Machine Learning Homework #4

Joe Ellis - jge2105

November 20, 2013

1 Problem 1 – EM Derivation

In this problem we will derive the Expectation-Maximization algorithm for a mixture of multinomials. Suppose, that x is represented as a vector such that $x(j) = 1$ if x takes the j^{th} value, and $\sum_j x(j) = 1$. The distribution of x is described by a mixture of K discrete multinomials such that:

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

where

$$p(x|\mu_k) = \prod_{j=1}^{M} \mu_k(j)^{x(j)}. \tag{2}$$

Where π_k is the mixing coefficients, and μ_k specifies the parameters of the k^th component. We then must derive the Expectation and Maximization steps to maximize the log-likelihood of an observed data set $\{x_i\}$.

1.1 Expectation Step

First we want to calculate what is referred to in EM as the responsibilities of each data point given the mixtures. To do this we want to find the $p(z_i|x_i;\theta)$, where z_i is the

mixture that this data point belongs to. In this way, we can estimate the likelikhood that each data point was generated by one of the given mixtures, and then weight our estimates in the maximization step accordingly.

$$p(z_i|x_i|;\theta) = \frac{p(x_i|z_i;\theta)p(z_i)}{p(xi)},\ byBayesRule \tag{3}$$

$$= \frac{p(x_i|z_i;\theta)\pi_k}{p(xi)} \tag{4}$$

$$= \frac{p(x_i|\mu_k)\pi_k}{\sum_{k=1}^{K}\pi_k p(x|\mu_k)} \tag{5}$$

$$= \frac{\pi_k \mu_k^{x_i(j)}}{\sum_{k=1}^{K}\pi_k \mu_k^{x_i(j)})}. \tag{6}$$

$$\tag{7}$$

Therefore, we have the responsibilities of each data point, and have calculated how likely each data point was generated from each mixture.

1.2 Maximization Step

Now that we have the responsibilities we will in turn now calculate the parameters μ_k and π_k that maximize the log-likelihood function of the seen data. Let's find the log-likelihood of the data given our hidden variables θ, to do this, we must maximize the general form of the equation below.

$$\theta := \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{z_i}(Q_i(z_i)\frac{p(x_i, z_i;\theta)}{Q_i(z_i)}). \tag{8}$$

In the above equation $Q_i(z_i)$ is the value for each mixture and data point that is calculated using the E-step from above. In the derivation below we will use the notation $\tau_{k,i}$, where $\tau_{k,i} = Q_i(z_k)$. Let's refer to the above equation as $l(\theta)$

$$l(\theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\frac{p(x_i, z_i; \theta)}{\tau_{k,i}} \tag{9}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(p(x_i, z_i; \theta)) - \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\tau_{k,i}) \tag{10}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(p(x_i, z_i; \theta)) - const \tag{11}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\pi_k \prod_{j=1}^{M} \mu_k(j)^{x_i(j)}) \tag{12}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\pi_k) + \sum_{j=1}^{M} log(\mu_k(j)^{x_i(j)}) \tag{13}$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\pi_k) + \tau_{k,i} \sum_{j=1}^{M} x_i(j) log(\mu_k(j)) \tag{14}$$

$$. \tag{15}$$

Now that we have derived the log likelihood we must maximize the above equation with respect to the parameters μ and π. However the derivation is not straight forward and we must use laplaclian variables to satisfy the constraints of, $\sum_j \mu_k(j) = 1$ and $\sum_k \pi_k = 1$.

Therefore, let's do the derivation. We set the derivative with respect to $\mu_k(j)$ and the constraints with laplace variable λ to zero and solve.

