Homework 3

Aniket Kesari
25948127
UC Berkeley

October 16, 2017

EM Algorithm

Suppose $X_1, ..., X_n$ are i.i.d. observations from a mixture of two Poisson distributions:

$P_0(X) = \frac{\mu_0^X e^{-\mu_0}}{X!}$

$P_1(X) = \frac{\mu_1^X e^{-\mu_1}}{X!}$

with mixing probabilities of π and $1 - \pi$ (i.e. there is an intial probability that an observation X_i is drawn from P_0 and a probability of $1 - \pi$ from P_1)

Define the complete data vector and the distribution of the missing random variable

The parameters I am estimating:

$\theta = (\pi, \mu_0, \mu_1)$

The complete data vector is:

$[X, Y, Z]$

Where X is the observed data, Y is the observed outcome data, and Z is the latent variable matrix.

Write down the E and M steps for estimating μ_0, μ_1, π

E-Step

1. First, define the likelihood of observing X given the parameters is defined as the sum (over the number of components) of the mixing distributions, multiplied by the density functions, for the observed data given the parameters.

$L(X; \theta) = \sum_{j=1}^{2} p_j g_j(X; \mu_k)$

2. Because both of the groups are coming from Poisson distributions, we can express this in terms of a product.

$L(X; \theta) = \prod_{i=1}^{n} \sum_{j=1}^{2} p_j g(X_i; \mu_j)$

3. Re-writing again into a log like-lihood:

$loglik(X; \theta) = log \prod_{i=1}^{n} \sum_{j=1}^{2} p_j g(X_i, \mu_j) = \sum_{i=1}^{n} log \sum_{j=1}^{2} p_j g(X_i; \mu_j)$

Again $g(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}$

4. The problem can be summarized as: $\theta^* = argmax_\theta loglik(X; \theta)$

5. Define:

$q(j,i) = p_j g(X_i, \mu_j)$ as the joint probability of selecting component p_j and selecting an observation from that component $(g(X_i; \mu_i))$.

6. Define:

$p(j|i) = \frac{q(j,i)}{\sum_m}^2 q(m,i)$ where m is the "membership indicator," therefore this entire expression can be considered the "membership probability." More simply, this states the probability that a given X_i is a member of component j.

7. Jensen's inequality states that:

$log \sum_{j=1}^2 \pi_j \alpha_j \geq \sum_{j=1}^2 \pi_j log \alpha_j$

This says that the log of a sum is always greater than or equal to the sum of a log. Given this, we can rewrite:

$log \sum_{j=1}^2 c_j = log \sum_{j=1}^2 c_k \frac{\pi_j}{\pi_j} = log \sum_{j=1}^2 \pi_j \frac{c_j}{\pi_j} \geq \sum_{j=1}^2 \pi_j log \frac{c_j}{\pi_j}$

Restating the E-step:

$p(j|i) = \frac{p_j g(X_i;\mu_j)}{\sum_m^2 p_j g(X_i;\mu_j)}$

Applying Jensen's inequality to this:

$loglik(X;\theta) = \sum_{i=1}^n log \sum_{j=1}^2 q(j,i) \geq \sum_{i=1}^n \sum_{j=1}^2 p(j|i) log \frac{q(j,i)}{p(j|i)} = b(\theta)$

In words, this basically says that the loglikelihood of observing the data given the parameters is bounded by the double summation of the member probability, multiplied by the likelihood of observing membership given the Poisson distribution over the total membership probability.

8. Expanding the log:

$b(\theta) = \sum_{i=1}^n \sum_{j=1}^2 p(k|i) log q(j,i) - \sum_{i=1}^n \sum_{j=1}^2 p(j|i) log p(j|i)$

The membership probability values are fixed, so we can drop these from the function. Now:

$Q(\theta) = \sum_{i=1}^n \sum_{j=1}^2 p(j|i) log q(j,i)$

M-Step

The process for maximizing is:

1. Differentiate Q with respect to θ_i
2. Set the derivative equal to 0
3. Solve for θ_i
4. Differentiate

$\frac{\delta Q}{\delta \mu_j} = \frac{\delta}{\delta \mu_j} \sum_{i=1}^n \sum_m^2 p(m|i) log q(m,i)$

2. The derivative of a sum if the sum of derivatives, and $p(j|i)$ is a constant. Any term that does not involve j is 0, so we can take these out:

$\frac{\delta Q}{\delta \mu_j} = \sum_{i=1}^n p(j|i) \frac{\delta}{\delta \mu_j} log q(j,i)$

3. Replace q (probability of component multiplied by density) and g (Poisson density) with their definitions

$\frac{\delta Q}{\delta \mu_j} = \sum_{i=1}^n p(j|i) \frac{\delta}{\delta \mu_j} log p_j g(X_i : \mu_j$

$\frac{\delta Q}{\delta \mu_j} = \sum_{i=1}^n p(j|i) \frac{\delta}{\delta \mu_j} log p_j \frac{\mu_j^{X_i}}{X_i!} e^{-\mu_j}$

4. Expand and simplify the log:

$\frac{\delta Q}{\delta \mu_j} = \sum_{i=1}^{n} p(j|i) \frac{\delta}{\delta \mu_j} log p_j + log \mu_k^{X_i} - log X_i! + log e^{-\mu_j}$

$\frac{\delta Q}{\delta \mu_j} = \sum_{i=1}^{n} p(j|i) \frac{\delta}{\delta \mu_j} log p_j + X_i log \mu_j - log X_i! - \mu_j$

5. Evaluate the derivative and set to 0

$\frac{Q}{\delta \mu_j} = \sum_{i=1}^{n} p(j|i)(\frac{X_i}{\mu_j} - 1) = 0$

Distribute $p(j|l)$ and expand the sum:

$0 = \sum_{l}^{n} p(j|i)(\frac{X_i}{\mu_j}) - \sum_{i=1}^{n} p(j|i)$

$\sum_{i=1}^{n} p(j|i) = \sum_{l}^{n} p(j|i)(\frac{X_i}{\mu_j})$

Pull out the constant:

$\sum_{i=1}^{n} p(j|i) = \frac{1}{\mu_j} \sum_{i=1}^{n} p(j|i) X_i$

Multiply both sides by μ_i, and divde $\sum_{i=1}^{n} p(j|i)$:

$\mu_j = \frac{\sum_{i=1}^{n} p(j|i) X_i}{\sum_{i=1}^{n} p(j|i)}$

We then iterate this procedure until we converge on an estimate for the parameter.

Give an initial estimator to start the EM algorithm

I use k-means with two clusters as a starting point. K-means should perform fairly well in creating an initial separation of the data into two groups, even if it's not perfect. Because each iteration

Write down the E and M steps if the second distribution is actually Bernoulli(p)

The major difference between this problem and when the likelihood function is defined as:

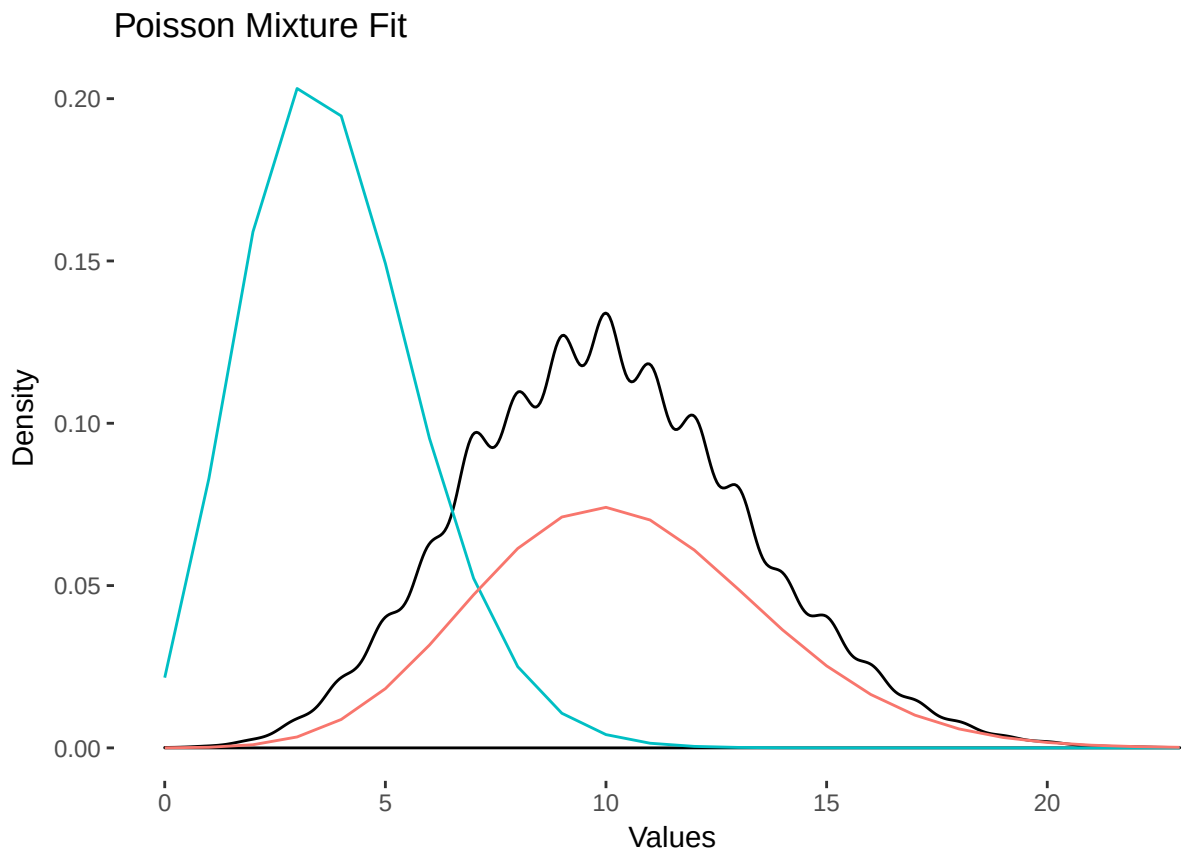
$L(X; \theta) = \sum_{j=1}^{2} p_j g_j(X; \mu_k)$

