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.

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 $\boldsymbol{\mu}$

$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{j=0}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}[j])}. \tag{3}$$

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

$$P(\mathbf{x}|\Theta) = \sum_{k=0}^{K-1} \pi_k P(\mathbf{x}|\boldsymbol{\mu}_k) \tag{4}$$

$$= \sum_{k=0}^{K-1} \pi_k \prod_{j=0}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}[j])} \tag{5}$$

$$\tag{6}$$

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

$$P(\mathbf{x}_a|\mathbf{x}_b) \tag{7}$$

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

Question 3 [4 points]

Recall that the low dimensional codes in PCA are defined as

$$\mathbf{z}_i = \mathbf{U}^\top(\mathbf{x}_i - \boldsymbol{\mu}), \tag{8}$$

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

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{x}_i \tag{9}$$

Show that

$$\bar{\mathbf{z}} = \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|-1} \mathbf{z}_i = 0. \tag{10}$$

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.