CS 7641 CSE/ISYE 6740 Homework 2

Prakash, Fall 2021

Deadline: Oct 18, 6pm ET

- Submit your answers as an electronic copy on Canvas and Gradescope.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit.
- Typing with Latex is highly recommended. Typing with MS Word is also okay. If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting. 5 points extra credit if you submit typeset answers.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 1.5, 1.6, 2.5, 9.2, 9.3

1 EM for Mixture of Gaussians

Mixture of K Gaussians is represented as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1}$$

where π_k represents the probability that a data point belongs to the kth component. As it is probability, it satisfies $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$. In this problem, we are going to represent this in a slightly different manner with explicit latent variables. Specifically, we introduce 1-of-K coding representation for latent variables $z^{(k)} \in \mathbb{R}^K$ for $k = 1, ..., K$. Each $z^{(k)}$ is a binary vector of size K, with 1 only in kth element and 0 in all others. That is,

$$\begin{aligned} z^{(1)} &= [1; 0; ...; 0] \\ z^{(2)} &= [0; 1; ...; 0] \\ &\vdots \\ z^{(K)} &= [0; 0; ...; 1]. \end{aligned}$$

For example, if the second component generated data point x^n, its latent variable z^n is given by $[0; 1; ...; 0] = z^{(2)}$. With this representation, we can express $p(z)$ as

$$p(z) = \prod_{k=1}^{K} {\pi_k}^{z_k},$$

where z_k indicates kth element of vector z. Also, $p(x|z)$ can be represented similarly as

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k}.$$

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

By the sum rule of probability, (1) can be represented by

$$p(x) = \sum_{z \in Z} p(z)p(x|z). \tag{2}$$

where $Z = \{z^{(1)}, z^{(2)}, ..., z^{(K)}\}$.

(a) Show that (2) is equivalent to (1). [5 pts]

↪ Consider $z = z^{(i)}$.

↪ Clearly, $p(z^{(i)}) = \pi_1^{z_1}...\pi_i^{z_i}...\pi_k^{z_k} = \pi_1^0...\pi_i^1...\pi_k^0 = \pi_i$

↪ $p(x \mid z^{(i)}) = \mathcal{N}(x \mid \mu_1, \Sigma_1)^{z_1}...\mathcal{N}(x \mid \mu_i, \Sigma_i)^{z_i}...\mathcal{N}(x \mid \mu_k, \Sigma_k)^{z_k} = \mathcal{N}(x \mid \mu_1, \Sigma_1)^0...\mathcal{N}(x \mid \mu_i, \Sigma_i)^1...\mathcal{N}(x \mid \mu_k, \Sigma_k)^0 = \mathcal{N}(x \mid \mu_i, \Sigma_i)$

↪ From the above two steps, we have $p(x) = \sum_{z \in Z} p(z)p(x|z) = p(z^{(1)})p(x|z^{(1)}) + ... + p(z^{(i)})p(x|z^{(i)}) + ... + p(z^{(k)})p(x|z^{(k)})$

↪ $p(x) = \pi_1\mathcal{N}(x \mid \mu_1, \Sigma_1) + ... + \pi_i\mathcal{N}(x \mid \mu_i, \Sigma_i) + ... + \pi_k\mathcal{N}(x \mid \mu_k, \Sigma_k)$

↪ Which means p(x) = $\sum_{k=1}^{K} \pi_k\mathcal{N}(x|\mu_k, \Sigma_k)$, and is hence proved to be equivalent.

(b) In reality, we do not know which component each data point is from. Thus, we estimate the responsibility (expectation of z_k^n) in the E-step of EM. Since z_k^n is either 1 or 0, its expectation is the probability for the point x_n to belong to the component z_k. In other words, we estimate $p(z_k^n|x_n)$. Derive the formula for this estimation by using Bayes rule. Note that, in the E-step, we assume all other parameters, i.e. π_k, μ_k, and Σ_k, are fixed, and we want to express $p(z_k^n|x_n)$ as a function of these fixed parameters. [10 pts]

↪ It is common to note from Bayes rule - $p(x) = \dfrac{p(x \mid z)p(z)}{p(x)} = \dfrac{p(x, z)}{\sum_{z'} p(x, z')}$

↪ Now we can obtain $p(z_k^n \mid x_n) = \dfrac{p(x_n \mid z_k^n)p(z_k^n)}{p(x_n)} = \dfrac{\pi_k\mathcal{N}(x \mid \mu_k, \Sigma_k)}{\sum_i \pi_i\mathcal{N}(x \mid \mu_i, \Sigma_i)}$

(c) In the M-Step, we re-estimate parameters π_k, μ_k, and Σ_k by maximizing the log-likelihood. Given N i.i.d (Independent Identically Distributed) data samples, derive the update formula for each parameter. Note that in order to obtain an update rule for the M-step, we fix the responsibilities, i.e. $p(z_k^n|x_n)$, which we have already calculated in the E-step. [15 pts]

Hint: Use Lagrange multiplier for π_k to apply constraints on it.

