CS 189 Introduction to Machine Learning
Spring 2018 DIS10

1 One Dimensional Mixture of Two Gaussians

Suppose we have a mixtures of two Gaussians in $\mathbb{R}$ that can be described by a pair of random variables (X, Z) where X takes values in $\mathbb{R}$ and Z takes value in the set $1, 2$. The joint-distribution of the pair (X, Z) is given to us as follows:

$$Z \sim \text{Bernoulli}(\beta),$$
$$(X|Z = k) \sim \mathcal{N}(\mu_k, \sigma_k), \quad k \in 1, 2,$$

We use $\boldsymbol{\theta}$ to denote the set of all parameters $\beta, \mu_1, \sigma_1, \mu_2, \sigma_2$.

(a) Write down the expression for the joint likelihood $p_{\boldsymbol{\theta}}(X = x_i, Z_i = 1)$ and $p_{\boldsymbol{\theta}}(X = x_i, Z_i = 2)$. What is the marginal likelihood $p_{\boldsymbol{\theta}}(X = x_i)$?

(b) What is the log-likelihood $\ell_{\boldsymbol{\theta}}(\mathbf{x})$? Why is this hard to optimize?

(c) (Optional) You'd like to optimize the log likelihood: $\ell_{\boldsymbol{\theta}}(x)$. However, we just saw this can be hard to solve for an MLE closed form solution. Show that a lower bound for the log likelihood is $\ell_{\boldsymbol{\theta}}(x_i) \geq \mathbb{E}_q \left[\log \left(\frac{p_{\boldsymbol{\theta}}(X=x_i, Z_i=k)}{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)} \right) \right]$.

(d) (Optional) The EM algorithm first initially starts with two randomly placed Gaussians (μ_1, σ_1) and (μ_2, σ_2), which are both particular realizations of $\boldsymbol{\theta}$.

- E-step: $\mathbf{q}_{i,k}^{t+1} = p_{\boldsymbol{\theta}}(Z_i = k|X = x_i)$. For each data point, determine which Gaussian generated it, being either (μ_1, σ_1) or (μ_2, σ_2).
- M-step: : $\boldsymbol{\theta}^{t+1} = \text{argmax}_\theta \sum_{i=1}^n \mathbb{E}_q \left[\log\big(p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)\big) \right]$. After labeling all data-points in the E-step, adjust (μ_1, σ_1) and (μ_2, σ_2).

Why does alternating between the E-step and M-step result in maximizing the lower bound?

(e) E-step: What are expressions for probabilistically imputing the classes for all the datapoints, i.e. $q_{i,1}^{t+1}$ and $q_{i,2}^{t+1}$?

(f) What is the expression for μ_1^{t+1} for the M-step?

(g) Compare and contrast k-means, soft k-means, and mixture of Gaussians fit with EM.

