Theory Assignment 3 (Practice)

COMP 451 - Fundamentals of Machine Learning

Prof. William L. Hamilton

Winter 2021

Question 1 [7 points]

Let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ denote a set of points in $\mathbb{R}^m$. Let

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{1}$$

Show that the following inequality holds

$$\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \leq \sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{q}\|^2, \forall \mathbf{q} \in \mathbb{R}^m. \tag{2}$$

In other words, show that taking the mean of a set of points is the optimal choice in order to minimize the average distance to the points in that set.

Solution.
For any $\mathbf{q} \in \mathbb{R}^m$ we have that

$$\begin{aligned}
\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{q}\|^2 &= \sum_{i=1}^{n}\|(\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mathbf{q})\|^2 && (3)\\
&= \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + \|\bar{\mathbf{x}} - \mathbf{q}\|^2 + 2(\mathbf{x}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}} - \mathbf{q}) && (4)\\
&= n\|\bar{\mathbf{x}} - \mathbf{q}\|^2 + \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + 2(\mathbf{x}_i - \bar{\mathbf{x}})^\top(\bar{\mathbf{x}} - \mathbf{q}) && (5)\\
&= n\|\bar{\mathbf{x}} - \mathbf{q}\|^2 + \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + 2\left(\mathbf{x}_i^\top\bar{\mathbf{x}} - \mathbf{x}_i^\top\mathbf{q} - \|\bar{\mathbf{x}}\|^2 + \bar{\mathbf{x}}^\top\mathbf{q}\right) && (6)\\
&= n\|\bar{\mathbf{x}} - \mathbf{q}\|^2 + 2\left(\left(\sum_{i=1}^{n}\mathbf{x}_i\right)^\top\bar{\mathbf{x}} - \left(\sum_{i=1}^{n}\mathbf{x}_i\right)^\top\mathbf{q} - m\|\bar{\mathbf{x}}\|^2 + m\bar{\mathbf{x}}^\top\mathbf{q}\right) + \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 && (7)\\
&= n\|\bar{\mathbf{x}} - \mathbf{q}\|^2 + 2\left(m\|\bar{\mathbf{x}}\|^2 - m\bar{\mathbf{x}}^\top\mathbf{q} - m\|\bar{\mathbf{x}}\|^2 + m\bar{\mathbf{x}}^\top\mathbf{q}\right) + \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 && (8)\\
&= n\|\bar{\mathbf{x}} - \mathbf{q}\|^2 + \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 && (9)\\
&\geq \sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 && (10)
\end{aligned}$$

Question 2 [7 points]

In class we introduced the Gaussian mixture model (GMM). In this question, we will consider a mixture of Bernoulli distributions. Here, our data points will be defined as m-dimensional vectors of binary values $\mathbf{x} \in \{0,1\}^m$.

First, we will introduce a single multivariate Bernoulli distribution, which is defined by a mean vector $\boldsymbol{\mu}$

$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{j=0}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}[j])}. \tag{11}$$

Thus, we see that a the individual binary dimensions are independent for a single multivariate Bernoulli. Now, we can define a mixture of K multivariate Bernoulli distributions as follows

$$\begin{align} P(\mathbf{x}|\Theta) &= \sum_{k=0}^{K-1} \pi_k P(\mathbf{x}|\boldsymbol{\mu}_k) \tag{12} \\ &= \sum_{k=0}^{K-1} \pi_k \prod_{j=0}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}[j])} \tag{13} \\ & \tag{14} \end{align}$$

where $\Theta = \{\boldsymbol{\mu}_k, \pi_k, k = 0, .., K-1\}$ are the parameters of the mixture and $P(\mathbf{x}|\boldsymbol{\mu}_k)$ is the probability assigned to the point by each individual component in the model.

Now, suppose that we partition the datapoints $\mathbf{x}$ into two parts $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_b]$ so that $\mathbf{x}_a \in \{0,1\}^{m-d}$ and $\mathbf{x}_b \in \{0,1\}^d$. Show that the conditional distribution

$$P(\mathbf{x}_a|\mathbf{x}_b) \tag{15}$$

is itself a Bernoulli mixture distribution and provide expressions for the mixing coefficients and the component/cluster densities.