↪ Likelihood (for n samples) L = $\prod_i^n p(x_i) = \prod_i^n \sum_k^K \pi_k\mathcal{N}(x_i \mid \mu_k, \Sigma_k)$

↪ $\log L = \sum_i^n \log(\sum_i^K \pi_k\mathcal{N}(x_k \mid \mu_k, \Sigma_k))$

- To update μ_k we see as follows -

↪ $\dfrac{\partial \log L}{\partial \mu_k} = \sum_i^n \dfrac{\pi_k \dfrac{\partial \mathcal{N}(x_k \mid \mu_k, \Sigma_k)}{\partial \mu_k}}{\sum_i^K \pi_k\mathcal{N}(x_k \mid \mu_k, \Sigma_k)} = \sum_i^n \dfrac{\pi_k\mathcal{N}(x_k \mid \mu_k, \Sigma_k)(x_i - \mu_k)}{\sum_i^K \pi_k\mathcal{N}(x_k \mid \mu_k, \Sigma_k)} = 0$

↪ The above follows since $\dfrac{\partial \mathcal{N}(x_k \mid \mu_k, \Sigma_k)}{\partial \mu_k} = \dfrac{exp(-\dfrac{(x_i - \mu - k)^T\Sigma^{-1}(x_i - \mu_k)}{2})(-\Sigma^{-1})(x_i - \mu_k}{((2\pi)^l|\Sigma|)^{0.5}}$

$\hookrightarrow$ Which implies $\sum_i^n p(z_k^i = 1 \mid x_i)(x_i - \mu_k) = 0$

$\hookrightarrow$ Update - $\mu_k = \dfrac{\sum_i^n p(z_k^i = 1 \mid x_i)(x_i)}{\sum_i^n p(z_k^i = 1 \mid x_i)}$

- To proceed with updating -

$\hookrightarrow$ Maximize $\sum_i^n \log(\sum_i^K \pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k))$ subject to $\sum_k^K \pi_k = 1$

$\hookrightarrow$ Using a lagrange multiplier, $\sum_i^n \log(\sum_i^K \pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k)) + \lambda(1 - \sum_k^K \pi_k)$

$\hookrightarrow$ Differentiate w.r.t π_k : $\sum_i^n \dfrac{\mathcal{N}(x_k \mid \mu_k, \Sigma_k)}{\sum_i^K \pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k)} - \lambda = 0$

$\hookrightarrow$ $\sum_i^n \dfrac{p(z_k^i = 1 \mid x_i)}{\pi_k} - \lambda = 0$

$\hookrightarrow$ $\pi_k = \sum_i^n \dfrac{p(z_k^i = 1 \mid x_i)}{\lambda}$

$\hookrightarrow$ But, $\sum_k^K \pi_k = 1$, therefore. $\sum_k^K \pi_k = \dfrac{\sum_i^n \sum_k^K p(z_k^i \mid x_i)}{\lambda} = 1 \implies \dfrac{n}{\lambda} = 1 \implies \lambda = n$

$\hookrightarrow$ Therefore, we update $\pi_k = \sum_i^n \dfrac{p(z_k^i = 1 \mid x_i)}{n}$

- Now, to update Σ_k, we do as below.

$\hookrightarrow$ $\dfrac{\partial \log L}{\partial \Sigma_k} = \sum_i^n \dfrac{\pi_k \dfrac{\partial \mathcal{N}(x_k \mid \mu_k, \Sigma_k)}{\partial \Sigma_k}}{\sum_i^K \pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k)}$

$\hookrightarrow$ $\dfrac{\partial \log L}{\partial \Sigma_k} = \sum_i^n \dfrac{\pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k) \dfrac{-1}{2}(\Sigma_k^{-1} - \Sigma_k^{-1}(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1})}{\sum_i^K \pi_k \mathcal{N}(x_k \mid \mu_k, \Sigma_k)}$

