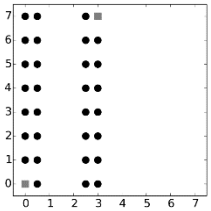
1. Koji od sledećih problema možemo rešiti klasterovanjem?
   a. Predikcija količine kišnih padavina bazirano na relevantnim obeležjima
   b. Obučavanje robota da prođe labirint
   c. Ispitivanje velike kolekcije *spam* poruka u cilju otkrivanja pod-tipova *spam*-a
   d. Dati su podaci o dečijoj visini i godinama. Predvideti visinu deteta u zavisnosti od njegove starosti.
   e. Data je kolekcija od 1000 eseja na temu ekonomije. Automatski grupisati slične eseje.
   f. Dato je 50 članaka napisanih od strane ženskih autora i 50 članaka napisanih od strane muških autora. Obučiti algoritam da određuje pol autora na osnovu teksta.

2. Tačno ili netačno – *K-means* algoritam:
   a. Zahteva da broj obeležja bude manji ili jednak broju primera
   b. Nema garanciju konvergencije
   c. Konvergira u lokalni optimum
   d. Uvek konvergira u globalni optimum
   e. Rezultati *K-means* algoritma veoma zavise on inicijalizacije centroida
   f. Rezultuje najmanjom vrednošću ciljne funkcije kada je K=1
   g. Konvergira u globalni optimum isključivo ako se inicijalni centroidi poklapaju sa nekim od instanci skupa podataka.

3. Dati su 2D podaci na slici. Ako su dve tačke označene kvadratima inicijalni centroidi, nacrtajte klastere nakon jedne iteracije *K-means* algoritma. Da li se rešenje izmeni nakon druge iteracije?




4. Opišite *K-means* algoritam:

5. Da li je neophodno pretprocesirati obeležja skupa podataka pre primene *K-means* algoritma ako upotrebljavamo Euklidsku distancu? Ako da, koja vrsta pretprocesiranja je potrebna?

6. Koji od sledećih algoritama je deterministički (ishodi se ne menjaju ponovnim pokretanjem algoritma):
   a. PCA; b. *K-means*; c. Nijedan od navedenih

7. Svrha *K-means*++ algoritma je da _____. Alternativa je da primenimo sledeći postupak: _____.

8. U *K-means* algoritmu model klastera definišemo putem _____. Posledica ovoga je da rezultujući klasteri imaju _____ oblik.

9. Opišite *K-means++* algoritam. Ovaj algoritam nam garantuje/ne garantuje da ćemo pronaći globalni optimum (zaokružiti).

10. Funkcija koju minimizujemo kod *K-means* algoritma je sledeća (opišite je rečima ili formulom sa označenim simbolima):

11. Ako uvećavamo broj klastera *K*, vrednost ciljne funkcije iz zadatka 10 se smanjuje/uvećava. Zato možemo/ne možemo koristiti vrednost ove ciljne funkcije da odredimo optimalan broj klastera (zaokružite tačno).

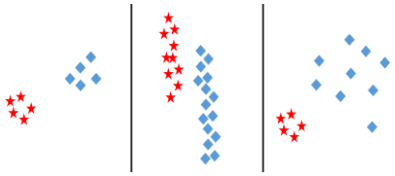
12. Recimo da ste izvršili *K-means* algoritam i dobili da je ciljna funkcija (zadatak 10) za jedno pokretanje K-*means* mnogo veća za *K*=5 nego za *K*=3. Zaokružite jednu opciju:
    a. Ovo nije moguće – mora da postoji greška u kodu
    b. Na osnovu rezultata, zaključujem da je optimalan broj klastera *K*=3
    c. U slučaju *K*=5, *K-means* je završio u lošem lokalnom minimumu. Treba ponovo da pokrenem *K-means* (više nasumičnih inicijalizacija)
    d. U slučaju *K*=3 je pokretanje bilo slučajno srećno. Treba da isprobam različite nasumične inicijalizacije za *K*=3 sve dok performanse za *K*=3 ne budu iste ili gore od slučaja *K*=5.

13. Kod *K-means* klasterovanja se metod „lakta na krivoj“ (*elbow method*) primenjuje radi _____. Ilustrujte kako bi izgledao rezultujući grafik i obeležite njegove ose. Tekstom opišite kako se dobija svaka tačka na grafiku.