$$\frac{\partial l(\theta)}{\partial \mu_k(j)} = \frac{\partial}{\partial \mu_k(j)} \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{k,i} log(\pi_k) + \tau_{k,i} \sum_{j=1}^{M} x_i(j) log(\mu_k(j)) - \sum_{k=1}^{K} \lambda_k (\sum_{j=1}^{M} \mu_k(j) - 1) \tag{16}$$

$$= \sum_{i=1}^{N} \tau_{k,i} \frac{1}{\mu_k(j)} - \lambda_k = 0. \tag{17}$$

$$\tag{18}$$

Thus,

$$\mu_k(j) = \frac{\sum_{i=1}^{N} x_i(j) \tau_{k,i}}{\lambda_k}. \tag{19}$$

Plugging $\mu_k(j)$ back into the equation, $\sum_{j=1}^{M} \mu_k(j) - 1)$ gives us the equation for λ_k below,

$$\lambda_k = \sum_{i=1}^{N} \sum_{m=1}^{M} \tau_{k,i} x_i(m). \tag{20}$$

So therefore we have the update for our $\mu_k(j)$ as,

$$\mu_k(j) = \frac{\sum_{i=1}^{N} \tau_{k,i} x_i(j)}{\sum_{i=1}^{N} \sum_{m=1}^{M} \tau_{k,i} x_i(m)} \tag{21}$$

Similarly, we can derive the value for the other parameter π_k, with laplacian constraints.

$$\frac{\partial l(\theta)}{\partial \pi_k} = \frac{\partial}{\partial \pi_k} \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{k,i} log(\pi_k) + \tau_{k,i} \sum_{j=1}^{M} x_i(j) log(\mu_k(j)) - \lambda(\sum_{k=1}^{K} \pi_k - 1) \tag{22}$$

$$= \sum_{i=1}^{N} \tau_{k,i} \frac{1}{\pi_k} - \lambda = 0. \tag{23}$$

$$\tag{24}$$

Thus,

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{k,i}}{\lambda}. \tag{25}$$

Plugging λ, back into the constraint equation, $\sum_{k=1}^{K} \pi_k - 1)$, we solve for lambda,

$$\lambda = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{k,i} \tag{26}$$

$$= N. \tag{27}$$

So, therefore we have the update for our mixing coefficients (π_k) as,

$$\pi_k = \frac{\sum_{i=1}^{N} \tau_{k,i}}{N} \tag{28}$$

$$\tag{29}$$

Therefore, we have derived the entire EM algorithm for a mixture of multinomials.

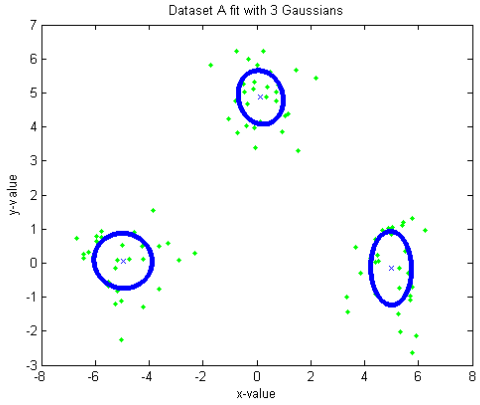

Figure 1: EM result on dataset A using 3 Gaussian distributions

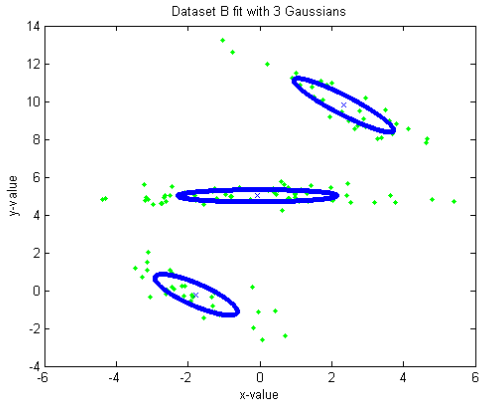

Figure 2: EM result on dataset B using 3 Gaussian distributions

2 Problem 2

2.1 Mixtures of Gaussians

Using the code provided to us in the tutorial section of the class website we were able to perform EM on a mixture of gaussians on two datasets. The figures generated from the EM algorithm and the gaussian distributions found are superimposed over the data points.

Figures 1 and 2 show the results of the EM calculation.

2.2 Mixtures of Multinomials

In this section we explore using the derivation provided in section 1 to create a mixture of multinomials model to differentiate between plays written by Shakespeare and Middleton. We have 9 plays written by Shakespeare and 9 plays written by Middleton, and we will

try to classify how we well we can seperate the two using multiple mixture models.

3 Problem 3