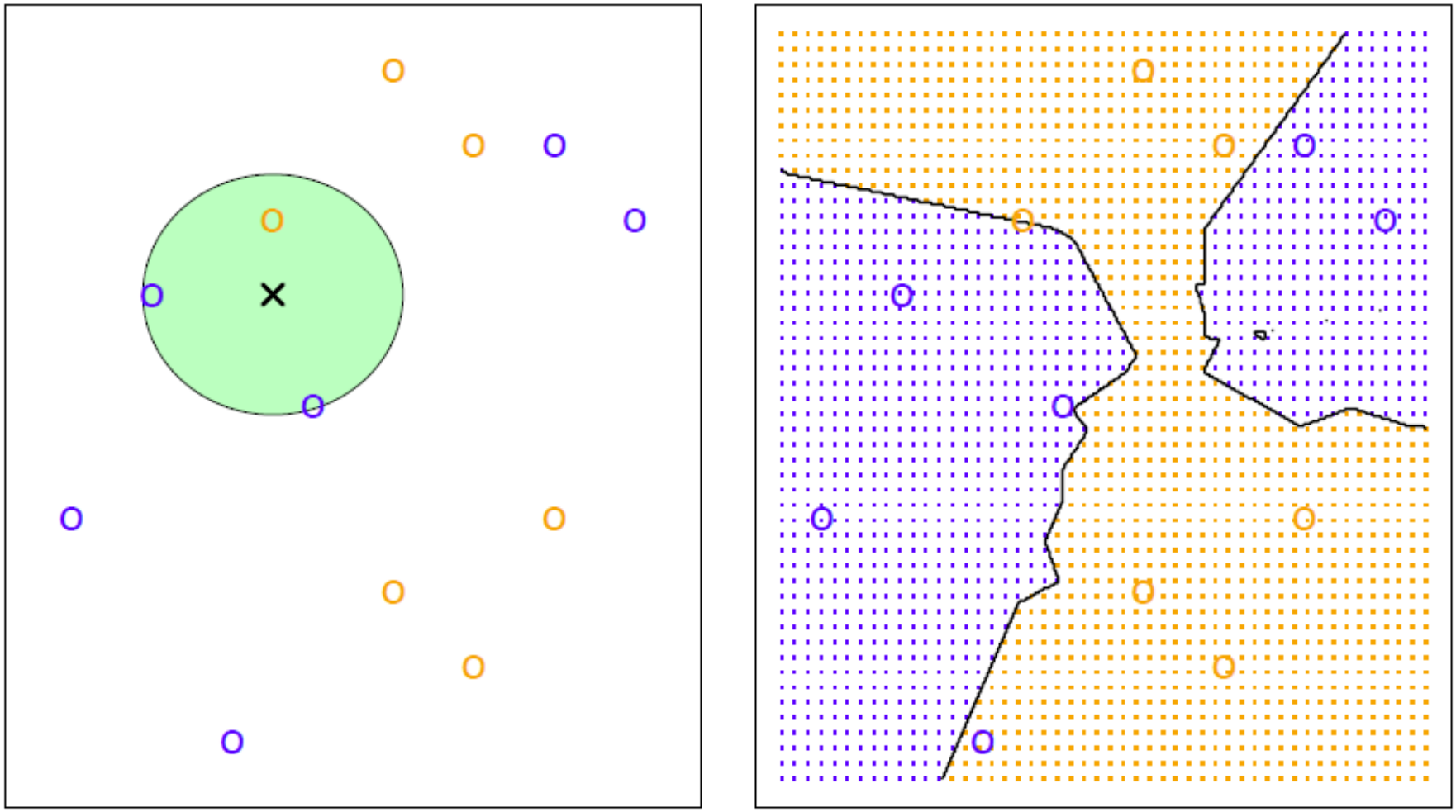
INSTANCE-BASED LEARNING

- kNN spada u kategoriju tzv. *Instance-based* metoda m. učenja
- Zajednička svojstva ovih metoda:
  - Sve instance iz skupa za trening se čuvaju u memoriji
  - Kada se pojavi nova instanca koju je potrebno klasifikovati
    - Pronalazi se *k* trening instanci koje su "najbliže" novoj instanci
    - Klasa nove instance se procenjuje na osnovu *k* najbližih trening instanci

