Theory Assignment 3

COMP 451 - Fundamentals of Machine Learning

Prof. William L. Hamilton

Winter 2021

Preamble The assignment is due April 6th at 11:59pm via MyCourses. Late work will be automatically subject to a 20% penalty, and can be submitted up to 5 days after the deadline. You may scan written answers or submit a typeset assignment, as long as you submit a single pdf file with clear indication of what question each answer refers to. You may consult with other students in the class regarding solution strategies, but you must list all the students that you consulted with on the first page of your submitted assignment. You may also consult published papers, textbooks, and other resources, but you must cite any source that you use in a non-trivial way (except the course notes). You must write the answer in your own words and be able to explain the solution to the professor, if asked.

Question 1 [13 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 $\boldsymbol{\mu}$

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

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

$$P(\mathbf{x}|\Theta) = \sum_{k=0}^{K-1} \pi_k P(\mathbf{x}|\boldsymbol{\mu}_k) \tag{2}$$

$$= \sum_{k=0}^{K-1} \pi_k \prod_{j=0}^{m-1} \boldsymbol{\mu}[j]^{\mathbf{x}[j]}(1-\boldsymbol{\mu}[j])^{(1-\mathbf{x}[j])} \tag{3}$$

$$\tag{4}$$

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.

Note that the mean of each individual component distribution $P(\mathbf{x}|\mu_k)$ is given by

$$\mathbb{E}_k[\mathbf{x}] = \boldsymbol{\mu}_k, \tag{5}$$

and the covariance matrix of each component is given by

$$\text{Cov}[\mathbf{x}] = \boldsymbol{\Sigma}_k = \text{diag}(\boldsymbol{\mu_k} \circ (1-\boldsymbol{\mu}_k)), \tag{6}$$

where $\circ$ denotes elementwise multiplication. In other words, the covariance matrix $\boldsymbol{\Sigma}_k$ for each component is a diagonal matrix with diagonal entries given by $\boldsymbol{\Sigma}_k[j, j] = \boldsymbol{\mu}[j](1 - \boldsymbol{\mu}[j])$. It is a diagonal matrix because each dimension is independent.

Part 1 [8 points]

Derive expression for the mean vector and the covariance matrix of the full mixture distribution defined in Equation 2. That is, give expressions for the following:

$$\mathbb{E}[\mathbf{x}] =? \qquad \text{Cov}[\mathbf{x}] =? \tag{7}$$

Hint: use the fact that

$$\begin{aligned}\text{Cov}[\mathbf{x}] &= \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top\right] \\ &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^\top.\end{aligned}$$

Part 2 [5 points]

Just as with a GMM, we can use the expectation maximization (EM) algorithm to compute learn the parameters of a Bernoulli mixture model. Here, we will provide you with the formula for the expectation step as well as the log-likelihood of the model. You must derive the formula for the maximization step.

Expectation step. In the expectation step of the Bernoulli mixture model, we compute scores $r(\mathbf{x}, k)$, which tell us how likely it is that point $\mathbf{x}$ belongs to component k. These scores are computed as follows:

$$r(\mathbf{x}, k) = \frac{\pi_k P(\mathbf{x}|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j P(\mathbf{x}|\boldsymbol{\pi}_j)}, \tag{8}$$

where $P(\mathbf{x}|\boldsymbol{\mu}_k)$ is defined as in Equation 2.

Log-likelihood. The log-likelihood of the Bernoulli mixture is defined as

$$\log(\mathcal{L}(\mathcal{D}, \Theta) = \sum_{\mathbf{x} \in \mathcal{D}} \sum_{k=1}^{K} r(\mathbf{x}, k) \left(\log(\pi_k) + \sum_{j=0}^{m-1} \mathbf{x}[j] \log(\boldsymbol{\mu}_k[j]) + (1 - \mathbf{x}[j]) \log(1 - \boldsymbol{\mu}_k[j]) \right) \tag{9}$$

Maximization step. You must find the formula for the $\boldsymbol{\mu}_k$ parameters in the maximization step:

$$\boldsymbol{\mu}_k = ? \tag{10}$$

Question 2 [5 points]

Recall that the low dimensional codes in PCA are defined as

$$\mathbf{z}_i = \mathbf{U}^\top (\mathbf{x}_i - \boldsymbol{\mu}), \tag{11}$$

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

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|-1} \mathbf{x}_i. \tag{12}$$

Recall that the reconstruction of a point $\mathbf{x}_i$ using its code $\mathbf{z}_i$ is given by

$$\tilde{\mathbf{x}}_i = \mathbf{U}\mathbf{z}_i + \boldsymbol{\mu}. \tag{13}$$

Show that

$$(\tilde{\mathbf{x}}_i - \mathbf{x}_i)^\top (\tilde{\mathbf{x}}_i - \boldsymbol{\mu}) = 0. \tag{14}$$

Question 3 [short answers; 2 points each]

Answer each question with 1-3 sentences for justification, potentially with equations/examples for support.
a) True or false: It is always possible to choose an initialization so that K-means converges in one iteration.

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:

$$[43+, 30-], \tag{15}$$

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 30 and would result in the following splits:

- num_words > 30 : $[5+, 15-]$
- num_words ≤ 30: $[38+, 15-]$

Test 2 (T2) tests whether the email contains an external URL link and would result in the following splits:

- has_link: $[25+, 5-]$
- not_has_link: $[18+, 25-]$

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

c) Which of the following statements is false:

1. If the covariance between two variables is zero, then their mutual information is also zero.
2. Adding more features is a useful strategy to combat underfitting.
3. Decision trees can learn non-linear decision boundaries.
4. The Gaussian mixture model contains more parameters than K-means.