Estimacija GRV i klasifikacija metodom *k* najbližih suseda (eng. *k nearest neighbours – kNN*)

- Estimacija GRV metodom *k* najbližih suseda
- Klasifikacija metodom *k* najbližih suseda
- Uticaj izbora parametra *k*
- Optimizacija kNN klasifikatora

Opšta formulacija neparametarske estimacije GRV

- Opšti izraz za neparametarsku estimaciju gustine verovatnoće je:

$$\hat{p}(\mathbf{x}) = \frac{k_N}{NV}$$

V – zapremina koja obuhvata **x**
N – ukupan broj uzoraka
k_N – broj uzoraka unutar V

- U praktičnoj primeni ovog izraza postoje dva osnovna pristupa
  - Estimacija GRV pomoću kernela (KDE) – fiksira se zapremina V i k_N se odredi na osnovu uzoraka (prebrojavanjem)
  - Estimacija GRV **metodom *k* najbližih suseda** (kNN) – fiksira se vrednost k_N i oko tačke estimacije se formira odgovarajuća zapremina (hipersfera), koja se proširuje sve dok se ne obuhvati k_N uzoraka
    - Jedna od najjednostavnijih metoda kasnog učenja
    - Kao i kod KDE, procena $p(\mathbf{x})$ konvergira ka stvarnoj gustini raspodele verovatnoće kada $N \to \infty$, $V \to 0$, $k_N \to \infty$ i $k_N/N \to 0$
    - Isti pristup može se koristiti i za klasifikaciju i za regresiju

kNN estimacija gustine raspodele verovatnoće

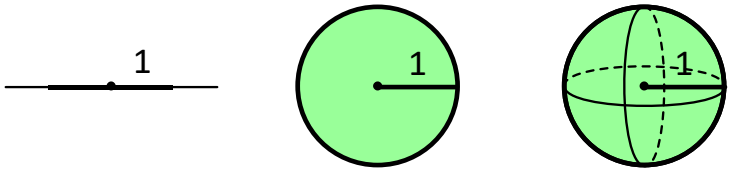
- Kod metode *k* najbližih suseda (kNN) zapremina oblasti oko tačke estimacije **x** proširuje se sve dok se ne obuhvati ukupno *k* uzoraka
- Estimacija gustine verovatnoće tada postaje

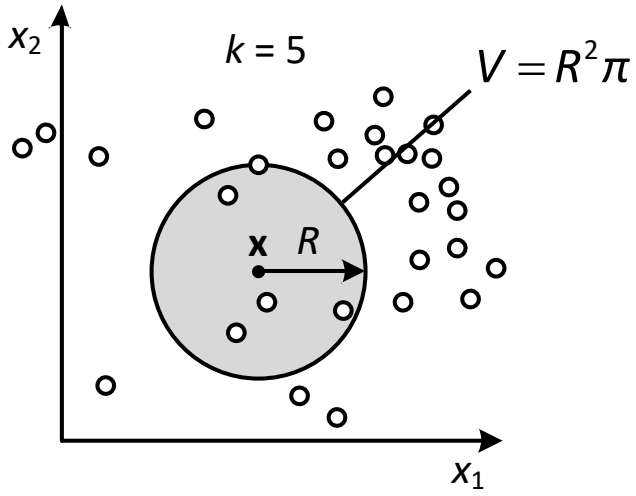
$$\hat{p}_{kNN}(\mathbf{x}) = \frac{k}{NV} = \frac{k}{Nc_D R_k^D(\mathbf{x})}$$

  - $R_k(\mathbf{x})$ je rastojanje između tačke estimacije **x** i njenog *k*-tog najbližeg suseda
  - c_D je zapremina jedinične sfere u *D*-dimenzionalnom prostoru:

$$c_D = \frac{\pi^{D/2}}{(D/2)!} = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$$

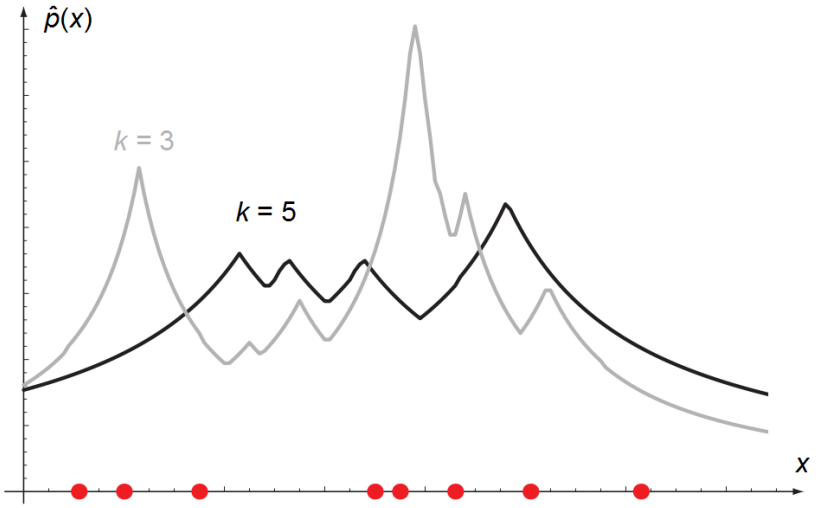
$c_1 = 2$ $\quad c_2 = \pi$ $\quad c_3 = 4\pi/3$





Nedostaci kNN estimacije GRV

- U opštem slučaju estimacija $\hat{p}_{kNN}(\mathbf{x})$ nije u potpunosti zadovoljavajuća
  - Podložna je lokalnom šumu, odnosno, zavisi od tačnog položaja uzoraka
  - Pošto funkcija $R_k(\mathbf{x})$ nije diferencijabilna, estimacija neće biti glatka
  - „Repovi“ estimacije mogu biti veoma izraženi
  - Estimacija ne zadovoljava uslov da je njen integral po celom uzoračkom prostoru jednak 1 (što je osnovni uslov koji svaka GRV treba da zadovolji)
    - Štaviše, integral estimacije po uzoračkom prostoru divergira