K NEAREST NEIGHBORS (KNN)

- Jedna od najbazičnijih *Instance-based* metoda m. učenja
- Posmatra sve instance kao tačke u *n*-dimenzionalnom prostoru
  - *n* je broj atributa kojima su instance opisane
- Koristi odgovarajuću metriku za računanje blizine/udaljenosti instanci
- Klasifikuje instancu tako što bira najpopularniju klasu među najbližim susedima te instance

KNN PRIMER ZA K = 3



Slika preuzeta iz knjige “An Introduction to Statistical Learning”, str. 40

METRIKE ZA PROCENU 'BLIZINE' INSTANCI

Euklidska distanca

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

- x_i i x_j su instance čija se udaljenost računa
- $a_1(x), a_2(x), \ldots, a_n(x)$ je vektor atributa (*features*) kojim je svaka instanca predstavljena

METRIKE ZA PROCENU 'BLIZINE' INSTANCI

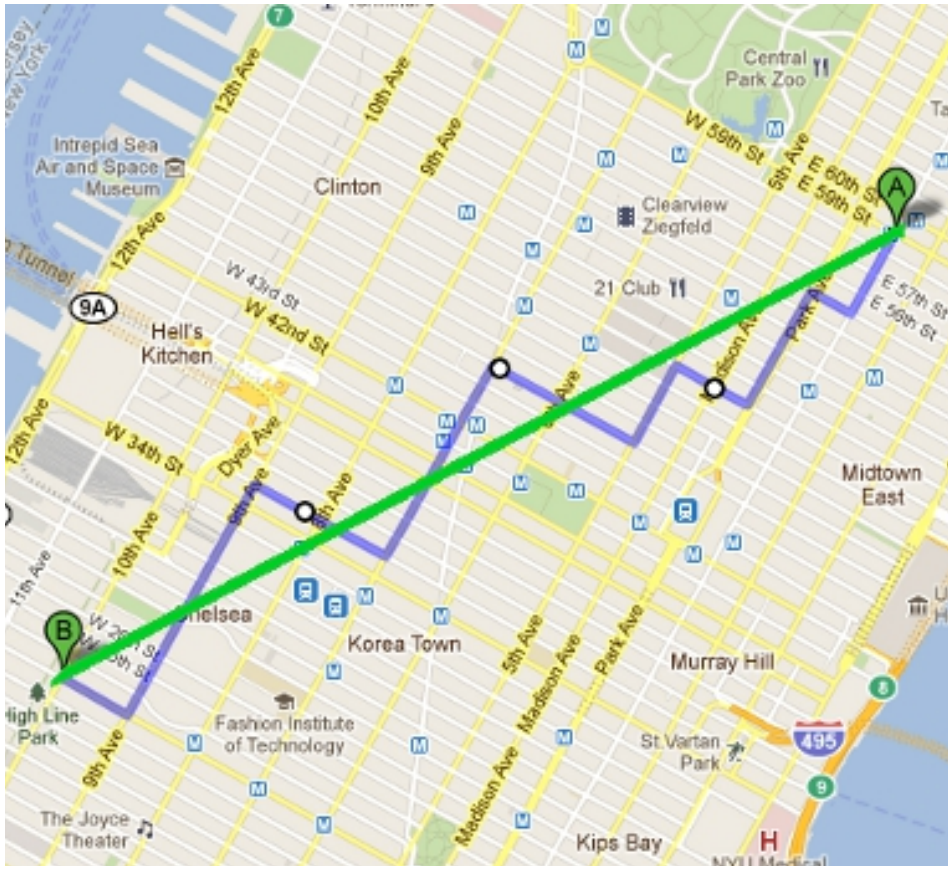
Manhattan (taxi-cab) distanca

$$d(x_i, x_j) = \sum_{k=1}^{n} |a_{ik} - a_{jk}|$$

x_i i x_j su instance;

