Homework 11

Due May 1 at 11pm

Unless stated otherwise, justify any answers you give. You can work in groups, but each student must write their own solution based on their own understanding of the problem.

When uploading your homework to Gradescope you will have to select the relevant pages for each question. Please submit each problem on a separate page (i.e., 1a and 1b can be on the same page but 1 and 2 must be on different pages). We understand that this may be cumbersome but this is the best way for the grading team to grade your homework assignments and provide feedback in a timely manner. Failure to adhere to these guidelines may result in a loss of points. Note that it may take some time to select the pages for your submission. Please plan accordingly. We suggest uploading your assignment at least 30 minutes before the deadline so you will have ample time to select the correct pages for your submission. If you are using LaTeX, consider using the minted or listings packages for typesetting code.

1. (Kernel k-means) The goal of this exercise is to show that we can "kernelize" Lloyd's algorithm. Let $x_1, \ldots, x_n$ be n points in a set $\mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated feature map $\Phi : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is the RKHS of K. Consider Lloyd's algorithm applied to the points $\Phi(x_1), \ldots, \Phi(x_n)$ with k centroids. We initialize the centroids as $y_1^0, \ldots, y_k^0$ being equal to $\Phi(x_1), \ldots, \Phi(x_k)$.

   (a) For $t \geq 1$, let I_l^t and n_l^t be defined as in Algorithm 1 in the lecture notes. For $t = 0$, we let $I_l^0 = \{l\}$ and $n_l^0 = 1$. The iterates of Lloyd's algorithm then satisfy for every $t \geq 1$, $y_l^t = \frac{1}{n_l^{t-1}} \sum_{i \in I_l^{t-1}} \Phi(x_i)$ (this follows from the definition of the algorithm and you do not have to prove this). Show that for any $t \geq 1$, the squared distance between a point $\Phi(x_j)$ and a centroid y_l^t is equal to

$$\|y_l^t - \Phi(x_j)\|_{\mathcal{H}}^2$$
$$= \frac{1}{(n_l^{t-1})^2} \sum_{i,i' \in I_l^{t-1}} (K(x_i, x_{i'}) + K(x_j, x_j) - K(x_i, x_j) - K(x_{i'}, x_j)).$$

(b) Show by induction on $t \geq 0$ that we can compute all the clusters I_l^t without ever having to compute any of the vectors $\Phi(x_i)$ (but only the kernel values $K(x_i, x_j)$).

2. In class, we stated that small eigenvalues of the Laplacian matrix of some graph $\mathcal{G}$ correspond to clusters in $\mathcal{G}$. The goal of this exercise is to check this intuition on a toy example. Let E_n be the $n \times n$ matrix with 1 in every entry. Consider the graph $\mathcal{G}$ with $2n$ vertices and weight matrix
$$W = \begin{pmatrix} E_n & qE_n \\ qE_n & E_n \end{pmatrix},$$
where $q \in [0, 1]$ is a parameter.

   (a) Depending on the value of q, what is the number of connected components of $\mathcal{G}$?

   (b) Assume that $q = 0$. Give an orthonormal basis of the eigenspace of L corresponding to the eigenvalue 0.

   (c) Same question for $q > 0$.

   (d) Assume that $q > 0$. Let $e \in \mathbb{R}^n$ be the vector with ones in all of its entries. Show that the vector $u = (e, -e) \in \mathbb{R}^{2n}$ is an eigenvector of L, with associated eigenvalue $\lambda = 2q/(q + 1)$.

   (e) By considering the regime $q \ll 1$, conclude that the existence of two small eigenvalues is related to the existence of two clusters in the graph $\mathcal{G}$.