CS 475 Machine Learning: Homework 4
Clustering, EM, Dimensionality Reduction, Graphical Models
Analytical Problems

Due: Wednesday April 15, 2020, 11:59 pm

50 Points Total Version 1.0

YOUR_NAME (YOUR_JHED)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the height of textboxes, font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

Notation

$\mathbf{x_i}$ One input data vector. $\mathbf{x_i}$ is M dimensional. $\mathbf{x_i} \in \mathbb{R}^{1 \times M}$.
We assume $\mathbf{x_i}$ is augmented with a 1 to include a bias term.

$\mathbf{X}$ A matrix of concatenated $\mathbf{x_i}$'s. There are N input vectors, so $\mathbf{X} \in \mathbb{R}^{N \times M}$

y_i The true label for input vector $\mathbf{x_i}$. In regression problems, y_i is a continuous value.
In general y_i can be a vector, but for now we assume y_i is a scalar. $y_i \in \mathbb{R}^1$.

$\mathbf{y}$ A vector of concatenated y_i's. There are N input vectors, so $\mathbf{y} \in \mathbb{R}^{N \times 1}$

$\mathbf{w}$ A weight vector. We are trying to learn the elements of $\mathbf{w}$.
$\mathbf{w}$ is the same number of elements as $\mathbf{x_i}$ because we will end up computing the dot product $\mathbf{x_i} \cdot \mathbf{w}$.
$\mathbf{w} \in \mathbb{R}^{M \times 1}$. We assume the bias term is included in $\mathbf{w}$.

Notes: In general, a lowercase letter (not boldface), a, indicates a scalar.
A boldface lowercase letter, $\mathbf{a}$, indicates a vector.
A boldface uppercase letter, $\mathbf{A}$, indicates a matrix.

1) K-Medoids Clustering (10 points)

K-medoids (`https://en.wikipedia.org/wiki/K-medoids`) is an algorithm similar to K-means, but changes the distance metric to the L1 distance. Both K-means and K-medoids attempt to minimize the squared error. In this case, we are minimizing:

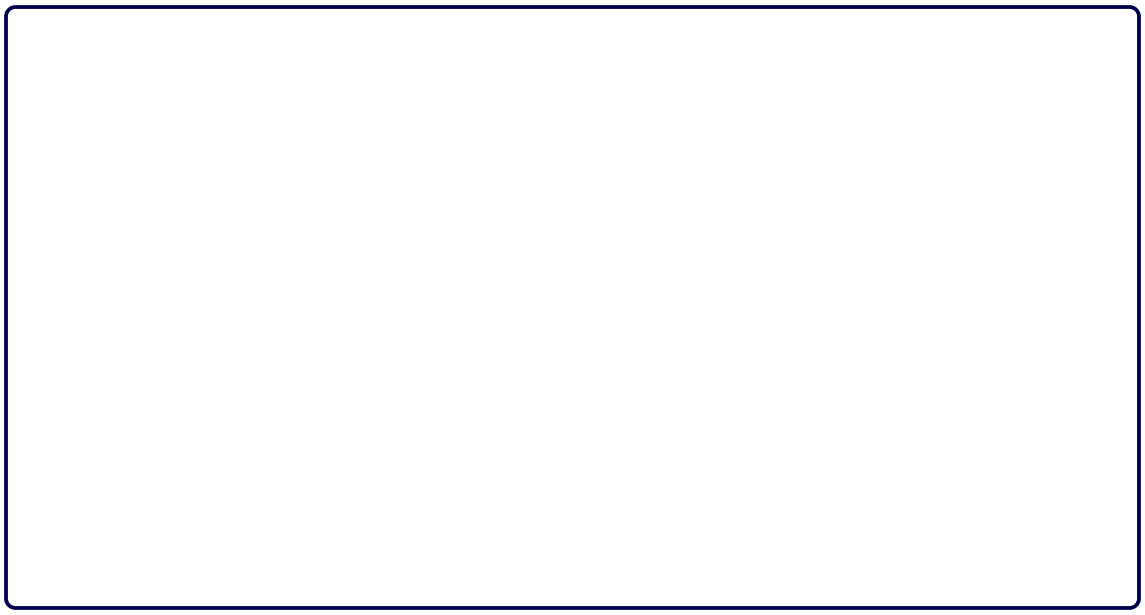
$$\min_{S=\{S_1,\ldots,S_k\}} \sum_{j=1}^{k} \sum_{x_i \in S_j} \|x_j - \mu_j\|_1 \tag{1}$$

