CSE 6740 Homework 2

Anqi Wu, Spring 2025

Deadline: Feb 27 Thursday, 9:30 am

- There are 2 sections in gradescope: Homework 2 and Homework 2 Programming. Submit your answers as a PDF file to Homework 2 (including report for programming) and also submit your code in a zip file to Homework 2 Programming.
- All Homeworks are due by the beginning of class. Homework is penalized by 20% for each day that it is late (this applies additively, meaning that no credit is gained after 5 late days).
- We strongly encourage the use of LaTeX for your submission. Unreadable handwriting is subject to zero credit.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 1.5, 1.6, 2.5, 9.2, 9.3, 9.4

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 $0 \le \pi_k \le 1$ 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^i, its latent variable z^i is given by $[0; 1; ...; 0] = z^{(2)}$. With this representation, we can express $p(z)$ as

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

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

where z_k indicates kth element of vector z.

Also, $p(x|z)$ can be represented similarly as

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

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]

(b) In reality, we do not know which component each data point is from. Thus, we estimate the responsibility (expectation of z_k^i) in the E-step of EM. Since z_k^i is either 1 or 0, its expectation is the probability for the point x_i to belong to the component z_k. In other words, we estimate $p(z_k^i = 1|x_i)$. [5 pts]

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 $p(z_k^i|x_i)$ as a function of these fixed parameters.
Hint: Derive the formula for this estimation by using Bayes rule.

(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 $x_1, ..., x_N$, write down the log likelihood function, and derive the update formula for each parameter. [8 pts]

Note that in order to obtain an update rule for the M-step, we fix the responsibilities, i.e. $p(z_k^i|x_i)$, which we have already calculated in the E-step.
Hint: Use Lagrange multiplier for π_k to apply constraints on it.

(d) Establish the convergence of the EM algorithm by showing that the log likelihood function will be nondecreasing in each iteration of this algorithm. [8 pts]

Hint: Use Jensen's inequality: for concave function $f(x)$, we have $f(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i f(x_i)$, where $\sum_i \alpha_i = 1, \alpha_i \geq 0$.

(e) EM and K-Means [5 pts]

K-means can be viewed as a particular limit of EM for Gaussian mixture. Briefly describe the expectation and maximization steps in K-Means Algorithm.

2 Density Estimation

(a) Given a sequence of m independently and identically distributed (iid) data points $D = \{x^1, x^2, x^3, \ldots, x^m\}$, where each x^i is a nonnegative integer (i.e., $x^i \in \{0, 1, 2, \ldots\}$) following a Poisson distribution. The Poisson distribution is given by:

$$P(x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{3}$$

where $\lambda > 0$ is both the mean and the variance of the distribution (rate parameter). Find the Maximum Likelihood Estimation (MLE) for λ. [8 pts]

(b) Given $D = \{x^1, x^2, x^3, \ldots, x^m\}$ as above (i.e., $x^i \in \{0, 1, 2, \ldots\}$ from a Poisson distribution), use results from (a), show the estimated mean and variance. [4 pts]

(c) Given $D = \{x^1, x^2, x^3, \ldots, x^m\}$, where each $x^i \in \{0, 1, 2, \ldots\}$, consider a simple frequency-based (non-parametric) estimator for the discrete pmf. Let C_k denote the number of samples in D that equal k (i.e., $C_k = \sum_{i=1}^{m} I\{x^i = k\}$). For $k = 0, 1, 2, \ldots$, define the estimator

$$p(k) \;=\; \frac{C_k}{m}. \tag{4}$$

Discuss the conditions for a valid pmf and prove that this frequency-based estimator satisfies these conditions. [6 pts]

3 Information Theory

In the lecture you became familiar with the concept of entropy for one random variable and mutual information. One property of mutual information is $I(X, Z) \geq 0$, where the larger the value, the greater the relationship between the two variables.

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) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$. Thus the addition of independent random variables adds uncertainty. [8 pts]

(b) Give an example of (necessarily dependent) random variables X, Y in which $H(X) > H(Z)$ and $H(Y) > H(Z)$. [2 pts]

(c) Let Z be a discrete random variable defined as $Z = X + Y$. Is the independence between X and Y a necessary condition, a sufficient condition, or both for the equality $H(Z) = H(X)+H(Y)$ to hold true? Justify your answer. [6 pts]

4 Programming: Gaussian Mixture Models

During lecture, we introduced Gaussian Mixture Models (GMMs). Here we ask you to implement your own GMM with `numpy`. Please follow the instruction in the `submission.py` to implement the required function. Specifically, you are asked to implement

1. `initialize_parameters` [5pt]: initialize parameters uniformly.
2. `compute_sigma` [5pt]: compute the covariance matrix.
3. `prob` [5pt]: compute probability of a n dimensional Gaussian.

4. `E_step` [5pt]: perform the expectation step.
5. `M_step` [5pt]: perform the maximization step.
6. `likelihood` [5pt]: compute the loglikelihood.
7. `train_model` [10pt]: train the GMM model until convergence (convergence criterion provided).

For submission, you are required to submit the `submission.py` file to Gradescope.