

CS573: Homework 4

Due date: Tuesday November 16, start of class

1 Between cluster distances (4 pts)

Let cluster C_i contain n_i samples, and let d_{ij} be some measure of distance between two clusters C_i and C_j . In general, one might expect that if C_i and C_j are merged to form a new cluster C_k , then the distance from C_k to some other cluster C_h is not simply related to d_{hi} and d_{hj} . However, consider the equation:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

Show that the following choices for the coefficients $\alpha_i, \alpha_j, \beta, \gamma$ lead to the distance functions indicated.

1. Single-link: $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = -0.5$
2. Complete-link: $\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma = +0.5$
3. Average-link: $\alpha_i = \frac{n_i}{n_i+n_j}, \alpha_j = \frac{n_j}{n_i+n_j}, \beta = \gamma = 0$
4. Between-cluster distance (i.e., squared Euclidean distance between centroids):
 $\alpha_i = \frac{n_i}{n_i+n_j}, \alpha_j = \frac{n_j}{n_i+n_j}, \beta = -\alpha_i \alpha_j, \gamma = 0$

2 Clustering theorem (4 pts)

Read the paper: J. Kleinberg (2002). An Impossibility Theorem for Clustering. In Proceedings of the 16th conference on Neural Information Processing Systems. Explain its main result. Is there a hope for providing a good clustering framework/algorithm?

3 Spectral Clustering (8 pts)

In this problem we will analyze the operation of a variant of spectral clustering methods on two datasets shown in Figure 1. For each of the datasets (unless directed otherwise) please answer the following questions.

1. The first step is to build an affinity matrix. The matrix defines the degree of similarity between points.
 - (a) Suppose we use the L2 norm to construct the following affinity matrix (let x_i denote the i th datapoint):

$$A(i, j) = A(j, i) = \begin{cases} 1 & \text{if } |x_i - x_j|_2 < \theta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

What θ value would you choose and why?

- (b) Suppose instead we use Gaussian kernel for our affinity matrix:

$$A(i, j) = \exp\left(-\frac{|x_i - x_j|_2}{2\sigma^2}\right) \quad (2)$$

What σ value would you choose and why?

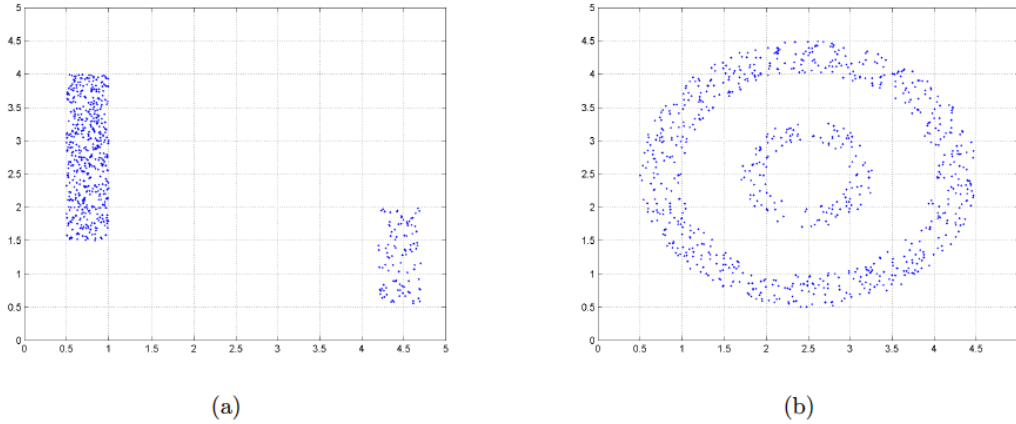


Figure 1: Examples datasets for spectral clustering.

