Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

Writing Assignment 4

Issued: Tuesday 1st December, 2020 **Due:** Monday 14th December, 2020

4.1. (3 points) *HGR Maximal Correlation* In the derivation of HGR maximal correlation analysis, given a feature function $f\colon \mathcal{X} \to \mathbb{R}$, we defined the corresponding *information vector* as the vector $\phi \in \mathbb{R}^{|\mathcal{X}|}$ with elements $\phi(x) = f(x)\sqrt{P_X(x)}$. This correspondence between function f and information vector ϕ is denoted by $\phi \leftrightarrow f(X)$. Show that

(a) $\phi_1 \leftrightarrow 1(X)$, where $\phi_1 = \left(\sqrt{P_X(1)}, \dots, \sqrt{P_X(|\mathcal{X}|)}\right)^{\mathrm{T}}$, and $1(x)$ is a constant function, i.e. $1(x) = 1$ for all $x \in \mathcal{X}$.

(b) The variance of a feature is the length of its corresponding information vector: $\mathbb{E}[f^2(X)] = \|\phi\|^2$, where $\phi \leftrightarrow f(X)$.

(c) The covariance of two features is the inner product of their information vectors: $\langle \phi_1, \phi_2 \rangle = \mathbb{E}[f_1(X)f_2(X)]$, where $\phi_1 \leftrightarrow f_1(X), \phi_2 \leftrightarrow f_2(X)$.

Solution:

(a) Easily verified using definition.

(b) $\mathbb{E}[f^2(X)] = \sum_{x\in\mathcal{X}} P_X(x) f^2(x) = \sum_{x\in\mathcal{X}} \left(\sqrt{P_X(x)} f(x)\right)^2 = \|\phi\|^2.$

(c)

$$\begin{aligned}
\mathbb{E}[f_1(X)f_2(X)] &= \sum_{x\in\mathcal{X}} P_X(x) f_1(x) f_2(x) \\
&= \sum_{x\in\mathcal{X}} \left(\sqrt{P_X(x)} f_1(x)\right) \cdot \left(\sqrt{P_X(x)} f_1(x)\right) \\
&= \langle \phi_1, \phi_2 \rangle.
\end{aligned}$$

4.2. (3 points) *ICA* In the lecture, we briefly discussed why Gaussian random variables are forbidden in ICA. To understand this limitation more formally, let's assume that the joint distribution of two independent components, say, s_1,s_2, are Gaussian.

$$P(\boldsymbol{s}_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s_i^2}{2}\right)$$

(a) Please find the joint pdf $P(s_1, s_2)$.

(b) Suppose that the mixing matrix $\boldsymbol{A}$ is orthogonal. For example, we could assume that this is so because the data has been whitened, which means $\boldsymbol{A}^{-1} = \boldsymbol{A}^{\mathrm{T}}$ holds. Please find the joint pdf $P(x_1, x_2)$ of the mixtures x_1 and x_2 and then explain why Gaussian variables are forbidden.

Solution:

(a)

$$P(s_1, s_2) = P(s_1)P(s_2) = \frac{1}{2\pi}\exp\left(-\frac{s_1^2 + s_2^2}{2}\right) = \frac{1}{2\pi}\exp\left(-\frac{\|\boldsymbol{s}\|^2}{2}\right)$$

(b)

$$P(x_1, x_2) = \frac{1}{2\pi}\exp\left(-\frac{\|\boldsymbol{A}^{\mathbf{T}}\boldsymbol{x}\|^2}{2}\right)\left|\det \boldsymbol{A}^{\mathrm{T}}\right| = \frac{1}{2\pi}\exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right)$$

It means you cannot tell any difference between $P(\boldsymbol{s})$ and $P(\boldsymbol{x})$. The original and mixed distributions are identical. Therefore, there is no way how we could infer the mixing matrix from the mixtures.

4.3. (4 points) *EM for Mixture of Gaussian (Soft k-Means)* We talked about EM for Mixture of Gaussians in class. Please repeat what have been done in this problem. Consider the case of a mixture of k Gaussians in which $\boldsymbol{\theta}$ is a triplet $(\boldsymbol{\phi}, \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_k\})$. For simplicity, we assume that $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{I}$, which don't need calculations in your EM steps. We have that

$$P_{\boldsymbol{\theta}^{(t)}}(Z = z) = \phi_z^{(t)}$$

$$P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{X} = \boldsymbol{x}|Z = z) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_z|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_z^{(t)})^{\mathrm{T}}\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_z^{(t)})\right)$$

(a) Please derive the updates in the E-step and M-step. *Hint: The E-step needs to write out* $P_{\boldsymbol{\theta}^{(t)}}(Z = z|\boldsymbol{X} = \boldsymbol{x}_i)$

(b) Write down the updated parameter $\boldsymbol{\theta}^{(t+1)}$ and compare your procedures with K-means.