$a(x_i)$ i $a(x_j)$ su vektori atributa (*features*) koji opisuju x_i i x_j

Euklidska vs Manhattan distanca



Izvor slike:
http://blog.csdn.net/kikitamoon/article/details/42119415

METRIKE ZA PROCENU 'BLIZINE' INSTANCI

- Euklidska i Manhattan metrika su najpopularnije, ali, zavisno od konkretnog problema klasifikacije, koristite se i druge metrike, izvedene iz ovih osnovnih
- Na primer, u slučaju da želimo da dodatno istaknemo velike razlike između instanci, a one manje stavimo u drugi plan, možemo koristiti metrike sa višim stepenom
  - Npr. metrike koje računaju sumu razlika atrubuta podignutih na 3. ili 4. stepen

NORMALIZACIJA VREDNOSTI ATRIBUTA

- Vrednosti različitih atributa su najčešće izražene u različitim skalama
  - Npr., površina, broj prostorija i procenjena vrednost (cena) stana
- Ukoliko bi direktno primenili Euklidsku ili neku drugu metriku, atributi sa višim opsegom vrednosti (cena) bi potpuno potisnuli atribute čiji se vrednosti kreću u značajno manjem opsegu (broj prostorija)
- Normalizacijom se vrednosti svih atributa svode na opseg [0,1]
- Tipičan pristup normalizaciji vrednosti atributa:

$$a_i = \frac{v_i - min\ v_i}{max\ v_i - \min v_i}$$

gde je v_i vrednost atributa a_i

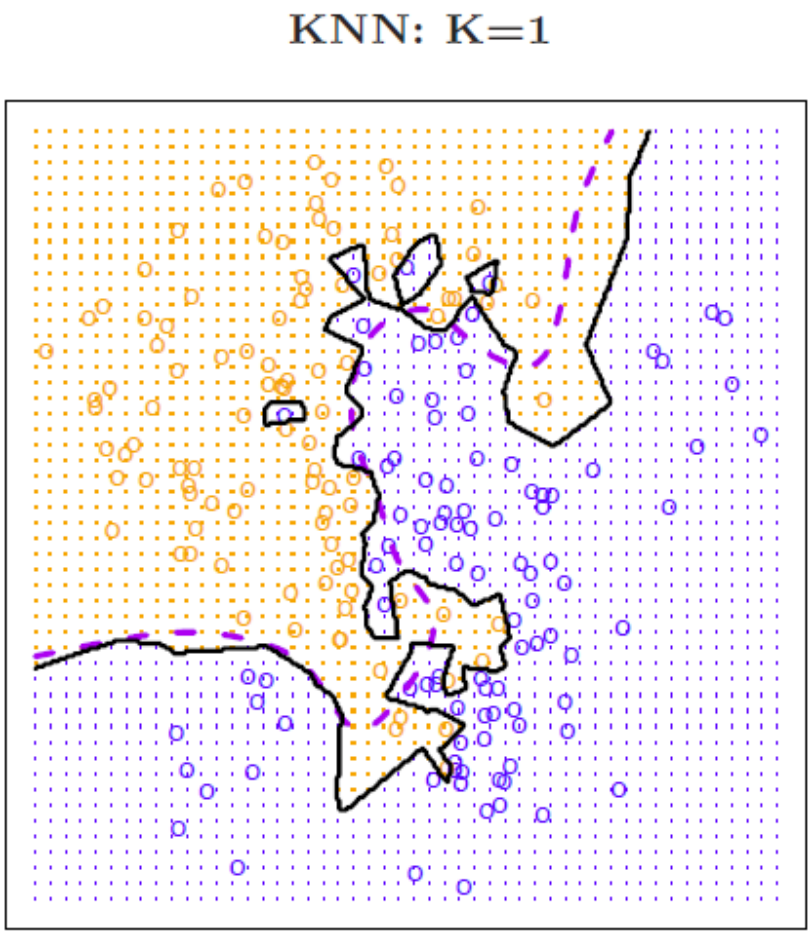
NOMINALNI ATRIBUTI

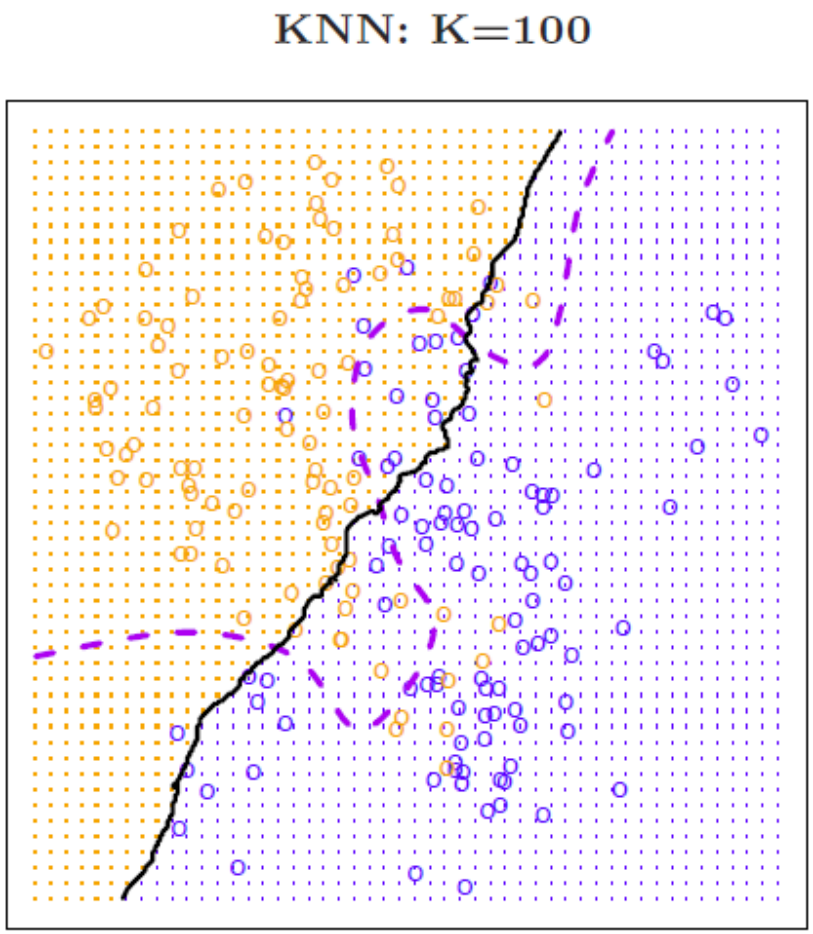
- U slučaju nominalnih atributa, tj. atributa čije su vrednosti simboličke, a ne numeričke, razlika vrednosti atributa a_p dveju instanci x_i i x_j je
  - 1, ukoliko $a_{pj} \neq a_{pi}$
  - 0, ukoliko $a_{pj} = a_{pi}$

KAKO ODREDITI PARAMETAR K?