Uticaj izbora parametra *k* na procenu GRV

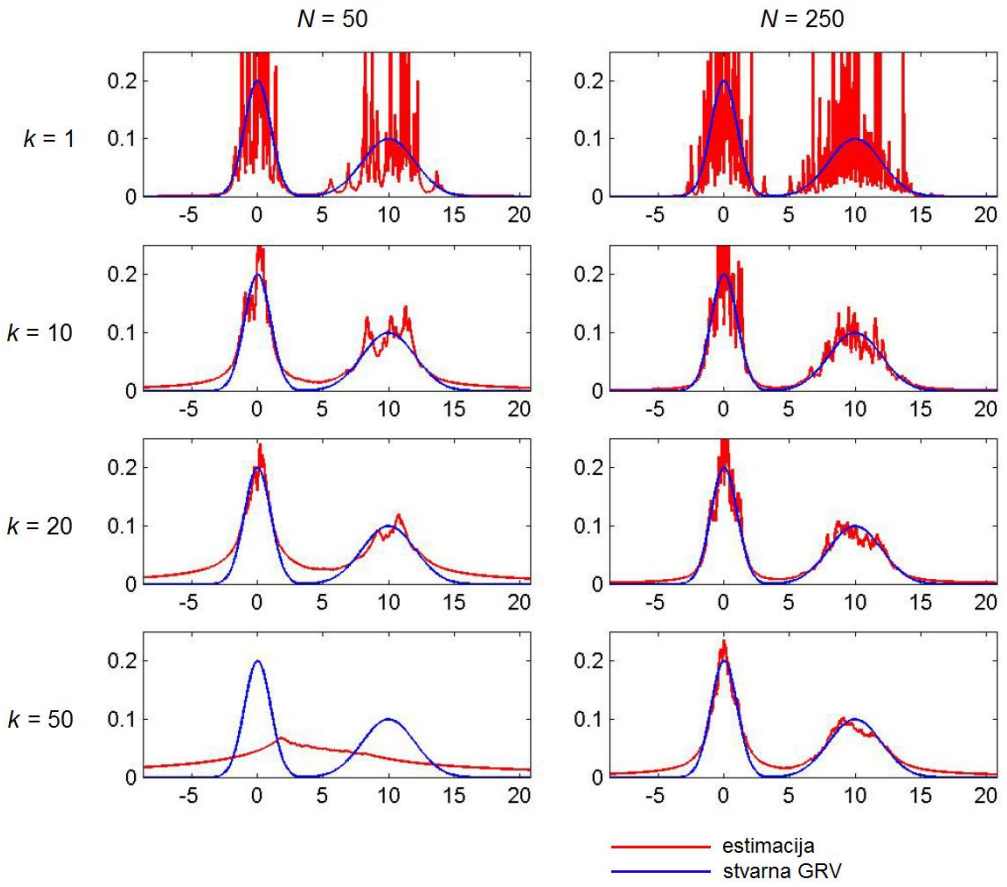
- Za velike vrednosti *k*
  - Uzimaju se u obzir veoma udaljeni uzorci
    - lokalni karakter procene GRV se gubi
  - Izbor velike vrednosti *k* povećava ionako veliku računsku složenost algoritma
- Za male vrednosti *k*
  - Uzimaju se u obzir samo najbliži uzorci, čiji je tačan položaj može biti u velikoj meri posledica slučajnosti (procena GRV sklona je šumu)

kNN estimacija GRV (primer 1)

- Bimodalna raspodela (mešavina dve 1-D Gaussove raspodele):

$$p(\mathbf{x}) = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(10,4)$$

  - Za suviše malo k procena ima nagle skokove jer se $R_k^D(\mathbf{x})$ skokovito menja od tačke do tačke
  - Za suviše veliko k gubi se lokalni karakter estimacije jer se uzimaju u obzir i veoma udaljeni uzorci
  - Estimacija je, naravno, sve približnija stvarnoj $p(\mathbf{x})$ sa porastom broja uzoraka N



kNN estimacija GRV (primer 2)

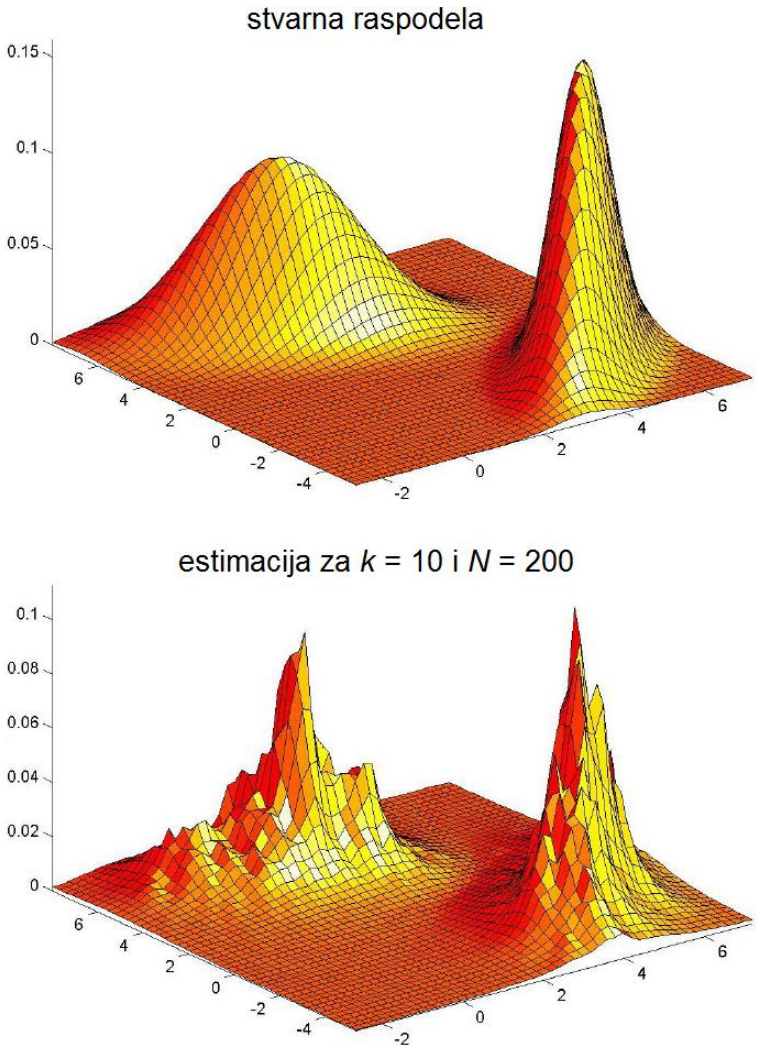
- Bimodalna raspodela (mešavina dve 2-D Gaussove raspodele):