Unlike K-means, K-medoids chooses a provided example as a cluster center (medoids) rather than the mean of a subset of the examples. Therefore, instead of selecting a new cluster center μ to be the mean of the datapoints assigned to the cluster, instead select the datapoint that has the closest average euclidean distance to the other points in the cluster. In case of a tie, select the datapoint with the lowest index (i.e. if x_0 and x_2 are tied, select x_0).

x_0	3	1
x_1	2	1
x_2	2	4
x_3	3	3
x_4	2	2
x_5	1	5

Table 1: Datapoints for Question 1

(a) For the dataset, run the K-medoids algorithm for two iterations, with $k = 2$ clusters. Select your initial cluster medoids to be $\mu_0 = x_0$ and $\mu_1 = x_1$ What are the final centers (i.e. μ_0 and μ_1)? Which data points are assigned to each cluster (cluster 0 and cluster 1)?



(b) For the dataset, run the K-means algorithm for two iterations, with $k = 2$ clusters. Continue to use the L1 distance. Select your initial cluster means to be $\mu_0 = x_0$ and $\mu_1 = x_1$ What are the final centers (i.e. μ_0 and μ_1)? Which data points are assigned to each cluster (cluster 0 and cluster 1)?

(c) What are the benefits of the K-medoids algorithm, compared to K-means (**briefly**, in no more than three sentences)?

2) Expectation-Maximization (15 points)

(a) Consider a dataset with N examples of D dimensions concatenated into the matrix $\mathbf{X} \in \{0,1\}^{N \times D}$. One example $\mathbf{x}_n \in \{0,1\}^D$ corresponds to a row of $\mathbf{X}$. x_{ni} corresponds to the binary element in row n, column i of $\mathbf{X}$. As the data is binary, $x_{ni} \in \{0,1\}$. We can model $\mathbf{X}$ using a mixture of K Bernoulli distributions, where the probability $x_{ni} = 1$ according to distribution k is μ_{ki}. $\boldsymbol{\mu}_k \in [0,1]^D$ is the vector describing the probability each dimension of $\mathbf{x}_n$ is equal to one, according to distribution k. We model the "responsibility" of distribution k for modeling the data as π_k, such that $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0, \forall k$. Let $\mathbf{z}_n \in \{0,1\}^K$ be a binary indicator assigning one distribution to each data point, such that exactly one element in $\mathbf{z}_n$ is equal to one. Let $\mathbf{Z} \in \{0,1\}^{N \times K}$ be the concatenation $\mathbf{z}_n \forall n$. We define $\gamma(z_{nk}) = \mathbb{E}[z_{nk}]$ and $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$. *For more details, see Bishop 9.3.3.*

Show that if we maximize the expected complete-data log-likelihood for a mixture of Bernoulli distributions (2) with respect to $\boldsymbol{\mu}_k$, we obtain the M-step equation (3).

$$\mathbb{E}_{\mathbf{Z}}\left[\log(p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}))\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\Big\{\log(\pi_k) + \sum_{i=1}^{D}\left[x_{ni}\log(\mu_{ki}) + (1 - x_{ni})\log(1 - \mu_{ki})\right]\Big\} \tag{2}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{3}$$

(b) Show that as a consequence of the constraint $0 \leq p(\mathbf{x}_n|\boldsymbol{\mu}_k) \leq 1$ for the discrete variable $\mathbf{x}_n$, the incomplete-data log-likelihood function for a mixture of Bernoulli distributions, $\log(p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\pi}))$, is bounded above and hence there are no singularities for which the likelihood goes to infinity.

