EECS 189 Introduction to Machine Learning
Fall 2020

# DIS10

This discussion was released **Friday, November 6**.

This discussion serves as an introduction to the EM algorithm. We will first start with this Jupyter notebook. We will then come back to this worksheet for a more theoretical understanding of EM algorithm.

As a reminder, if you have questions, we will answer them via the queue at oh.eecs189.org. Once you complete the Jupyter notebook, please return to this worksheet.

## 1 Jupyter Notebook

## 2 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:

$$Z \sim \text{Bernoulli}(0.5),$$
$$(X|Z = k) \sim \mathcal{N}(\mu_k, \sigma_k), \quad k \in 1, 2,$$

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

(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) It turns out that we can compute the log-likelihood easily.

$$\begin{aligned}
\ell_{\boldsymbol{\theta}}(\mathbf{x}) &= \ln(p_{\boldsymbol{\theta}}(X = x_1, \ldots, X = x_n)) \\
&= \sum_{i=1}^{n} \ln(p_{\boldsymbol{\theta}}(X = x_i)) \\
&= \sum_{i=1}^{n} \ln\left[\frac{1}{2}\mathcal{N}(x_i|\mu_1, \sigma_1^2) + \frac{1}{2}\mathcal{N}(x_i|\mu_2, \sigma_2^2)\right]
\end{aligned}$$

Here, we use $\mathcal{N}(x|\mu, \sigma^2)$ as shorthand for the Gaussian density evaluated at $x$ for a Normal random variable with mean $\mu$ and variance $\sigma^2$.

This log-likelihood can be optimized, but not analytically. Taking the derivative with respect to $\mu_1$, for example, would give:

$$\frac{\partial \ell_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \mu_1} = \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_1, \sigma_1^2)}{\mathcal{N}(x_i|\mu_1, \sigma_1^2) + \mathcal{N}(x_i|\mu_2, \sigma_2^2)}(\frac{x_i - \mu_1}{\sigma_1^2})$$

This ratio of exponentials and linear terms makes it difficult to analytically solve for the maximum likelihood estimate.

Anyway, we still want 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 a single term in the log likelihood is** $\ell_{\boldsymbol{\theta}}(x_i) \geq \mathbb{E}_q\left[\ln\left(\frac{p_{\boldsymbol{\theta}}(X=x_i,Z_i=k)}{q_{\boldsymbol{\theta}}(Z_i=k|X=x_i)}\right)\right]$**.** Here, this bound should hold for any distribution $q_{\boldsymbol{\theta}}(Z_i = k|X = x_i)$. The expectation in the expression above to the right is using $q$ to treat $k$ as a random variable.

Here, you should start with:

$$\ell_{\boldsymbol{\theta}}(x_i) = \ln\left(\sum_k p_{\boldsymbol{\theta}}(X = x_i, Z_i = k)\right)$$

and go from there. You don't have to worry about the details of Gaussians for this problem.

*(HInt: At a high level, look at what you are trying to prove. There are three things you clearly need to do just by looking at the patterns: (1) Somehow introduce the distribution q into the problem; (2) Somehow turn the sum over k into an expectation; (3) Somehow get that expectation/sum outside the logarithm.*

We will stop here due to time limit and continue the theoretical setup in your homework.

Contributors:

- Anant Sahai
- Jane Yu
- Peter Wang