$\hookrightarrow$ $\dfrac{\partial \log L}{\partial \Sigma_k} = -\dfrac{\Sigma_k^{-1}}{2} \sum_i^n p(z_k^i = 1/x_i)(1 - (x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}) = 0$

$\hookrightarrow$ Multiplying both sides by Σ_k, we get : $I \sum_i^n p(z_k^i = 1/x_i) = \sum_i^n p(z_k^i = 1/x_i)(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}$

$\hookrightarrow$ Therefore, we update $\Sigma_k = \dfrac{\sum_i^n p(z_k^i = 1/x_i)(x_i - \mu_k)(x_i - \mu_k)^T}{\sum_i^n p(z_k^i = 1/x_i)}$

(d) EM and K-Means [10 pts]

K-means can be viewed as a particular limit of EM for Gaussian mixture. Considering a mixture model in which all components have covariance ϵI, show that in the limit $\epsilon \to 0$, maximizing the expected complete data log-likelihood for this model is equivalent to minimizing objective function in K-means:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \|x_n - \mu_k\|^2,$$

where $\gamma_{nk} = 1$ if x_n belongs to the k-th cluster and $\gamma_{nk} = 0$ otherwise.

- Using - All covariance matrices are ϵI, we get the following. (Double summation is represented with double indices on one summation - sum up separately)

$\hookrightarrow$ $\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{0.5}} exp \frac{||x - \mu_k||^2}{2\epsilon}$

$\hookrightarrow$ $p(z_k^n \mid x_n) = \frac{\pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)(x, z)}{\sum_i \pi_i \mathcal{N}(x \mid \mu_i, \Sigma_i)} = \frac{\pi_k exp \frac{||x - \mu_k||^2}{2\epsilon}}{\sum_i \pi_i exp \frac{||x - \mu_k||^2}{2\epsilon})}$

- Consider ϵ tends to 0, the term above (in the denominator) for which $||x_n - \mu_k||^2$ is the least will go to 0 the slowest. We get the following :

Eq.1 $p(z_k^n = 1 \mid x_n) = r_{nk}$ which is the assignment to clusters in k-means.

$\hookrightarrow$ Total likelihood is $p(X, Z \mid \mu, \Sigma, \pi) = \prod_{n,k}^{N,K} \pi_k^{z_{nk}} \mathcal{N}(x_n \mid \mu_k, \Sigma_k)^{z_{nk}}$

$\hookrightarrow$ Log likelihood is $\log(p(X, Z \mid \mu, \Sigma, \pi)) = \sum_{n,k}^{N,K} z_{nk}(\log \pi_k + \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k))$

$\hookrightarrow$ $E[\log(p(X, Z \mid \mu, \Sigma, \pi))] = \sum_{n,k}^{N,K} p(z_k^n = 1 \mid x_n)(\log \pi_k + \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k))$

$\hookrightarrow$ From **Eq.1**, we find that the expected log-likelihood simplifies to $-\sum_{n,k}^{N,K} r_{nk} \frac{||x_n - \mu_k||^2}{2\epsilon} + const.$

$\hookrightarrow$ It is thus clear to see that on maximizing expected log likelihood, we minimize the distortion function.

2 Density Estimation

Consider a histogram-like density model in which the space x is divided into fixed regions for which density $p(x)$ takes constant value h_i over ith region, and that the volume of region i in denoted as Δ_i. Suppose we have a set of N observations of x such that n_i of these observations fall in regions i.

(a) What is the log-likelihood function? [8 pts]

$\hookrightarrow$ Given that there are P regions, the Log-likelihood can be written as below.

$\hookrightarrow$ $\log(\prod_{n=1}^{N}(p(x^n))) = \log \prod_{i=1}^{L} h_i^{n_i} = \sum_{i=1}^{P} n_i \log h_i$.

(b) Derive an expression for the maximum likelihood estimator for h_i. [10 pts]

Hint: This is a constrained optimization problem. Remember that $p(x)$ must integrate to unity. Since $p(x)$ has constant value h_i over region i, which has volume Δ_i. The normalization constraint is $\sum_i h_i \Delta_i = 1$. Use Lagrange multiplier by adding $\lambda(\sum_i h_i \Delta_i - 1)$ to your objective function.

$\hookrightarrow$ We have $L = \sum_{i=1}^{P} n_i \log h_i + \lambda(\sum_i h_i \Delta_i - 1)$.

$\hookrightarrow$ Apply partial derivative w.r.t h_i and put it to 0.

