Expectation-Maximization for Estimating Parameters for a Mixture of Poissons

Brandon Malone
Department of Computer Science
University of Helsinki

February 18, 2014

Abstract

This document derives, in excrutiating detail at some points, the EM update rules for a simple mixture of Poisson distributions. This document is largely based on [1] and [2].

1 Introduction

This work assumes a dataset is distributed according to a mixture of Poisson distributions. The goal is to learn the mixing distributions and Poisson parameters. Broadly, this can be considered the *density estimation* problem.

2 Density Estimation

The density estimation problem is a common problem in machine learning. Broadly, the problem is formulated as: given a set of N observations, $\mathcal{D}$, in some space, and a family Θ of probability density functions, find $\theta \in \Theta$ which most likely generated the data.

In this work, we will assume that Θ is a family of mixture models, so the likelihood of its members are defined as:

$$\mathcal{L}(\mathcal{D} : \theta) = \sum_{k}^{K} p_k g_k(\mathcal{D} : \lambda_k), \tag{1}$$

where $\theta = \{p_1 \dots p_k, \lambda_1 \dots \lambda_k\}$ is a set of parameters, K is the number of mixture components, p_k are the mixing probabilities ($P(z_l = k)$), g_k is the density function of the k^{th} component, and λ_k are the set of parameters for g_k. The p_ks define a probability distribution, so we also have that $0 \leq p_k \leq 1$ and $\sum_k p_k = 1$.

If we further assume that each component has the same density function (but different parameters) and that the observations are exchangeable, then we can

write:

$$\mathcal{L}(\mathcal{D} : \theta) = \prod_{l}^{N} \sum_{k}^{K} p_k g(D_l : \lambda_k). \tag{2}$$

Typically, we will work in log-space because it is more amenable to analysis and computation. Consequently, we define the log-likelihood of θ as:

$$\ell(\mathcal{D} : \theta) = \log \prod_{l}^{N} \sum_{k}^{K} p_k g(D_l; \lambda_k) = \sum_{l}^{N} \log \sum_{k}^{K} p_k g(D_l : \lambda_k). \tag{3}$$

The density function of the Poisson distribution is:

$$g(x : \lambda) = \frac{\lambda^x}{x!} \exp\{-\lambda\} \tag{4}$$

With definitions, we can define our version of the density estimation problem as:

$$\theta^* = \arg\max_{\theta \in \Theta} \ell(\mathcal{D} : \theta) \tag{5}$$

The goal is to find θ^*.

3 Notation

In order to express ideas succinctly, we use the following notation.

$$q(k, l) = p_k g(D_l : \lambda_k) \tag{6}$$

This is the joint probability of selecting component (p_k) and selecting an observation from that component ($g(D_l : \lambda_k)$). This makes explicit the assumption that selecting a component is independent of selecting an observation from a component.

We now define the conditional probability of selecting component k, given the observation D_l. First, recall one definition of conditional probability:

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{7}$$

Additionally, recall that we can find prior probabilities for B by taking the joint probability of all variables, and then summing (for discrete variables) or integrating (for continuous variables) the others out. Supposing A is the only other variable and that it is discrete, then:

$$P(B) = \sum_{A} P(A, B) \tag{8}$$

With these basic ideas of probability in mind, then, we can write the conditional probability of selecting component k given the observation D_l as:

$$p(k|l) = \frac{q(k,l)}{\sum_m^K q(m,l)} \tag{9}$$

We call these the *membership probabilities* because they give the probability that D_l is a member of component k.

One other useful observation is that $\sum_k p(k|l) = 1$. That is to say, *some* component generated each observation.

4 Jensen's Inequality

A potential difficulty in evaluating the log-likelihood (Equation 3) is that it contains the logarithm of a sum. Jensen's inequality shows that:

$$\log \sum_k^K \pi_k \alpha_k \geq \sum_k^K \pi_k \log \alpha_k, \tag{10}$$

when the π_ks define a probability distribution (i.e., $0 \leq \pi_k \leq 1$ and $\sum_k \pi_k = 1$).

With this inequality, we can define a probability distribution π and rearrange terms to see that for any values of c_k:

$$\log \sum_k^K c_k = \log \sum_k^K c_k \frac{\pi_k}{\pi_k} = \log \sum_k^K \pi_k \frac{c_k}{\pi_k} \geq \sum_k^K \pi_k \log \frac{c_k}{\pi_k}. \tag{11}$$

5 The EM Algorithm

With these tools, we can now derive the expection-maximization algorithm for finding the parameters of the Poisson mixture model. The EM algorithm begins with a (bad) set of estimates for the p_ks and λ_ks and then alternates between two steps. In the first "E" step, we use our parameter estimates to construct a bound b on ℓ using Jensen's inequality. Then, in the "M" step, we find parameter estimates which maximize the bound.

Concretely, during the "E" step, we compute the expectation that each observation came from each component. That is, we use the existing estimates to calculate the new membership probabilities:

$$p(k|l) = \frac{p_k g(D_l : \lambda_k)}{\sum_m^K p_k g(D_l : \lambda_k)} \tag{12}$$

As mentioned in Section 3, and explicit because of the normalization in Equation 12, the $p(k|l)$s sum to 1. Therefore, we can use Jensen's inequality to bound the likelihood:

$$\ell(\mathcal{D} : \theta) = \sum_l^N \log \sum_k^K q(k,l) \geq \sum_l^N \sum_k^K p(k|l) \log \frac{q(k,l)}{p(k|l)} = b(\theta) \tag{13}$$

by plugging in $p(k|l)$ for π_k and $q(k,l)$ for c_k.

