Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Problem Set 3

Issued: Tuesday 20th November, 2018 **Due:** Monday 26th November, 2018

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

ii. Does this algorithm always converge? Prove it or give a counterexample.

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

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

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.