CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi

DIS7

1 Gaussian Mixture Models

Let Z represent the (unobserved) assignment of a given observation to one of the K clusters:

$$Z \sim \text{Categorical}(\pi_1, \ldots, \pi_K),$$

where π_k is the probability that a randomly selected observation is assigned to cluster k. Conditioned on Z, observations are assumed to be Gaussian distributed,

$$X \mid Z = i \sim \mathcal{N}(\mu_i, \Sigma_i).$$

Here, μ_i and Σ_i are the mean and covariance matrix of the i-th cluster.

We let $(X_1, Z_1), \ldots, (X_n, Z_n)$ denote the set of observations and their corresponding cluster assignments, under i.i.d. assumptions.

(a) What is the set of parameters θ that we can learn from the data?

(b) Write down the joint log-likelihood function for a single observation X_i and its corresponding cluster assignment Z_i, $\log p_\theta(X_i, Z_i)$.

(c) Why is maximizing $\sum_{i=1}^{n} \log p_\theta(X_i, Z_i)$ impossible?

(d) Instead, we consider the marginalized log-likelihood function, $\ell_{\text{marginal}}(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i)$. Write down a formula for $\ell_{\text{marginal}}(\theta)$.

(e) Suggest an iterative strategy to learn θ? What guarantees would this approach provide?

2 The EM algorithm

This question is the second part of the previous question; all notations and assumptions are the same.

Another prevalent approach for fitting Gaussian Mixture Models, and other latent variable models, is to use the so-called Expectation-Minimization algorithm. While we won't cover the details of the EM implementation in this discussion, we here provide a high-level overview of the algorithm.

Instead of maximizing the marginalized log-likelihood function, the EM algorithm aims to maximize a lower bound $\mathcal{F}(q, \theta)$ on the marginalized log-likelihood function, such that

$$\ell_{\text{marginal}}(\theta) \geq \mathcal{F}(q, \theta) = \sum_{i=1}^{n} \mathcal{F}_i(q_i, \theta), \tag{1}$$

where $\mathcal{F}_i(q_i, \theta) := \sum_{z=1}^{K} q_i(z) \log \frac{p_\theta(X_i, Z_i = z)}{q_i(z)}$.

Here, q_i can be seen as an arbitrary distribution over the K clusters for the i-th observation. Because $\mathcal{F}$ has two arguments, the EM algorithm will iteratively aim to optimize over both q_i and θ, as we will see later.

(a) We will first demonstrate Equation (1). Show that for an arbitrary data point i, the following inequality holds for any distribution $q_i(z)$ over cluster assignments:

$$\log p_\theta(X_i) \geq \sum_{z=1}^{K} q_i(z) \log \frac{p_\theta(X_i, Z_i = z)}{q_i(z)} = \mathcal{F}_i(q_i, \theta).$$

Then, show that Equation (1) holds. This inequality is extremely important, and serves as the basis of the EM algorithm, as well as other important algorithms in machine learning, such as variational autoencoders.

Hint: You might find the following application of Jensen inequality useful. For any $\alpha_1, \dots \alpha_K$ s.t. $\sum_i \alpha_i = 1$, and any positive $f_1, \dots f_K$

$$\sum_i \alpha_i \log f_i \leq \log \sum_i \alpha_i f_i.$$

(b) The inequality we have showed holds for any distributions $q_1, \dots q_n$. For a fixed θ, the EM algorithm aims to optimize over the distributions $q_1, \dots q_n$, in order to make $\mathcal{F}(q, \theta)$ as close as possible to $\ell_{\text{marginal}}(\theta)$. Once the optimal q_i are found, θ is updated to maximize $\mathcal{F}(q, \theta)$. This yields the following iterative update at iteration t of the algorithm:

$$\begin{cases} q_i^{(t+1)} & = \arg\max_{q_i} \mathcal{F}(q_i, \theta^{(t)}) \quad \text{(E-step)} \\ \theta^{(t+1)} & = \arg\max_{\theta} \mathcal{F}(q^{(t+1)}, \theta) \quad \text{(M-step)} \end{cases}$$

Let θ be fixed. Show that when for any $i \leq N$ and $z \leq K$, $q_i(z) = p_\theta(Z_i = z \mid X_i)$,

$$\ell_{\text{marginal}}(\theta) = \sum_{i=1}^{n} \mathcal{F}_i(q_i, \theta).$$

What optimal q_i should be used in the E-step?