- Postupkom kros-validacije pronaći *K* koje garantuje dobre performance klasifikatora, a pri tome ne dovodi do problema *over-fitting*-a
- Najčešće se za K bira neparan broj, kako bi se bez dileme mogla izabrati dominantna klasa među *K* najbližih suseda
- Generalno, što *K* ima manju vrednost, to je kNN metoda fleksibilnija, odnosno sklonija *over-fitting*-u

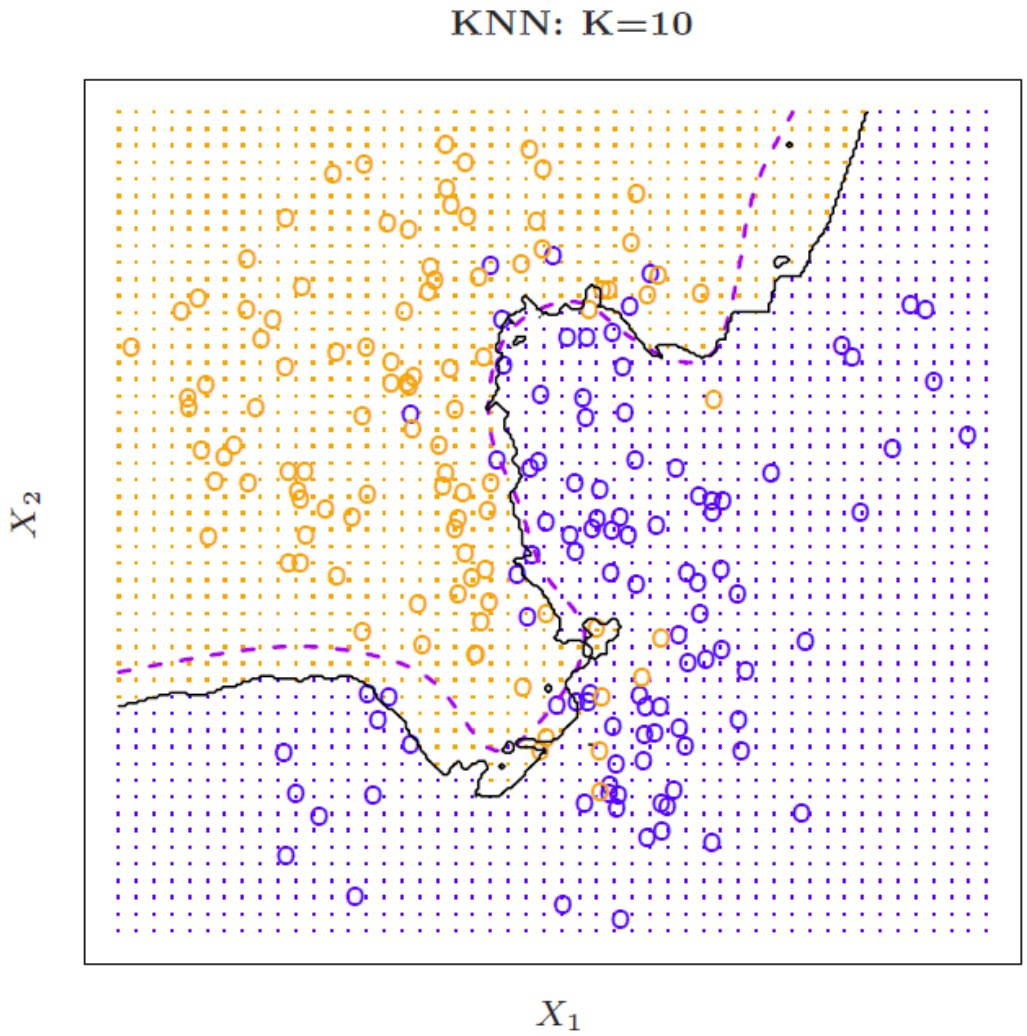
PRIMER OVER-FITTING-A (K = 1) I UNDER-FITTING-A (K = 100)