14. Pored računarske zahtevnosti, nedostatak *elbow* metode u praksi je _____.

15. Kako evaluiramo rezultat klasterovanja? Objasnite na primeru klasterovanja tekstualnih dokumenata po temi.

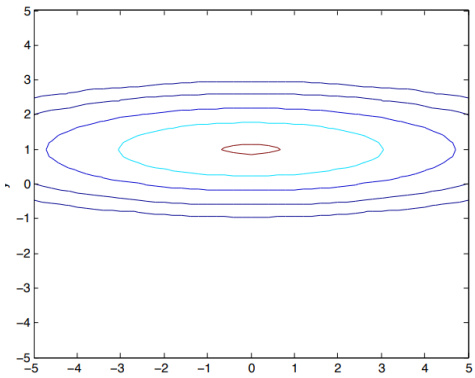
16. Za slučajeve prikazane na slici označite koji algoritmi (od *K-means* bez otežinjenih dimenzija i GMM) su mogli rezultovati dobijenim klasterima.



17. Tačno ili netačno:
    a. Može se desiti da EM algoritam postavi varijanse klastera na 0.
    b. *K-means* se često zaglavljuje u lokalnom optimumu, dok ovo nije veliki problem za EM.
    c. GMM bolje modeluje klastere različitih veličina i orijentacija u odnosu na *K-means*.
    d. GMM bolje modeluje klastere koji se preklapaju od *K-means.*
    e. EM je manje podložan overfitting-u od *K-means*.
    f. *K-means* je identičan GMM modelu gde smo dijagonalne komponente u matrici kovarijanse postavili na vrednosti bliske 0.
    g. GMM model je nenadgledan.
    h. GMM model svaku instancu skupa podataka pridružuje isključivo jednom klasteru.

18. Koje su prednosti GMM (*Gaussian Mixture Model*) algoritma nad *K-means* algoritmom?

19. Šta su $\mu$ i $\Sigma$ Gausove distribucije prikazane na grafiku?

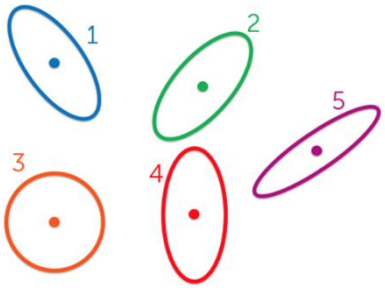


a. $\mu = [0,0]^T, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

b. $\mu = [0,1]^T, \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

c. $\mu = [0,1]^T, \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 0.25 \end{bmatrix}$

d. $\mu = [0,1]^T, \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 0.25 \end{bmatrix}$




20. Recimo da imamo podatke opisane sa 3 obeležja. Postavimo broj klastera na 4 i koristimo pune matrice kovarijanse. Koliko parametara imamo u GMM modelu? Objasnite.

21. Recimo da imamo podatke opisane sa 4 obeležja. Postavimo broj klastera na 5 i pretpostavimo dijagonalne matrice kovarijanse. Koliko parametara imamo u GMM modelu? Objasnite.

22. Koje od kontura na slici kod zadatka 19 opisuju GMM sa matricom kovarijanse koja ima ne-nula elemente samo na glavnoj dijagonali?

23. Šta su u EM (*Expectation-Maximization*) algoritmu E i M koraci?
    a. E – proceniti verodostojnost preko parametara klastera, M – Maksimizovati odgovornosti klastera
    b. E – proceniti odgovornost klastera, M – maksimizovati verodostojnost preko parametara klastera
    c. E – proceniti verodostojnost preko parametara klastera, M – Maksimizovati broj parametara klastera
    d. E – proceniti broj parametara, M – maksimizovati verodostojnost preko parametara klastera.

24. Uopštenije, E i M koraci su:
    a. E – proceniti nedostajuće/latentne varijable skupa podataka; M – maksimizovati verodostojnost preko parametara modela
    b. E – proceniti broj nedostajućih/latentnih varijabli skupa podataka; M – maksimizovati verodostojnost preko parametara modela
    c. E – proceniti verodostojnost preko parametara modela; M – maksimizovati broj nedostajućih/latentnih varijabli u skupu podataka
    d. E – proceniti verodostojnost preko parametara modela; M – maksimizovati broj parametara modela

25. Pretpostavimo da podaci dolaze iz 6 Gausijana (ovo je stvarna struktura podataka). Za koji model, čiji su parametri određeni EM algoritmom, očekujete da će imati najveću verodostojnost:

    a. Mešavina 2 Gausijana; b. Mešavina 6 Gausijana; c. Mešavina 8 Gausijana; d. Mešavina 10 Gausijana.