$$p(\mathbf{x}) = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

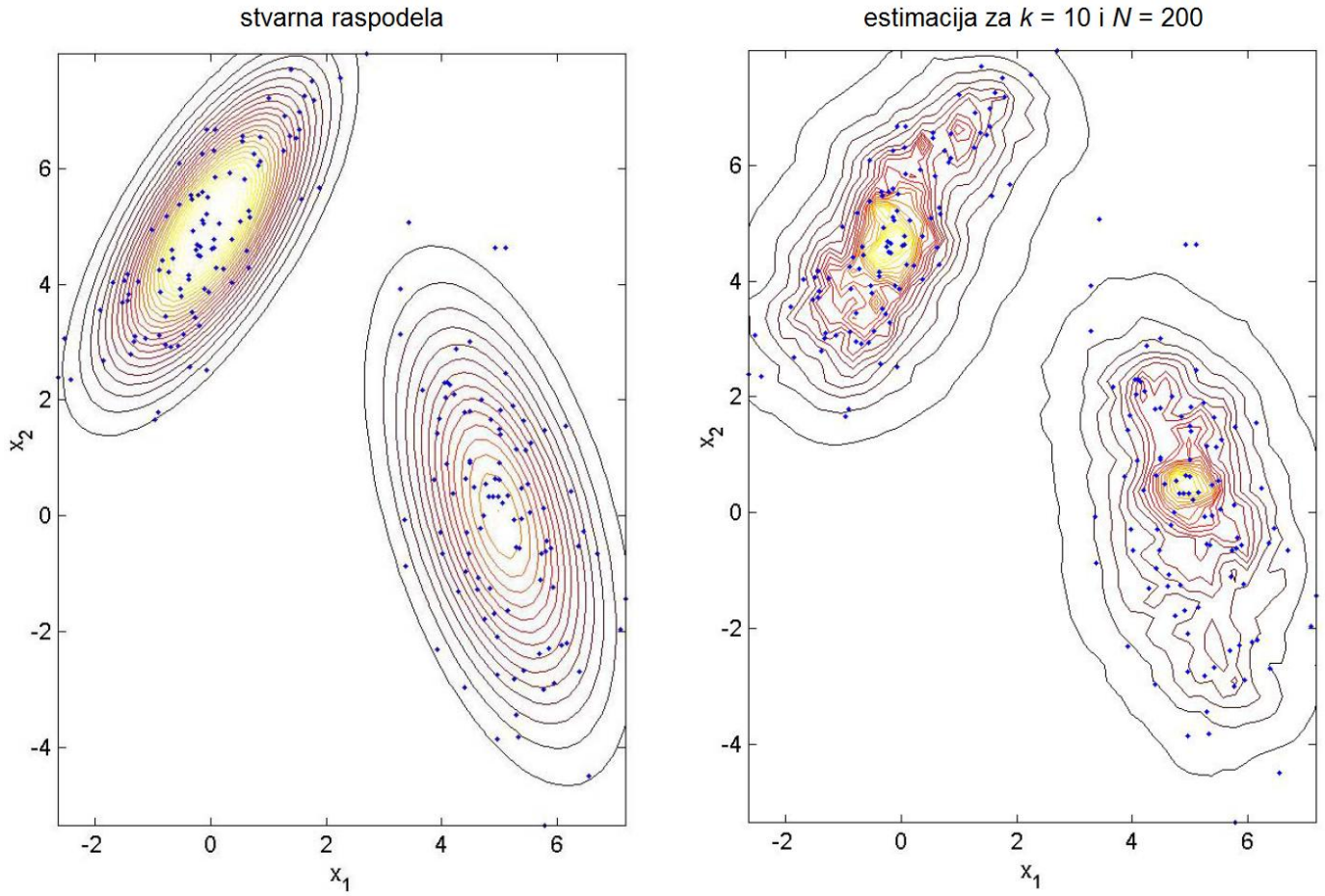
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 5 \end{bmatrix} \qquad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

  - Važe isti zaključci kao i u prethodnom slučaju



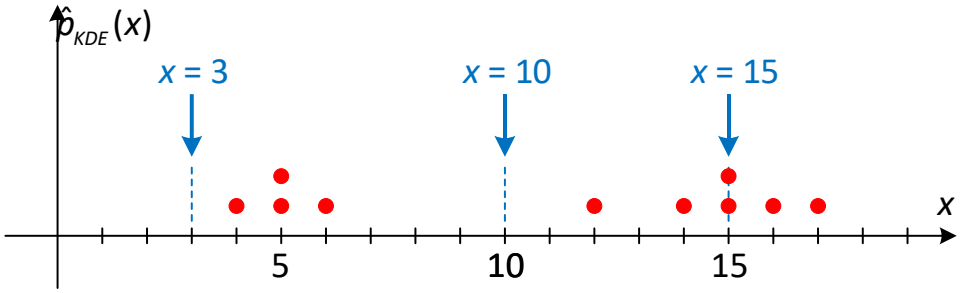
kNN estimacija GRV (primer 2)



Tačke predstavljaju uzorke iz skupa za obuku a linije predstavljaju geometrijska mesta tačaka iste (estimirane) gustine verovatnoće

Primer

- Na osnovu datog skupa uzoraka $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$ estimirati gustinu raspodele verovatnoće $p(x)$ u tačkama $x = 3, 10, 15$ pomoću metode k najbližih suseda za $k = 4$.



$$\hat{p}_{kNN}(x)\big|_{x=3} = \frac{k_N}{Nc_1R_4(3)} = \frac{4}{10 \cdot 2 \cdot 3} = \frac{1}{15}$$

$$\hat{p}_{kNN}(x)\big|_{x=10} = \frac{k_N}{Nc_1R_4(10)} = \frac{4}{10 \cdot 2 \cdot 5} = \frac{1}{25}$$

$$\hat{p}_{kNN}(x)\big|_{x=15} = \frac{k_N}{Nc_1R_4(15)} = \frac{4}{10 \cdot 2 \cdot 1} = \frac{1}{5}$$

