Tsinghua-Berkeley Shenzhen Institute
Learning from Data
Fall 2019

Written Assignment 3

Issued: Sunday 10th November, 2019 **Due:** Sunday 24th November, 2019

3.1. (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) i. (1 point) 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. (1 point) 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.

3.2. (PCA) The covariance matrix of a random vector $\mathbf{x} \in \mathbb{R}^d$ is defined as

$$\mathrm{Cov}(\mathbf{x}) \triangleq \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}] \in \mathbb{R}^{d \times d},$$

where $\mathbb{E}[\cdot]$ is the mathematical expectation.

(a) (1 point) Show that

i. $\boldsymbol{u}^{\mathrm{T}} \mathrm{Cov}(\mathbf{x}) \boldsymbol{u} \geq 0, \forall\, \boldsymbol{u} \in \mathbb{R}^d$.

ii. $\mathrm{tr}(\mathrm{Cov}(\mathbf{x})) = \mathbb{E}[\|\mathbf{x} - \mathbb{E}[\mathbf{x}]\|^2]$.

(b) (1 point) Suppose we want to estimate the covariance matrix $\hat{\boldsymbol{C}}$ of the dataset $\mathcal{X}$ in 3.1 using the following formula:

$$\hat{\boldsymbol{C}} = \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{\mu}})(\boldsymbol{x}^{(i)} - \hat{\boldsymbol{\mu}})^{\mathrm{T}},$$

p+1

where

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}^{(i)}.$$

Give the minimum value of m required such that $\hat{\boldsymbol{C}}$ is non-singular.

3.3. (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) (0.5 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) (0.5 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 $\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$

(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 $\mathbf{\Sigma}$. 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 .

3.4. (Kernel PCA 1 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 $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{x}') = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$.

3.5. (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

$$\|A\|_2 \triangleq \max_{x\in\mathbb{R}^n:\, \|x\|>0} \frac{\|Ax\|}{\|x\|}.$$

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.