We can't pull the g_j terms into a product because they are not coming from the same distribution with different parameters.

Write R code to implement the E and M steps. Run it on some simulated data where you know the true parameters. Show the accuracy of clustering as you vary the values of μ_0 and μ_1

```
## [1] -26323.9
```

-26323.90
-25891.30
-25801.65
-25769.29
-25754.20
-25746.02
-25741.12
-25737.97
-25735.84
-25734.33
-25733.23
-25732.40
-25731.77
-25731.27
-25730.88
-25730.56
-25730.30
-25730.08
-25729.90
-25729.75
-25729.63
-25729.52
-25729.42
-25729.34
-25729.27
-25729.21
-25729.16
-25729.11
-25729.07
-25729.03
-25729.00
-25728.97
-25728.94
-25728.92
-25728.90
-25728.88
-25728.86
-25728.85
-25728.83
-25728.82
-25728.81
-25728.80
-25728.79
-25728.78
-25728.77
-25728.76
-25728.75
-25728.75
-25728.74
-25728.74



The "loglik.vector" object (printed here) shows that the EM algorithm converges pretty quickly to approximately -25,600 (across several simulations).

How would you create confidence intervals? Can you use asymptotic normality?

I can create confidence intervals for the inferred parameters μ_0 and μ_1 by calculating the Information Matrix for the sample, and then applying the following formula:

$\hat{\theta} \pm Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{I(\theta)}}$

Where $\hat{\theta}$ is the predicted parameter. The logic here is that the distribution of Maximum Likelihood Estimators are asymptoptically normal with mean θ and variance θ. The inverse of the Fisher Information Matrix ($I(\theta)$) of the sample approximates these features, therefore using asymptotic normality to derive the estimate.

The Linear Model in the Neyman-Rubin Framework

Basic model: $Y_i = T_i a_i + (1 - T_i) b_i$

Rewritten model: $Y_i = \bar{a} T_i + \bar{b}(1 - T_i) + Q_a(z_i = \bar{z}) T_i + Q_b(z_i - \bar{z})(1 - T_i) + \epsilon_i$

In your own words, give an interpretation of each term in this decomposition. What are the predictors in the model? What is the response? What are the error terms, and are they i.i.d.?

The predictors in the model are the population average under treatment or control (meaning either a or b, depending on the indicator on treatment), and the covariate under treatment or control (again depending on the indicator). The response is the individual "y," or the observed outcome for the unit, and the error terms are deviations between the observed individual point (again under treatment or control) and the predicted one (found by adding the population average to the constant adjusted covariate). The error terms are not i.i.d. because they are dependent on the indicator.

How should you choose Q_a and Q_b? Justify your choice.

Q_a and Q_b could be chosen with a propensity score matching method. Essentially, the idea behind the constant adjustment is to make sure that the covariates are balanced across both treatment and control groups (thus making the only difference between paired units whether they received treatment or not). Propensity score matching provides a useful way to match observations under control to observations under treatment based on similarity distance across the covariates.