Klasifikacija na osnovu kNN

- kNN estimacija GRV omogućuje vrlo jednostavnu aproksimaciju Bayesovog klasifikatora, za koji je pokazano da je optimalan
  - Neka je dat skup od N uzoraka za obuku, pri čemu ih iz svake klase ω_i ima po N_i, i neka je potrebno klasifikovati nepoznati uzorak $\mathbf{x}$
  - Oko uzorka $\mathbf{x}$ postavlja se hipersfera zapremine V, koja sadrži ukupno k uzoraka, od čega ih iz svake klase ω_i ima po k_i
  - Raspodela obeležja za svaku klasu može se estimirati kNN metodom:

$$\hat{p}_{kNN}(\mathbf{x} \mid \omega_i) = \frac{k_i}{N_i V},$$

    dok se bezuslovna gustina raspodele verovatnoće obeležja estimira kao:

$$\hat{p}_{kNN}(\mathbf{x}) = \frac{k}{NV}$$

  - Apriorne verovatnoće se estimiraju kao:

$$\mathrm{P}(\omega_i) = \frac{N_i}{N}$$

  - Bayesov klasifikator može se, prema tome, dobiti na sledeći način:

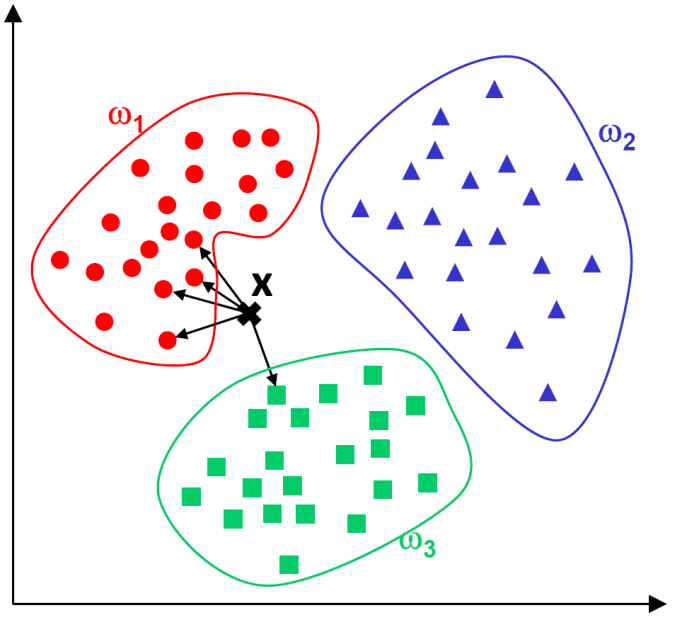
$$\hat{\mathrm{P}}_{kNN}(\omega_i \mid \mathbf{x}) = \frac{\hat{p}_{kNN}(\mathbf{x} \mid \omega_i)\mathrm{P}(\omega_i)}{\hat{p}_{kNN}(\mathbf{x})} = \frac{\frac{k_i}{N_i V} \cdot \frac{N_i}{N}}{\frac{k}{NV}} = \frac{k_i}{k}$$

Klasifikacija na osnovu kNN

- kNN pravilo klasifikacije je intuitivna metoda koja klasifikuje nepoznate uzorke prema klasnoj pripadnosti bliskih uzoraka iz skupa za obuku

> Za dati uzorak $\mathbf{x} \in \mathbb{R}^D$ pronaći k „najbližih“ uzoraka iz skupa za obuku i svrstati $\mathbf{x}$ u klasu koja je među tih k uzoraka najzastupljenija.

- kNN zahteva jedino:
  - izbor celobrojnog parametra k
  - skup za obuku, gde je klasna pripadnost svakog uzorka poznata
  - metriku za utvrđivanje stepena bliskosti dva vektora obeležja
- U primeru sa slike $k = 5$, tako da se nepoznati uzorak svrstava u klasu ●
- Na sličan način realizuje se i regresija
  - Izlazna vrednost uzorka $\mathbf{x}$ predstavlja prosek izlaznih vrednosti njegovih k najbližih suseda