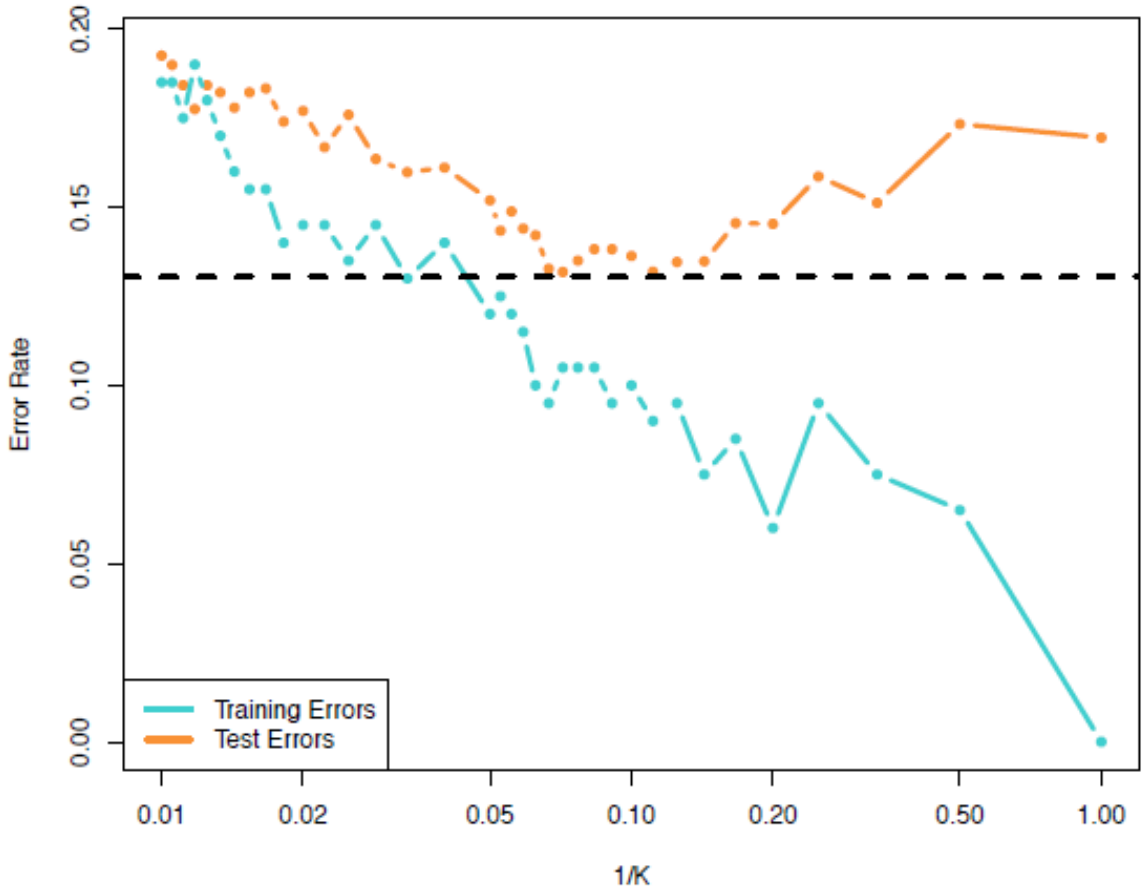
Slika preuzeta iz knjige “An Introduction to Statistical Learning”, str. 41

PRIMER OPTIMALNE VREDNOSTI ZA K (K = 10) ZA DATI SKUP PODATAKA



Slika preuzeta iz knjige “An Introduction to Statistical Learning”, str. 41

POREĐENJE GREŠKE NA TRENING I TEST PODACIMA



Kako se *K* smanjuje, tako i greška na trening setu (kontinuirano) opada; međutim, na test setu, to smanjenje greške u jednom trenutku prestaje (K=10) i od tada greška samo raste

Slika preuzeta iz knjige “An Introduction to Statistical Learning”, str. 42

ZAHVALNICE

Ovi slajdovi su delimično zasnovani na:

- Poglavlju 2 knjige “An Introduction to Statistical Learning” (link)
- Poglavlju 4.7 knjige “Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition” (link)
- Prezentaciji “k Nearest Neighbor” preuzetoj sa SlideShare.net (link)