Solution.
First, we can apply the rules of probability to find that

$$\begin{align} P(\mathbf{x}_a|\mathbf{x}_b) &= \frac{P(\mathbf{x}_a, \mathbf{x}_b)}{P(\mathbf{x}_b)} \tag{16} \\ &= \frac{P(\mathbf{x})}{P(\mathbf{x}_b)} \tag{17} \\ &= \frac{\sum_{k=0}^{K-1} \pi_k P(\mathbf{x}|\boldsymbol{\mu}_k)}{P(\mathbf{x}_b)} \tag{18} \\ &= \sum_{k=0}^{K-1} \frac{\pi_k}{P(\mathbf{x}_b)} P(\mathbf{x}|\boldsymbol{\mu}_k). \tag{19} \end{align}$$

Thus, we see that $P(\mathbf{x}_a|\mathbf{x}_b)$ is a mixture distribution with the same component density as before but with the mixture coefficients given by $\frac{\pi_k}{P(\mathbf{x}_b)}$. The term $P(\mathbf{x}_b)$ can be computed via marginalization as

$$P(\mathbf{x}_b) = \sum_{\mathbf{c}\in\{0,1\}^{m-d}} \sum_{k=0}^{K-1} \pi_k \prod_{j=0}^{m-d-1} \boldsymbol{\mu}[j]^{\mathbf{c}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{c}[j])} \prod_{j=d}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}_b[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}_b[j])}. \tag{20}$$

Question 3 [4 points]

Recall that the low dimensional codes in PCA are defined as

$$\mathbf{z}_i = \mathbf{U}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \tag{21}$$

where $\mathbf{U}$ is a matrix containing the top-k eigenvectors of the covariance matrix as rows and

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{x}_i \tag{22}$$

Show that

$$\bar{\mathbf{z}} = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{z}_i = 0. \tag{23}$$

Solution. *We have that*

$$\begin{aligned}
\frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{z}_i &= \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \left(\mathbf{U}^\top(\mathbf{x}_i - \boldsymbol{\mu})\right) && (24)\\
&= \mathbf{U}^\top\left(\frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{x}_i - \boldsymbol{\mu}\right) && (25)\\
&= \mathbf{U}^\top\left(\frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{x}_i - \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \boldsymbol{\mu}\right) && (26)\\
&= \mathbf{U}^\top(\boldsymbol{\mu} - \boldsymbol{\mu}) && (27)\\
&= 0 && (28)
\end{aligned}$$

Question 4 [short answers; 2 points each]

a) True or false: Soft K-means and a Gaussian mixture model are equivalent.

b) Suppose you are learning a decision tree for email spam classification. Your current sample of the training data has the following distribution of labels:

$$[43+, 30-]$$

i.e., the training sample has 43 examples that are spam and 30 that are not spam. Now, you are choosing between two candidate tests.

Test 1 (T1) tests whether the number of words in the email is greater than 20 and would result in the following splits:

- num_words > 20 : $[13+, 20-]$
- num_words ≤ 20: [30+, 10-]

Test 2 (T2) tests whether the email contains spelling errors and would result in the following splits:

- spelling_error: $[30+, 15-]$
- no_spelling_error: $[13+, 15-]$

Which test should you use to split the data? I.e., which test provides a higher information gain?

c) True or false: If we transform some input features using PCA, then the covariance matrix of the resulting transformed features is diagonal.

Solution.
a) *False. The GMM also includes the covariance matrices as parameters.*

b) *The first test (T1) is better, since it has a lower conditional entropy (and thus higher information gain). The conditional entropy of T1 is*

$$H(data \,|\, T1) = -\frac{33}{73}\left(\frac{13}{33}\log_2\left(\frac{13}{33}\right) + \frac{20}{33}\log_2\left(\frac{20}{33}\right)\right) - \frac{40}{73}\left(\frac{30}{40}\log_2\left(\frac{30}{40}\right) + \frac{10}{40}\log_2\left(\frac{10}{40}\right)\right) \approx 0.882 \tag{29}$$

while the conditional entropy of T2 is

$$H(data \,|\, T2) = -\frac{45}{73}\left(\frac{30}{45}\log_2\left(\frac{30}{45}\right) + \frac{15}{45}\log_2\left(\frac{15}{45}\right)\right) - \frac{28}{73}\left(\frac{13}{28}\log_2\left(\frac{13}{28}\right) + \frac{15}{28}\log_2\left(\frac{15}{28}\right)\right) \approx 0.948 \tag{30}$$

c) *True, since the basis vectors in PCA are all orthogonal.*