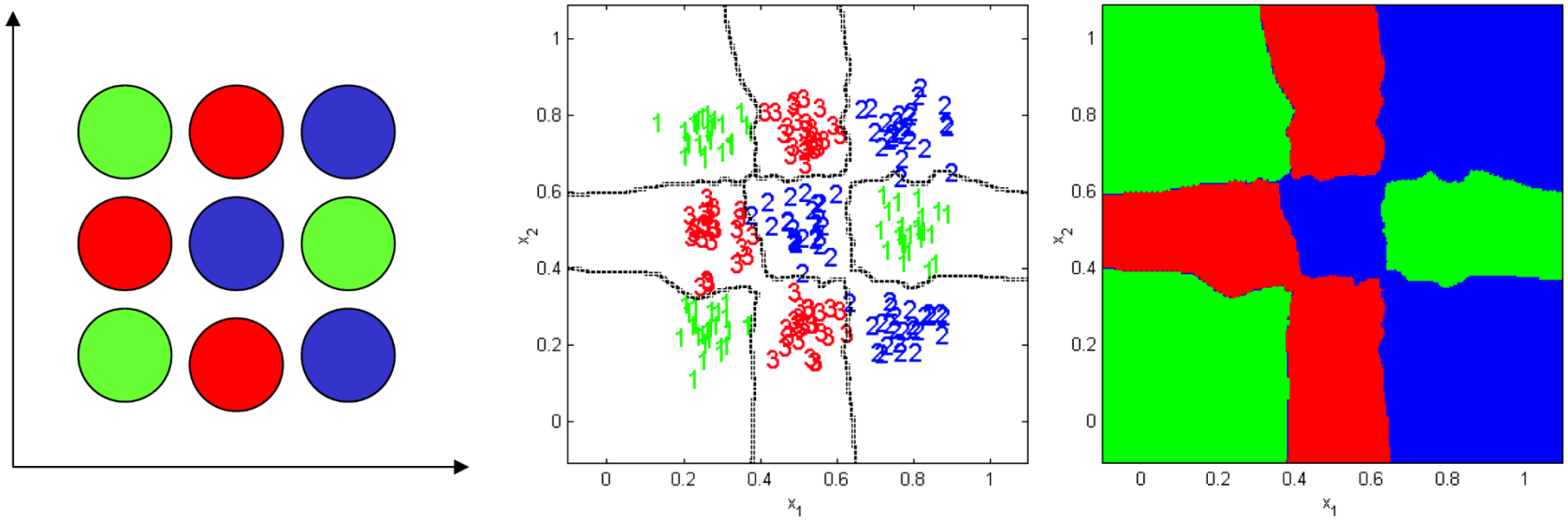


kNN klasifikacija kao metoda kasnog učenja

- Metode mašinskog učenja mogu se načelno podeliti na metode kasnog učenja (eng. *lazy learning*) i metode ranog učenja (eng. *eager learning*)
- Algoritmi sa kasnim učenjem (u koje spada i kNN klasifikacija)
  - Obrada uzoraka za obuku odlaže se do trenutka kada se javi zahtev za klasifikacijom nepoznatog uzorka, dok klasične obuke često i nema
  - U odgovoru na zahtev za klasifikacijom koristi se celokupni skup za obuku
    - Ove metode imaju vrlo visoke zahteve za memorijskim prostorom i računski su veoma zahtevne u toku klasifikacije
  - Nakon klasifikacije svi rezultati do kojih se došlo odbacuju se i sve se radi iznova za svaki naredni zahtev, bez obzira koliko je već puta urađena klasifikacija
- Algoritmi sa ranim učenjem (podsećanje)
  - Uzorci za obuku ne čuvaju se u izvornom obliku, već se prevode u odgovarajući *model*, koji predstavlja njihov komprimovan opis
    - Estimacija parametara gustine verovatnoće
    - Određivanje grafovske strukture sa odgovarajućim težinama (npr. kod neuralnih mreža)
  - Nakon dobijanja modela skup za obuku može se odbaciti jer više nije potreban
    - klasifikacija nepoznatih primeraka vrši se isključivo na osnovu kreiranog modela
  - Obuka može biti izuzetno složena, ali je klasifikacija računski mnogo manje zahtevna
  - Postoji opasnost od loše formiranog modela

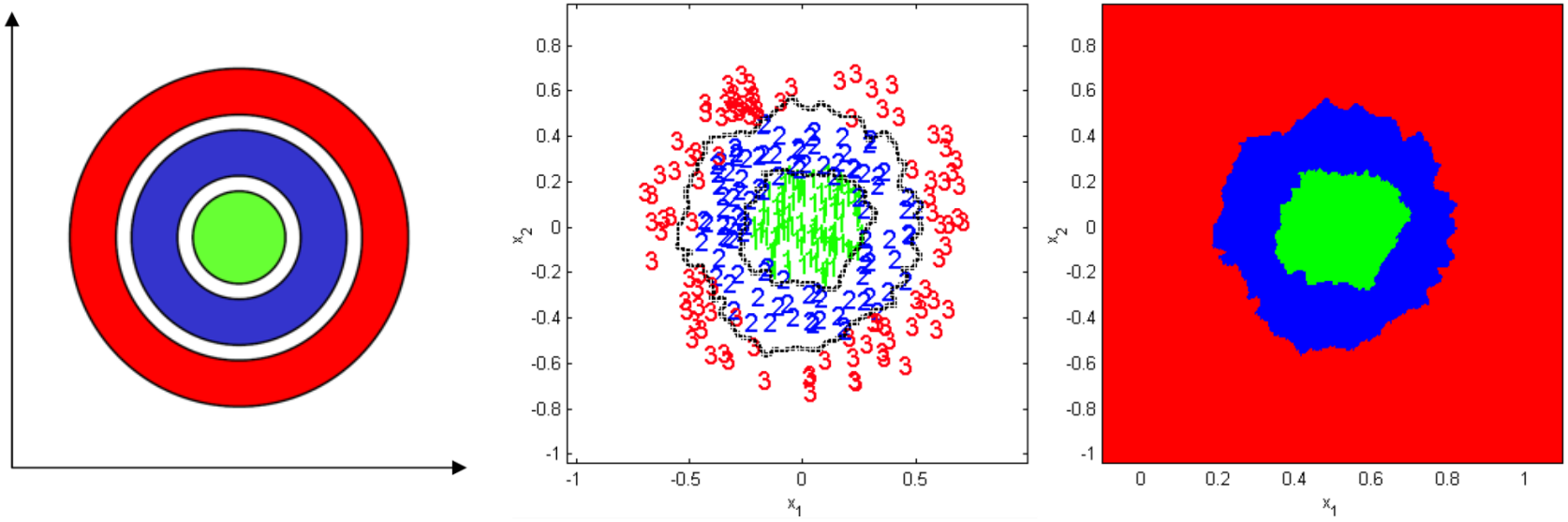
Klasifikacija na osnovu kNN (primer 1)

- Tri multimodalne nelinearno separabilne klase
- kNN pravilo sa k = 5 i euklidskim rastojanjem kao metrikom



Klasifikacija na osnovu kNN (primer 2)

- Tri nelinearno separabilne klase
- kNN pravilo sa k = 5 i euklidskim rastojanjem kao metrikom

