STAT 535 Homework 7
Out December 4, 2020
Due December 10, 2020
©Marina Meilă
mmp@stat.washington.edu

Problem 1 – K-means facts

a. With the notation from the lecture notes, prove that the K-means loss can be written as

$$\mathcal{L}(\Delta, \mathcal{D}) = \frac{1}{2} \operatorname{trace}(DX) \tag{1}$$

where $D_{ij} = \|x_i - x_j\|^2$, $X = X(\Delta)$ is the clustering matrix defined on page 19 of `l15-nov19` and $\mathcal{L}(\Delta, \mathcal{D})$ is defined in equation (2). You will need to prove equation (4) along the way (do not assume it's a given fact).

b. Assume that $n = mK$ and that the data set contains K equal clusters. You initialize by picking K data points at random, and assigning them to centers $\mu_{1:K}$. For simplicity, we write this as $\mu_{1:K} \sim uniform(\mathcal{D})$. What is the probability that each of the K clusters contains exactly 1 μ_k? Compute the numerical values of this probability for $K = 2, 5, 10$. Assume sampling is *with replacement* (a simplification that everyone uses in practice).

c. – Where is $K \ln K$ coming from? Now you pick $K' = cK \ln K$ centers at random i.e. $\mu_{1:K'} \sim uniform(\mathcal{D})$ (still with replacement), where $c > 1$ is a constant. Calculate the probability that the first cluster contains none of the K' centers (another simplification), using the approximation $(1 + \frac{1}{z})^z \approx e$ for $z > 0$ large. What is the value you obtain (should be a simple formula depending on K and c).

Problem 2 – Stochastic Block Model

The Stochastic Block Model (SBM) is a probabilistic model for networks (i.e. undirected graphs). For a graph with n nodes, the probability of an edge between $i, j \in V = \{1, 2, \dots n\}$ is $S_{ij} \in [0, 1]$. Hence $S = [S_{ij}]$ can be considered a similarity matrix. In a SBM, we assume that there are K clusters, not necessarily equal, and that

$$S_{ij} = \begin{cases} a, & \text{if } i, j \text{ in the same cluster} \\ b, & \text{if } i, j \text{ in different clusters.} \end{cases} \tag{2}$$

If one samples edges with probabilities given by S, one obtains a random graph. In this problem, we will work only with the matrix of probabilities S; we will show that given S, we can recover the clusters by spectral clustering. *Spectral Clustering is also a method to recover clusters from a graph sampled from S.*

a. Show that if i, j are in the same cluster, their degress are the same, i.e. $D_i = D_j$. Moreover, show that if $D_i = D_j$ when $i \in C_k$ and $j \in C_{k'}$, then $n_k = n_{k'}$; the notation is the same as in the lecture notes.

[b. – Extra credit] Show that the matrix $P = D^{-1}S$ is lumpable. What is rank P?

c. Calculate the probability π_{C_k} of the Markov Chain defined by P being in cluster C_k. Calculate the probability $\hat{P}_{kk'}$ of a transition from C_k to $C_{k'}$, given that the Markov chain is in C_k. For simplicity, assume all clusters have equal sizes.

[**d. − Extra credit**] Show that for $a > b$, the eigenvalues of $\hat{P}$ are strictly greater than 0 (assume all cluster sizes equal). Find the eigenvectors of $\hat{P}$ as functions of a, b, K.