Internet Analytics (COM-308)

Problem Set 6

Problem 1

In this exercise we verify the two formulas which we used (in the lecture notes and lab) for updating μ_k and Σ_k in the M–step of EM for GMM.

We know that the log–likelihood function for a GMM is

$$L = \log p(X_1, \ldots, X_n | \pi, \mu, \Sigma) = \sum_n \log \sum_k \pi_k N(X_n | \mu_k, \Sigma_k).$$

(a) Show that $\mu_k = \frac{1}{N_k} \sum_n \gamma_{nk} x_n$ with $\gamma_{nk} = \frac{\pi_k N(X_n|\mu_k, \Sigma_k)}{\Sigma_j \pi_j N(X_n|\mu_j, \Sigma_j)}$ and $N_k = \sum_n \gamma_{nk}$ is the solution to $\nabla_{\mu_k} L = 0$.

(b) Show that $\Sigma_k = \frac{1}{N_k} \sum_n \gamma_{nk}(X_n - \mu_k)(X_n - \mu_k)^\intercal$ is the solution of $\nabla_{\Sigma_k} L = 0$.

Hint: Use $\frac{\partial \log|\det(X)|}{\partial X} = (X^{-1})^\intercal = (X^\intercal)^{-1}$ and $\frac{\partial \mathrm{Tr}(AX^{-1}B)}{\partial X} = -(X^{-1}BAX^{-1})^\intercal$.

(c) Consider what happens when we fit a GMM to a set of data points in such a way that one cluster (mixture component) "focuses" on only one of the points. What happens to Σ_k for this cluster, and how does the likelihood evolve?

Problem 2

(a) Recall that the modularity of a graph is given by

$$Q = \frac{1}{2m} \sum_{i=1}^{|C|} \sum_{u,v \in c_i} \left(\mathbf{1}_{u,v} - \frac{d_u d_v}{2m} \right),$$

where $\mathbf{1}_{u,v} = 1$ if there is an edge between nodes u and v, d_u is the degree of node u, m is the number of edges in the graph, and c_i is the set of nodes in community i.

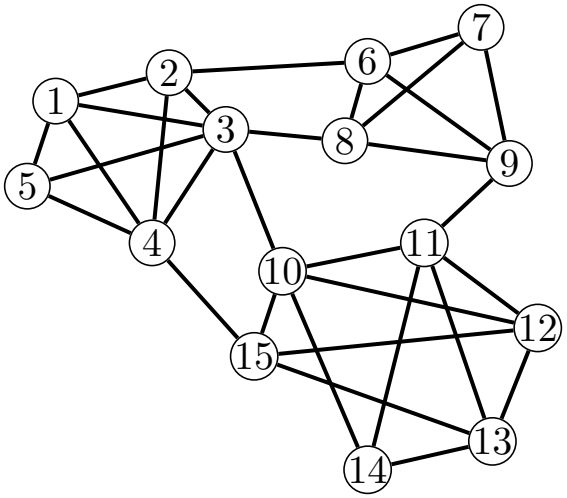
Show that the above definition of modularity can also be expressed more compactly via the following useful formula:

$$Q = \sum_{i=1}^{|C|} \left[\frac{a_i}{m} - \left(\frac{b_i}{2m} \right)^2 \right].$$

Explain what the values of a_i and b_i are.

Hint: Recall that $(\sum_i x_i)^2 = \sum_i \sum_j x_i x_j$.

(b) Find the partitioning which maximizes the modularity of the graph G in Figure 1. What is the maximum modularity value Q?



Graph G

(c) We want to increase the value of modularity by removing one edge from graph G. Guess the edge whose deletion results in the largest increase, and compute the new Q.