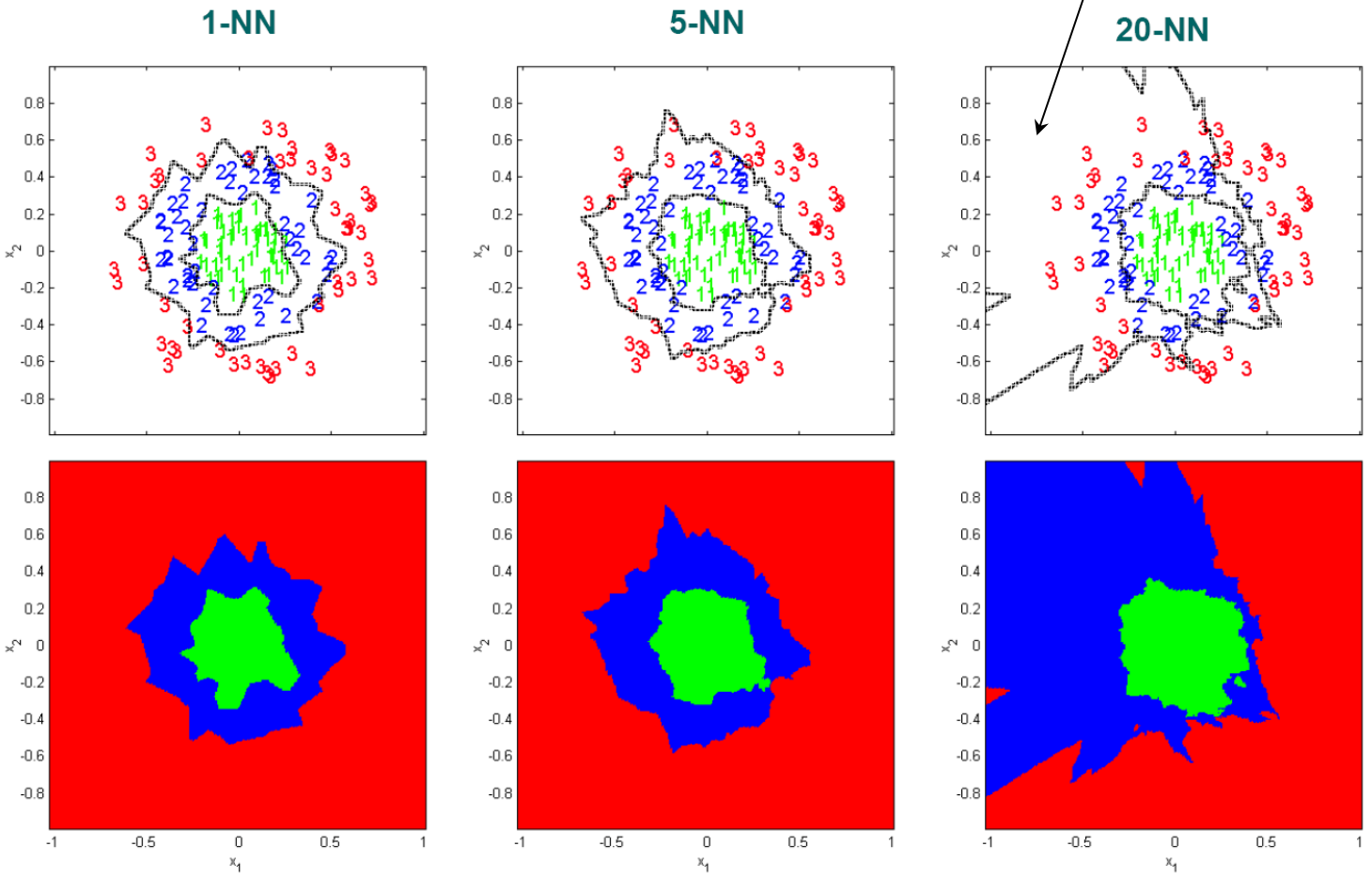


Uticaj izbora parametra *k* na klasifikaciju

- Za velike vrednosti *k*
  - Uzimaju se u obzir veoma udaljeni uzorci
    - tačnost klasifikacije počinje da opada
  - Izbor velike vrednosti *k* povećava ionako veliku računsku složenost algoritma
  - Moguće je dobiti procenu o pouzdanosti odluke klasifikacije (na osnovu odnosa k_i i k)
- Za male vrednosti *k*
  - Uzimaju se u obzir samo najbliži uzorci, čiji je tačan položaj može biti u velikoj meri posledica slučajnosti (granice odlučivanja manje su glatke)
  - Poseban slučaj ove metode za $k = 1$ naziva se *metoda najbližeg suseda* (eng. *nearest neighbour*), i predstavlja izuzetno jednostavan pristup, koji uz dovoljno velik skup za obuku ipak postiže solidne rezultate

Uticaj izbora parametra *k* na klasifikaciju

k ne sme biti preveliko u odnosu na ukupan broj uzoraka!