Solution:

(a) **E-step**

$$\begin{aligned} P_{\boldsymbol{\theta}^{(t)}}(Z = z|\boldsymbol{X} = \boldsymbol{x}_i) &= \frac{1}{C_i^{(t)}}P_{\boldsymbol{\theta}^{(t)}}(Z = z)P_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{X} = \boldsymbol{x}_i|Z = z) \\ &= \frac{1}{C_i^{(t)}}\phi_z^{(t)}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_z^{(t)})^{\mathrm{T}}\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_z^{(t)})\right) \end{aligned}$$

M-step

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m}\sum_{z=1}^{k}\left(\log(\phi_z) - \frac{1}{2}\|\boldsymbol{x}_i - \boldsymbol{\mu}_z\|^2\right)$$

(b)

$$\boldsymbol{\mu}_y^{(t+1)} = \frac{\sum_{i=1}^{m} P_{\boldsymbol{\theta}^{(t)}}(Z = z|\boldsymbol{X} = \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{i=1}^{m} P_{\boldsymbol{\theta}^{(t)}}(Z = z|\boldsymbol{X} = \boldsymbol{x}_i)}$$

$$\phi_y = \frac{\sum_{i=1}^{m} P_{\boldsymbol{\theta}^{(t)}}(Z = z|\boldsymbol{X} = \boldsymbol{x}_i)}{\sum_{z'=1}^{k}\sum_{i=1}^{m} P_{\boldsymbol{\theta}^{(t)}}(Z = z'|\boldsymbol{X} = \boldsymbol{x}_i)}$$

In the K-means algorithm, we first assign each example to a cluster according to the distance $\|\boldsymbol{x}_i - \boldsymbol{\mu}_z\|$. Then, we update each center according to the average of the examples assigned to this cluster. In the EM approach, however, we determine the probability that each example belongs to each cluster. Then, we update the centers on the basis of a weighted sum over the entire sample. For this reason, the EM approach for K-means is sometimes called "soft K-means".

4.4. (2 points) (Bonus question) *Weyl's Theorem* This problem introduces you to perturbation theory in PCA. Perturbation theory is useful in many real world problems, for instance, suppose we have computed the largest eigenvalue of the covariance matrix of some original samples. Then suddenly a bunch of new data come in and the covariance matrix should be like

$$\boldsymbol{\Sigma} = \frac{n_{origin}\boldsymbol{\Sigma}_{origin} + n_{new}\boldsymbol{\Sigma}_{new}}{n_{origin} + n_{new}}$$

Let's note it as

$$\boldsymbol{\Sigma} = \boldsymbol{A} + \boldsymbol{B}$$

Define $\lambda(M)$ as the eigenvalue operator of matrix M. Our target is to bound the eigenvalues $\lambda(\boldsymbol{\Sigma})$ given some knowledge about $\lambda(\boldsymbol{A})$.
Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and their eigenvalues denoted by $\{\lambda_i(\boldsymbol{A})\}_{i=1}^{n}$, $\{\lambda_i(\boldsymbol{B})\}_{i=1}^{n}$ with $\lambda_1 > \cdots > \lambda_n$. Please prove that for any $1 \le k \le n$

$$\lambda_k(\boldsymbol{A}) + \lambda_n(\boldsymbol{B}) \le \lambda_k(\boldsymbol{A}+\boldsymbol{B}) \le \lambda_k(\boldsymbol{A}) + \lambda_1(\boldsymbol{B})$$

Hint: you should first prove that for any $\boldsymbol{v} \in \mathbb{R}^n$

$$\lambda_n(\boldsymbol{B}) \le \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{v}}{\|\boldsymbol{v}\|^2} \le \lambda_1(\boldsymbol{B})$$

Solution: The lemma is easy to prove. If $k = 1$

$$\begin{aligned}\lambda_1(\boldsymbol{A}+\boldsymbol{B}) &= \max_{\boldsymbol{v}} \frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v}}{\|\boldsymbol{v}\|^2}\\ &= \max_{\boldsymbol{v}} \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}}{\|\boldsymbol{v}\|^2} + \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{v}}{\|\boldsymbol{v}\|^2}\\ &\le \max_{\boldsymbol{v}} \frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{v}}{\|\boldsymbol{v}\|^2} + \lambda_1(\boldsymbol{B})\\ &= \lambda_1(\boldsymbol{A}) + \lambda_1(\boldsymbol{B})\end{aligned}$$

Then set the largest eigenvector of $\boldsymbol{A}$ as $\boldsymbol{v}_1$

$$\begin{aligned}\lambda_2(\boldsymbol{A}+\boldsymbol{B}) &\le \max_{\boldsymbol{v}:\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}_1=0} \frac{\boldsymbol{v}^{\mathrm{T}}(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v}}{\|\boldsymbol{v}\|^2}\\ &\le \lambda_2(\boldsymbol{A}) + \lambda_1(\boldsymbol{B})\end{aligned}$$

Then you can prove $\lambda_k(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_k(\boldsymbol{A})+\lambda_1(\boldsymbol{B})$ The other half should be proved from n to 1. It is the same.