Tsinghua-Berkeley Shenzhen Institute
Learning From Data
Fall 2020

Writing Assignment 3

Issued: Saturday 14th November, 2020 **Due:** Friday 27th November, 2020

POLICIES

- **Acknowledgments:** We expect you to make an honest effort to solve the problems individually. As we sometimes reuse problem set questions from previous years, covered by papers and web pages, we expect the students **NOT** to copy, refer to, or look at the solutions in preparing their answers (relating to an unauthorized material is considered a violation of the honor principle). Similarly, we expect to not to google directly for answers (though you are free to google for knowledge about the topic). If you do happen to use other material, it must be acknowledged here, with a citation on the submitted solution.
- **Required homework submission format:** You can submit homework either as one single PDF document or as handwritten papers. Written homework needs to be provided during the class in the due date, and PDF document needs to be submitted through Tsinghua's Web Learning (`http://learn.tsinghua.edu.cn/`) before the end of due date.

  It is encouraged you LaTeX all your work, and we would provide a LaTeX template for your homework.
- **Collaborators:** In a separate section (before your answers), list the names of all people you collaborated with and for which question(s). If you did the HW entirely on your own, **PLEASE STATE THIS**. Each student must understand, write, and hand in answers of their own.

2.1. (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:

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

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

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

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

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

(b) Define the distortion of k-means clustering as

$$J(\{c^{(i)}\}_{i=1}^{m},\{\boldsymbol{\mu}_j\}_{j=1}^{k})=\sum_{i=1}^{m}\|\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_{c^{(i)}}\|^2.$$

i. (0.5 points) Show that the distortion J does not increase in each step of Lloyd's algorithm (refer to the lecture slides).

ii. (0.5 points) Does this algorithm always converge? Prove it or give a counterexample.

2.2. (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) (1 point) 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

$$\operatorname*{arg\,min}_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1}\left\|\boldsymbol{x}^{(i)}-(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u})\boldsymbol{u}\right\|^2=\operatorname*{arg\,max}_{\boldsymbol{u}:\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}=1}\left(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}\right)^2$$

(b) (1 point) 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 $\boldsymbol{\Sigma}$.

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

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

(c) (1 point) Now for a n-dimensional projection where the basis is a complete orthonormal set $\{\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots,\boldsymbol{u}_n\}$ that satisfies $\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}$, where

$$\delta_{ij}=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases}$$

If we pick up a k-dimension projection, the residual will be the linear combination of the remaining bases.

$$\boldsymbol{x}^{(i)}-\sum_{j=1}^{k}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j=\sum_{j=k+1}^{n}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j$$

Please show that

$$\min_{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_k:\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j=\delta_{ij}} \frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}^{(i)}-\sum_{j=1}^{k}(\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{u}_j)\boldsymbol{u}_j\right\|^2=\sum_{i=k+1}^{n}\lambda_i,$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ is the eigenvalues of $\boldsymbol{\Sigma}$. It leads to the conclusion that the minimum average projection error is the sum of the eigenvalues of those eigenvectors that are orthogonal to the principal subspace .

2.3. (Kernel PCA 2 ponit) Show that the conventional linear PCA algorithm is recovered as a special case of kernel PCA if we choose the linear kernel function given by $k(\boldsymbol{x},\boldsymbol{x}')=\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}')=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$.

2.4. (Bonus Question) (SVD) In the CCA and maximal correlation lecture, we used singular value decomposition (SVD)[1] to extract important features from data. The following exercise explores several properties of SVD in details.

Suppose a rank-r matrix $A \in \mathbb{R}^{m\times n}$ has the singular value decomposition: $A = U\Sigma V^{\mathrm{T}}$, where $U = [u_1,\ldots,u_r] \in \mathbb{R}^{m\times r}, \Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_r), V = [v_1,\ldots,v_r] \in \mathbb{R}^{n\times r}, U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_r, \sigma_1 \geq \cdots \geq \sigma_r > 0$.

(a) (1 point) Show that $Av_i = \sigma_i u_i, A^{\mathrm{T}}u_i = \sigma_i v_i, i = 1,\ldots,r$.

(b) (1 point) The 2-norm of A is defined as

$$\|A\|_2 \overset{\text{def}}{=} \max_{x\in\mathbb{R}^n:\ \|x\|>0}\frac{\|Ax\|}{\|x\|}.$$

Prove that $\|A\|_2 = \sigma_1$. (*Hint: If* $U^{\mathrm{T}}U = I$, *then* $\|Ux\| = \|x\|$.)

[1]See https://en.wikipedia.org/wiki/Singular_value_decomposition for reference on SVD.