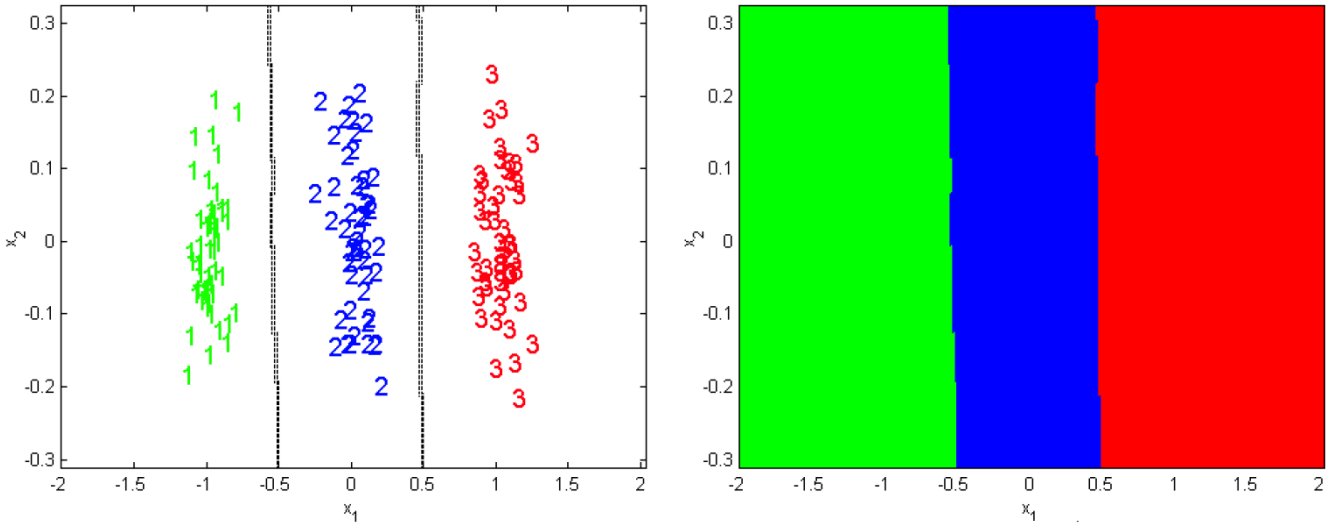


Karakteristike kNN klasifikatora

- Prednosti
  - Jednostavan za implementaciju i interpretaciju
  - Blago suboptimalan za velike skupove za obuku
    - Ako je P_B verovatnoća greške Bayesovog klasifikatora, već i 1-NN klasifikator za veliko N ima grešku između P_B i $2P_B$, dok je za $k > 1$ greška još manja
  - Neosetljiv na složenost stvarnih raspodela po pojedinim klasama
  - Veoma pogodan za paralelnu implementaciju
- Nedostaci
  - Zahteva velik skladišni prostor
  - Zahteva obimna izračunavanja prilikom klasifikacije
    - Potrebno je identifikovati koji uzorci predstavljaju k najbližih (naročito problematično za visokodimenzionalne prostore)
    - Proces je sve složeniji i sporiji što je broj uzoraka za obuku veći (dakle, što tačniji to sporiji)
  - Sa smanjenjem skupa za obuku performanse mogu drastično opasti
    - Pored generalnih tehnika koje se primenjuju kada je skup za obuku mali (augmentacija podataka, redukcija dimenzionalnosti) moguća rešenja obuhvataju izbor metrika optimizovanih za dati skup za obuku, odnosno, za dati problem

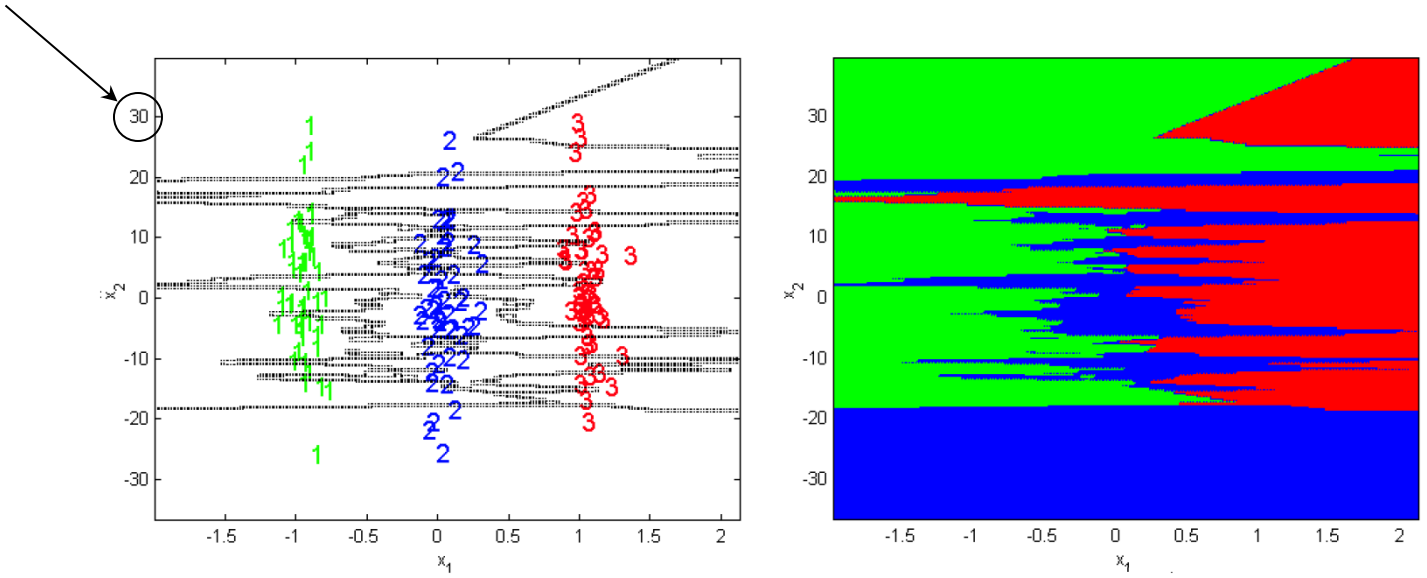
Problem nejednakog uticaja pojedinih obeležja

- Osnovno kNN pravilo utvrđuje bliskost uzoraka na osnovu euklidskog rastojanja
  - Ovo čini kNN metodu veoma osetljivom na zašumljena obeležja
- Primer: 3 klase u 2 dimenzije
  - Prva dimenzija (prvo obeležje) sadrži svu diskriminatornu informaciju, i gledano po ovoj dimenziji separabilnost klasa je odlična, dok drugu dimenziju čini beli šum, koji ne doprinosi klasifikaciji
  - Kada su obeležja približno jednakih dimenzija (istog reda veličine), kNN za k = 5 pronalazi granice odlučivanja izuzetno blizu optimalnih



Problem nejednakog uticaja pojedinih obeležja

- Osnovno kNN pravilo utvrđuje bliskost uzoraka na osnovu euklidskog rastojanja
  - Ovo čini kNN metodu veoma osetljivom na zašumljena obeležja
- Primer: 3 klase u 2 dimenzije
  - Prva dimenzija (prvo obeležje) sadrži svu diskriminatornu informaciju, i gledano po ovoj dimenziji separabilnost klasa je odlična, dok drugu dimenziju čini beli šum, koji ne doprinosi klasifikaciji
  - Ako je irelevantno obeležje većih dimenzija od relevantnog (a ovde je x_2 uvećano za dva reda veličine), kNN sa euklidskim rastojanjem kao metrikom daje veoma loše rezultate



Ponderisanje obeležja

- Prethodni primer je ilustrovao osetljivost kNN klasifikatora na zašumljena obeležja
  - Moguće rešenje bila bi normalizacija svih obeležja tako da bude $\mu_i = 0$, $\sigma_i^2 = 1$
  - Međutim, u višedimenzionalnom prostoru euklidsko rastojanje je veoma zašumljeno ako samo mali broj (od ukupnog broja) obeležja nosi informaciju korisnu za klasifikaciju

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{k=1}^{D} (x_k - x'_k)^2}$$

  - Obeležja se mogu ponderisati spram kvaliteta informacije koju svako od njih pruža:

$$d_w(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{k=1}^{D} (w_k (x_k - x'_k))^2}$$