$\hookrightarrow$ $\frac{\partial L}{\partial h_i} = \frac{n_i}{h_i} + \lambda \Delta_i = 0 \implies h_i = -\frac{n_i}{\lambda \Delta_i}$

$\hookrightarrow$ Similarly apply the partial derivative w.r.t λ and put it to 0.

$\hookrightarrow$ $L = \sum_{i=1}^{P} n_i(-\log n_i - \log \lambda - \log \Delta_i) + \lambda(-\sum_{i=1}^{P} \frac{n_i}{\lambda} - 1)$

↪ $\frac{\partial L}{\partial \lambda} = -\frac{\sum_{i=1}^{P} n_i}{\lambda} - 1 = -\frac{N}{\lambda} - 1 = 0 \implies \lambda = -N$

↪ From the above two equations, we get $h_i = \frac{n_i}{N\Delta_i}$

(c) Mark T if it is always true, and F otherwise. Briefly explain why. [12 pts]

- Non-parametric density estimation usually does not have parameters.

↪ **F** - Non parametric does not mean it does not have parameters, it means it cannot be described by a fixed number of parameters.

- The Epanechnikov kernel is the optimal kernel function for all data.

↪ **F** - This kernel is optimal only in the sense of MSE (Mean squared error). Optimality depends on the data.

- Histogram is an efficient way to estimate density for high-dimensional data.

↪ **F** - If you take into account computation, if the number of bins n^d is more than number of samples, then most bins are empty. This is clearly sub-optimal.

- Parametric density estimation assumes the shape of probability density.

↪ **T** - Parametric density estimation assumes the shape of probability density by choosing appropriate parameters.

3 Information Theory

In the lecture you became familiar with the concept of entropy for one random variable and mutual information. For a pair of discrete random variables X and Y with the joint distribution $p(x, y)$, the *joint entropy* $H(X, Y)$ is defined as

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y) \tag{3}$$

which can also be expressed as

$$H(X, Y) = -\mathbb{E}[\log p(X, Y)] \tag{4}$$

Let X and Y take on values $x_1, x_2, ..., x_r$ and $y_1, y_2, ..., y_s$ respectively. Let Z also be a discrete random variable and $Z = X + Y$.

(a) Prove that $H(X, Y) \leq H(X) + H(Y)$ [4 pts]

- We prove the above as follows :

↪ We know that $H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$

↪ $-H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x)p(y \mid x)]$

↪ $-H(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(x)] + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log[p(y \mid x)]$

$\hookrightarrow$ $-H(X,Y) = \sum_{x\in X} \log[p(x)] \sum_{y\in Y} p(x,y) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log[p(y \mid x)]$

$\hookrightarrow$ $-H(X,Y) = \sum_{x\in X} \log[p(x)]p(x) + \sum_{x\in X}\sum_{y\in Y} p(x,y)\log[p(y \mid x)]$

$\hookrightarrow$ Finally, we get $H(X,Y) = H(Y \mid X) + H(X)$

- Consider the quantity $H(Y \mid X) - H(Y)$.

$\hookrightarrow$ $H(Y \mid X) - H(Y) = -\sum_{x\in X}\sum_{y\in Y} p(x)p(y \mid x)\log[p(y \mid x)] + \sum_{y\in Y} p(y)\log[p(y)]$

$\hookrightarrow$ $H(Y \mid X) - H(Y) = -\sum_{x\in X} p(x)\sum_{y\in Y} p(y \mid x)\log[p(y \mid x)] + \sum_{y\in Y} p(y)\log[p(y)]\sum_{x} p(x \mid y)$

$\hookrightarrow$ We know $p(x \mid y)p(y) = p(y \mid x)p(x) = p(x,y)$

$\hookrightarrow$ $H(Y \mid X) - H(Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)(\log(p(y \mid x)) - log(p(y)))$

$\hookrightarrow$ $H(Y \mid X) - H(Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\dfrac{p(y \mid x)}{p(y)}$

$\hookrightarrow$ $H(Y \mid X) - H(Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\dfrac{p(x,y)}{p(x)p(y)}$

- Consider some random variable A that takes value $\dfrac{p(x)p(y)}{p(x,y)}$ with probability $p(x,y)$.
- Using Jensen's inequality we see as below:

$\hookrightarrow$ $\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\dfrac{p(x)p(y)}{p(x,y)} \leq \log\sum_x\sum_y \dfrac{p(x)p(y)}{p(x,y)}p(x,y) = \log 1 = 0$

$\hookrightarrow$ Hence, we have $H(Y \mid X) - H(Y) \leq 0$ which implies $H(X,Y) \leq H(X) + H(Y)$

(b) Show that $I(X;Y) = H(X) + H(Y) - H(X,Y)$. [2 pts]

$\hookrightarrow$ We know that $H(X,Y) = H(X) + H(Y \mid X)$.

