Writing Assignment 4

Issued: Sunday 28th November, 2021 **Due:** Sunday 12th December, 2021

4.1. (K-means) Given input data $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$, $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$, the k-means clustering partitions the input into k sets $C_1, \ldots, C_k$ to minimize the within-cluster sum of squares:

$$\arg\min_{C} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2,$$

where $\boldsymbol{\mu}_j$ is the center of the j-th cluster:

$$\boldsymbol{\mu}_j \triangleq \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}, \quad j = 1, \ldots, k.$$

(a) (2 points) Show that the k-means clustering problem is equivalent to minimizing the pairwise squared deviation between points in the same cluster:

$$\sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{\boldsymbol{x}, \boldsymbol{x}' \in C_j} \|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

(b) (2 points) Show that the k-means clustering problem is equivalent to maximizing the between-cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{j=1}^{k} |C_i||C_j|\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2.$$

4.2. (PCA) We will talk about a natural way to define PCA called Projection Residual Minimization. Suppose we have m samples $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(m)} \in \mathbb{R}^n\}$, then we try to use the projections or image vectors to represent the original data. There will be some errors (projection residuals) and naturally we hope to minimize such errors.

(a) (2 points) First consider the case with one-dimentional projections. Let $\boldsymbol{u}$ be a non-zero unit vector. The projection of sample $\boldsymbol{x}^{(i)}$ on vector $\boldsymbol{u}$ is represented by $(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}$. Therefore the residual of a projection will be

$$\left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|.$$

Please show that

$$\arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2 = \arg\max_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2$$

(b) (2 points) Follow the proof above and the discussion of the variance of projections in the lecture. Please show that minimizing the residual of projections is equivalent to finding the largest eigenvector of covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{u}^\star = \arg\min_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1} \frac{1}{m} \sum_{i=1}^{m} \left\|\boldsymbol{x}^{(i)} - (\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2$$

then $\boldsymbol{u}^\star$ is the largest eigenvector of $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$.

4.3. (Ncut) Recall the NCut problem is defined as minimizing

$$NCut(A_1, \dots, A_k) = \sum_{i=1}^{k} \frac{Cut(A_i, \bar{A}_i)}{vol(A_i)}$$

with respect to partition $A_1, \dots, A_k$ of G, and $vol(A) = \sum_{v_i \in A, v_j \in V} w_{ij}$.

We derive the normalized spectral clustering as relaxation of minimizing Ncut for the case $k = 2$.

(a) (1 point) For each vertex v_i, let $f_i = \begin{cases} \sqrt{\frac{vol(\bar{A})}{vol(A)}} & v_i \in A \\ \sqrt{-\frac{vol(A)}{vol(\bar{A})}} & v_i \notin A \end{cases}$ be the vertex label function. Show that the Ncut problem is equivalent to the following optimization problem:

$$\begin{aligned} \min_{A} \quad & \boldsymbol{f}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{f}, \\ s.t. \quad & \boldsymbol{D} \boldsymbol{f} \perp \mathbf{1} \\ & \boldsymbol{f}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{f} = vol(V) \end{aligned}$$

where D, L are the degree matrix and Laplacian matrix of G. Hint: First show that $(D\boldsymbol{f})^T \mathbf{1} = 0$ and $\boldsymbol{f}^T D \boldsymbol{f} = vol(V)$.

(b) (2 points) Now we relax the above optimization problem such that $\boldsymbol{f} \in \mathbb{R}^n$. Please prove that the optimal f^* is the second eigenvector of $L_{rw} := D^{-1}L$, and that it is the generalized eigenvector: $Lf = \lambda D f$.