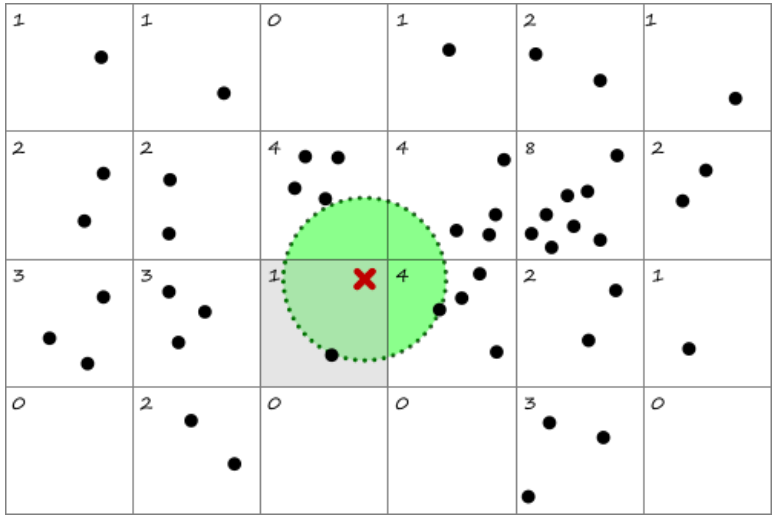
  - Ovaj postupak odgovara linearnoj transformaciji obeležja, pri čemu je transformaciona matrica dijagonalna, sa težinskim faktorima postavljenim na glavnoj dijagonali
    - Ponderisanje obeležja se može posmatrati kao specijalan slučaj izdvajanja obeležja bez njihovog kombinovanja (elementi van glavne dijagonale su jednaki 0)
    - Težinski faktori mogu se otkriti ili iterativnom postupkom, prateći performanse klasifikatora, ili analizom korelacije pojedinih obeležja s oznakom klase (što je brže)
- Ponderisanje obeležja nije isto što i ponderisanje rastojanja (varijanta kNN koja tretira bliže susede kao pouzdanije i dodeljuje im veću težinu)
  - Ponderisanje rastojanja generalno *ne poboljšava* performanse kNN klasifikatora

Računarska složenost kNN metode

- Metoda ima velike zahteve u pogledu računarske složenosti kako pri proceni GRV tako i prilikom klasifikacije
  - Potrebno je naći *k* najbližih suseda datoj tački u prostoru obeležja (tački procene GRV ili nepoznatom uzorku koji treba klasifikovati)
  - Naivan pristup: naći rastojanja do svih suseda, rangirati ih i odabrati *k* najmanjih
    - Vrlo nepraktično za velike vrednosti *N* i *D*
    - Dva standardna pristupa za prevazilaženje ovog problema su tzv. *bucketing* (poznat i kao Eliasov algoritam) i *k*-d (*k*-dimenzionalna) stabla

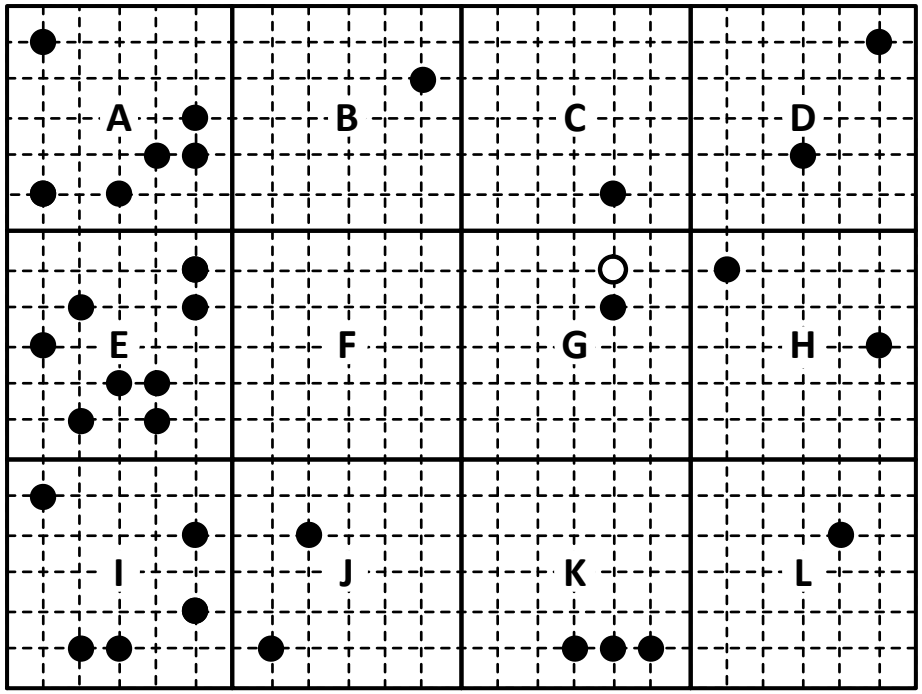
Bucketing

- Prostor je podeljen na identične ćelije, i svi uzorci koje određena ćelija sadrži čuvaju se u listi
- Ćelije se ispituju u redosledu rastućeg rastojanja od tačke **x** čijih *k* najbližih suseda treba identifikovati (gleda se rastojanje od **x** do najbliže tačke ćelije)
- Za svaku ćeliju koja se ispituje računaju se rastojanja od tačke **x** do svih uzoraka unutar ćelije
- Pretraga se završava čim rastojanje od tačke **x** do neke ćelije koju bi trebalo pretražiti postane veće od rastojanja do uzorka koji je trenutno evidentiran kao *k*-ti najbliži sused



Bucketing (primer 1)

- Neka je potrebno klasifikovati uzorak obeležen belim kružićem na osnovu $k = 4$ najbliža suseda iz skupa od $N = 32$ uzorka obeleženih crnim kružićima. Koje će sve ćelije biti ispitane u proceduri pretrage i kojim redosledom?



k-d (*k*-dimenzionalna) stabla

- *k*-d stabla predstavljaju način podele *k*-dimenzionalnog prostora tako da broj uzoraka u svim ćelijama bude otprilike isti
  - Dobijena podela je finija tamo gde je gustina uzoraka veća
- *k*-d stablo konstruiše se na osnovu skupa za obuku:
  - Izabere se proizvoljna dimenzija
  - Identifikuje se uzorak koji predstavlja medijan i prostor se po izabranoj dimenziji podeli na dva potprostora na tom mestu
  - Postupak se iterativno ponavlja na dobijenim potprostorima dok broj tačaka u nekom potprostoru ne padne ispod određenog praga
- *k* najbližih suseda identifikuje se na način sličan kao u slučaju *bucketing*-a