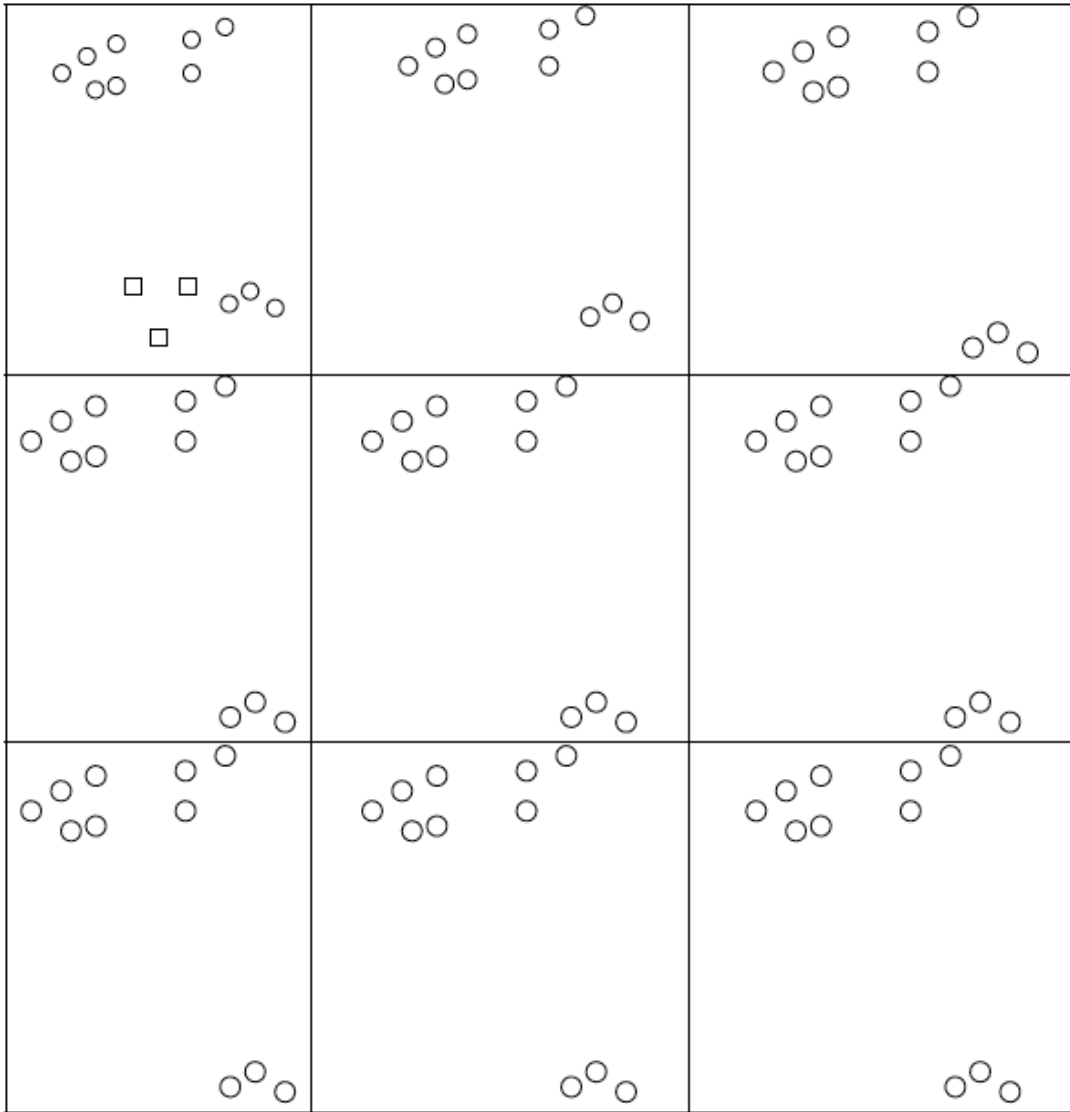
2. The second step is to compute first k dominant eigenvectors of the affinity matrix, where k is the number of clusters we want to have. For the dataset in Figure 1(a) and the affinity matrix defined by equation 1 is there a value of θ for which you can compute analytically eigenvalues corresponding to the first two dominant eigenvectors? If not, explain why not. If yes, compute and write these eigenvalues down, and describe the corresponding eigenvectors.
3. The third step is to construct a matrix Y by placing k dominant eigenvectors (from above) into columns and re-normalizing the rows (to make each row a unit vector). Then we can cluster the rows of Y into k clusters using K-means (or a similar algorithm). For the dataset in Figure 1(a), the affinity matrix defined by equation 1, and the eigenvectors from above, write down your best guess for the coordinates of $k = 2$ cluster centers.
4. Finally, given the clusters on matrix Y , a point x_i is declared to be in cluster j iff the i th row of Y is in cluster j .
 - (a) What are the final clusters you would expect to obtain for each of the datasets? Provide a rough sketch of the clusters to give an idea.
 - (b) What are the clusters you would expect to obtain if using EM algorithm for Gaussian Mixture Models with 2 clusters? Also provide a rough sketch of the clusters.