(c) The lower bound $\mathcal{L}(q, \theta)$ given by (4) with $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ has the same gradient with respect to θ as the log-likelihood function $\log(p(\mathbf{X}|\theta))$ at the point $\theta = \theta^{\text{old}}$. Explain why this is the case for a general model, not a specific GMM, BMM, or other.

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right) \tag{4}$$

3) Dimensionality Reduction (10 points)

PCA is often a pre-processing step before for classification. Consider a binary classification task, where each class is generated from a separate Gaussian. The first class uses a Gaussian with $\mu_1 = \{25, 2\}$ and covariance

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 0.005 & 0 \\ 0 & 4 \end{pmatrix}$$

and the second class has $\mu_2 = \{28, 2\}$ and covariance

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 0.0004 & 0 \\ 0 & 2 \end{pmatrix}$$

For the questions below, it may help to generate and plot some samples from these two distributions in Python.

(a) Suppose you train an SVM on this data (with equal numbers of samples from both Gaussians). How well would it distinguish data from these two classes? Why?

(b) Suppose we run PCA on this data and produce a one dimensional representation. Describe the principal component that PCA would select.

(c) We now train an SVM on the new one dimensional PCA representation. How well would it distinguish data from these two classes. Why?

4) Graphical Models (15 points)

1. Consider the Bayesian Network given in Figure 1. Are the sets **A** and **B** d-separated given set **C** for each of the following definitions of **A**, **B** and **C**? Justify each answer.

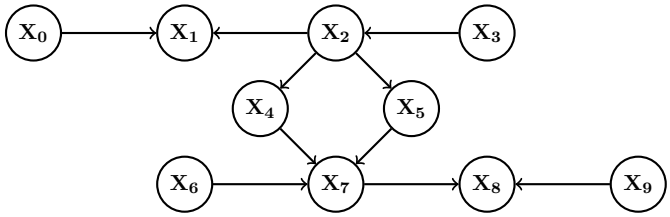


Figure 1: A directed graph

(a) $\mathbf{A} = \{X_3\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_1, X_2\}$

(b) $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2, X_7\}$

(c) $\mathbf{A} = \{X_4\}$, $\mathbf{B} = \{X_6\}$, $\mathbf{C} = \{X_8\}$

(d) $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2\}$

(e) $\mathbf{A} = \{X_6, X_9\}$, $\mathbf{B} = \{X_8\}$, $\mathbf{C} = \{X_3, X_1\}$

2. Now consider a Markov Random Field as given in Figure 2, where each edge is undirected. Are the sets **A** and **B** d-separated given set **C** for each of the following definitions of **A**, **B** and **C**? Justify each answer.

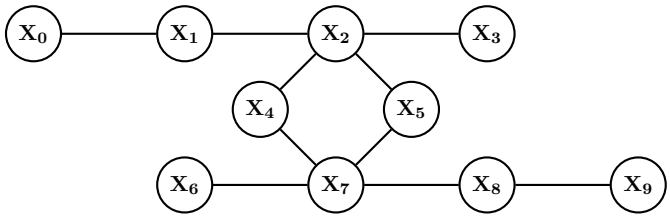


Figure 2: An undirected graph

(a) $\mathbf{A} = \{X_3\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_1, X_2\}$

(b) $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2, X_7\}$

(c) $\mathbf{A} = \{X_4\}$, $\mathbf{B} = \{X_6\}$, $\mathbf{C} = \{X_8\}$

(d) $\mathbf{A} = \{X_5\}$, $\mathbf{B} = \{X_4\}$, $\mathbf{C} = \{X_2\}$

(e) $\mathbf{A} = \{X_6, X_9\}$, $\mathbf{B} = \{X_8\}$, $\mathbf{C} = \{X_3, X_1\}$