$\hookrightarrow$ Also, $I(X;Y) = H(Y) - H(Y \mid X)$

$\hookrightarrow$ Adding the above two, we get as below.

$\hookrightarrow$ $I(X,Y) + H(X,Y) = H(X) + H(Y)$ or equivalently $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(c) Under what conditions does $H(Z) = H(X) + H(Y)$. [4 pts]

- Given that $Z = X + Y$, we have $p(Z = z \mid X = x) = p(Y = z - x \mid X = x)$

$\hookrightarrow$ $H(Z \mid X) = -\sum_{x\in X} p(X = x)\sum_{z\in Z} p(Z = z \mid X = x)\log p(Z = z \mid X = x)$

$\hookrightarrow$ $H(Z \mid X) = -\sum_{x\in X} p(X = x)\sum_{z\in Z} p(Y = z - x \mid X = x)\log p(Y = z - x \mid X = x)$

$\hookrightarrow$ $H(Z \mid X) = H(Y \mid X)$

$\hookrightarrow$ Similarly, it is clear to see that $H(Z \mid Y) = H(X \mid Y)$

$\hookrightarrow$ Therefore, $H(Z) = H(Z \mid X) + H(Z \mid Y) = H(Y \mid X) + H(X \mid Y)$

$\hookrightarrow$ So if X and Y are independent, $H(X \mid Y) = H(X)$ and $H(Y \mid X) = H(Y)$

$\hookrightarrow$ Therefore, in this case - $H(Z) = H(X) + H(Y)$

4 Programming: Text Clustering

Multinomial Distribution

The simplest distribution representing a text document is multinomial distribution(Bishop Chapter 2.2). The probability of a document D_i is:

$$p(D_i) = \prod_{j=1}^{n_w} \mu_j^{T_{ij}}$$

Here, μ_j denotes the probability of a particular word in the text being equal to w_j, T_{ij} is the count of the word in document. So the probability of document D_1 would be $p(D_1) = \mu_1^2 \cdot \mu_2^6 \cdot ... \cdot \mu_{n_w}^4$.

Mixture of Multinomial Distributions

EM for Mixture of Multinomials

In order to cluster a set of documents, we need to fit this mixture model to data. In this problem, the EM algorithm can be used for fitting mixture models. This will be a simple topic model for documents. Each topic is a multinomial distribution over words (a mixture component). EM algorithm for such a topic model, which consists of iterating the following steps:

1. Expectation

   Compute the expectation of document D_i belonging to cluster c:

$$\gamma_{ic} = \frac{\pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}{\sum_{c=1}^{n_c} \pi_c \prod_{j=1}^{n_w} \mu_{jc}^{T_{ij}}}$$

2. Maximization

   Update the mixture parameters, i.e. the probability of a word being W_j in cluster (topic) c, as well as prior probability of each cluster.

$$\mu_{jc} = \frac{\sum_{i=1}^{n_d} \gamma_{ic} T_{ij}}{\sum_{i=1}^{n_d} \sum_{l=1}^{m_w} \gamma_{ic} T_{il}}$$

$$\pi_c = \frac{1}{n_d} \sum_{i=1}^{n_d} \gamma_{ic}$$

Task [20 pts]

Implement the algorithm and run on the toy dataset `data.mat`. You can find detailed description about the data in the `homework2.m` file. Observe the results and compare them with the provided true clusters each document belongs to. Report the evaluation (e.g. accuracy) of your implementation.

Hint: We already did the word counting for you, so the data file only contains a count matrix like the one shown above. For the toy dataset, set the number of clusters $n_c = 4$. You will need to initialize the parameters. Try several different random initial values for the probability of a word being W_j in topic c, μ_{jc}. Make sure you normalized it. Make sure that you should not use the true cluster information during your learning phase.

- The details are as below:

↪ The μ matrix was initialized randomly with np.random.rand and then normalized

↪ The π matrix was initialized to [0.25, 0.25, 0.25, 0.25] because $n_c = 4$ is given.

↪ The results over 20 runs (100 iterations each run) are as follows : **Min acc.** = 59.75%; **Max acc.** = 91.25%; **Mean acc.** = 78.83%.

↪ It is also clear to see that the code depends on initial conditions. For different initial conditions, it gives varying results.