4 Gaussian Mixture Models and K-means (6 pts)

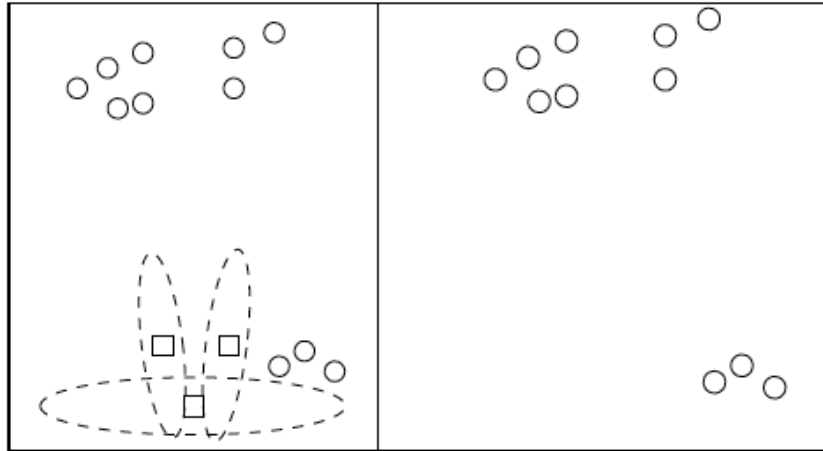
1. You are given a Gaussian mixture model, and all its class probabilities and Gaussian mean locations are learned using EM, but the covariance matrices are forced to be the identity matrices for each class. Rather than using a mixture of Gaussians, you evaluate the probability of each of the K classes given the datapoint and take the the cluster with the highest probability to be the cluster that produced the point. Is this equivalent to doing using a K-means model?
2. Suppose youve done K-means and your K is equal to your number of data points with each cluster defined by a single datapoint. Say that you classify test data points

as part of the cluster that they would belong to according to your distance metric. What model is this equivalent to?

- Run K-means manually for the following dataset. Circles are data points and squares are the initial cluster centers. Draw the cluster centers and the decision boundaries that define each cluster. Use as many pictures as you need until convergence. *Note:* Execute the algorithm such that if a mean has no points assigned to it, it stays where it is for that iteration.



- Now run a Gaussian mixture model of three Gaussians on the same dataset. The initial cluster centers are the same as for the k-means problem, and the dashed-line ellipses represent the size and shape of the initial covariance matrices. Assume that the model puts no restrictions on the form of the covariance matrices and that EM updates both the means and covariance matrices. Draw (approximately) the cluster centers and the size/shape of the covariance matrices of the final converged GMM.



5. Are the group memberships given by the mixture model the same as the memberships given by k-means? Why or why not?

5 Bayes nets (8 pts)

A analyst notices that people that drive SUVs (S) consume large amounts of gas (G) and are involved in more accidents (A) than the national average. They construct the Bayesian network in Figure 2.

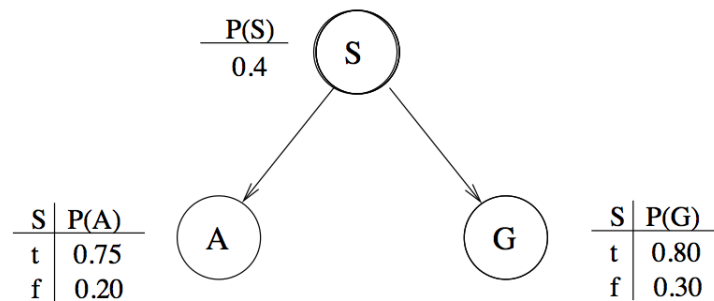


Figure 2: Bayesian network #1.

1. Compute $P(A)$ in two ways:
 - (a) By generating the *entire joint distribution* over these variables and explicitly summing the appropriate entries.
 - (b) Using the variable elimination algorithm.
2. Using conditional independence, compute $P(\neg g, a|s)$ and $P(\neg g, a|\neg s)$. Then use Bayes rule to compute $P(s|\neg g, a)$.

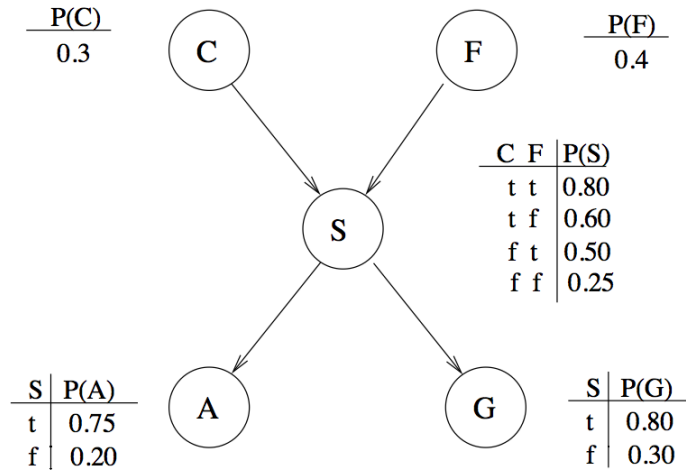


Figure 3: Bayesian network #2.

3. The analyst then notices that there are two types of people that drive SUVs, people from California (C) and people with large families (F). After collecting some statistics, the analyst arrives at the Bayesian network in Figure 3. Using the chain rule from probability, compute the probability $P(\neg g, a, s, c, \neg f)$.
4. Without explicitly generating the entire probability distribution compute $P(s)$ and $P(s|c)$ and $P(s|\neg c)$. (Hint: consider different values of F.)
5. Show how you can compute the value of $P(G)$ in the complete distribution without doing any additional work (i.e., based solely on the work you have already done). Explain.