We can then expand the logarithm, distribute $p(k|l)$, and split the sum to see that:

$$b(\theta) = \sum_{l}^{N}\sum_{k}^{K} p(k|l)\log q(k,l) - \sum_{l}^{N}\sum_{k}^{K} p(k|l)\log p(k|l). \tag{14}$$

The $p(k|l)$ values are fixed, so we can maximize the bound by focusing only on the first term of the summation. We call this function Q:

$$Q(\theta) = \sum_{l}^{N}\sum_{k}^{K} p(k|l)\log q(k,l) \tag{15}$$

We calculate the new parameters by differentiating Q with respect to each parameter, setting the derivative equal to 0, and solving for the parameters. We now derive these results in quite a bit of detail.

5.1 Updating the Poisson parameters

$$\frac{\partial Q}{\partial \lambda_k} = \frac{\partial}{\partial \lambda_k}\sum_{l}^{N}\sum_{m}^{K} p(m|l)\log q(m,l) \tag{16}$$

The derivative of a sum is the sum of derivatives, and $p(k|l)$ is a constant here. Also, the derivative of every term which does not involve k is 0, so we can ignore those.

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{l}^{N} p(k|l)\frac{\partial}{\partial \lambda_k}\log q(k,l) \tag{17}$$

Replace q with its definition, Equation 6.

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{l}^{N} p(k|l)\frac{\partial}{\partial \lambda_k}\log p_k g(D_l : \lambda_k) \tag{18}$$

Replace g with its definition (the Poisson density function).

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{l}^{N} p(k|l)\frac{\partial}{\partial \lambda_k}\log p_k \frac{\lambda_k^{D_l}}{D_l!}\exp\{-\lambda_k\} \tag{19}$$

Expand the logarithm.

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{l}^{N} p(k|l)\frac{\partial}{\partial \lambda_k}\log p_k + \log \lambda_k^{D_l} - \log D_l! + \log\exp\{-\lambda_k\} \tag{20}$$

Simplify the logarithm of exponential, and the exponentiation in the logarithm.

$$\frac{\partial Q}{\partial \lambda_k} = \sum_{l}^{N} p(k|l)\frac{\partial}{\partial \lambda_k}\log p_k + D_l \cdot \log \lambda_k - \log D_l! - \lambda_k \tag{21}$$

Evaluate the derivative.

$$\frac{\partial Q}{\partial \lambda_k} = \sum_l^N p(k|l)(\frac{D_l}{\lambda_k} - 1) \tag{22}$$

Set the derivative to 0. We now simply have some arithmetic to do.

$$0 = \sum_l^N p(k|l)(\frac{D_l}{\lambda_k} - 1) \tag{23}$$

Distribute $p(k|l)$ and expand the sum.

$$0 = \sum_n^N p(k|l)(\frac{D_l}{\lambda_k}) - \sum_l^N p(k|l) \tag{24}$$

For convenience, we will define $Z(k) \coloneqq \sum_l^N p(k|l)$. Add it to both sides.

$$Z(k) = \sum_l^N p(k|l)(\frac{D_l}{\lambda_k}) \tag{25}$$

λ_k is constant in the sum, so pull it out.

$$Z(k) = \frac{1}{\lambda_k} \sum_l^N p(k|l) D_l \tag{26}$$

Multipy both sides by λ_k, and divide both sides by $Z(k)$.

$$\lambda_k = \frac{\sum_l^N p(k|l) D_l}{Z(k)} \tag{27}$$

Thus, we have our new estimate of λ_k.

5.2 Updating the mixing probabilities

There derivation for updating the p_ks is similar to that of λ_k, but one key difference is the constraint that the p_ks are a probability distribution. To address this constraint, we use a Lagrange multiplier δ to constrain the p_k values.

$$\frac{\partial Q}{\partial p_k} = \frac{\partial}{\partial p_k} \sum_l^N \sum_m^K p(m|l) \log q(m,l) + \delta(\sum_k p_k - 1) \tag{28}$$

The derivative of sums is the sum of derivatives. The derivative of every term which does not involve k is 0.

$$\frac{\partial Q}{\partial p_k} = \sum_l^N \frac{\partial}{\partial p_k} p(k|l) \log q(k,l) + \frac{\partial}{\partial p_k} \delta(p_k - 1) \tag{29}$$

Evaluate the derivative.

$$\frac{\partial Q}{\partial p_k} = \sum_l^N \frac{p(k|l)}{p_k} + \delta \tag{30}$$

Set the derivative equal to 0 and solve for p_k.

$$0 = \sum_{l}^{N} \frac{p(k|l)}{p_k} + \delta \tag{31}$$

First, we need to find the value of δ. By summing over k, we see that δ is $-N$.

$$N = \sum_{l}^{N} \frac{p(k|l)}{p_k} \tag{32}$$

Since $\frac{1}{p_k}$ is constant, we can pull it out of the sum and move it to the other size. Then, we divide both sizes by N.

$$p_k = \frac{\sum_{l}^{N} p(k|l)}{N} \tag{33}$$

Recall that we defined $Z(k) := \sum_{l}^{N} p(k|l)$, so rewrite the equation using $Z(k)$.

$$p_k = \frac{Z(k)}{N} \tag{34}$$

Thus, we have our new estimate of p_k.

References

[1] J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report TR-97-021, International Computer Science Institute, 1998.

[2] C. Tomasi. Estimating Gaussian mixture densities with EM - a tutorial.