EM Correctness

Eric Walton

1 Background

1.1 Gibb's Inequality

If P and Q are discrete probability distributions, each with n items, then:

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

With equality only if

$$p_i = q_i$$

Proof

First off, for all x:

$$\log_2 x \leq x - 1$$

Therefore, since p_i is nonnegative:

$$\sum p_i \log(q_i/p_i) \leq \sum p_i(q_i/p_i - 1)$$

Notice that:

$$\sum p_i(q_i/p_i - 1) = \sum q_i - p_i = 1 - 1 = 0$$

So:

$$\sum p_i \log(q_i/p_i) \leq 0$$

$$\sum p_i(\log q_i - \log p_i) \leq 0$$

$$\sum p_i \log q_i \leq \sum p_i \log p_i$$

Multiply by negative 1:

$$-\sum p_i \log q_i \geq -\sum p_i \log p_i$$

2 EM Algorithm

2.1 Problem statement

Given a set X of observed data. Our goal is to calculate both θ, the parameters of the model, and Z, a set of hidden variables which affect probabilities within the model.

2.2 Optimization

Our goal is to choose θ and Z such that our observed data X exhibits maximum likelihood. In other words, we want to maximize:

$$P(X|\theta) = \sum_Z P(X, Z|\theta) = \sum_Z P(Z|X, \theta)P(X|\theta)$$

This quantity is often intractable...there may be infinitely many possible solutions for Z. Instead, we hold the distribution of Z constant for a moment. Let $\theta^{(t)}$ be the current estimate of θ. We define:

$$Q(\theta|\theta^{(t)}) = E_{Z|X,\theta^{(t)}}[\log P(X, Z|\theta)]$$

which is the expected value of the logarithm in the brackets over the conditional distribution of Z, where Z is conditional on the choice of $\theta^{(t)}$ and the observed X.

In the *expectation step* we calculate a closed form for $Q(\theta|\theta^{(t)})$ as a function of θ.

To *maximize*, find a new value for θ that maximizes $Q(\theta|\theta^{(t)})$.

Repeat until $Q(\theta|\theta^{(t)})$ converges to within a desired accuracy.

3 Proof of correctness

Let $\theta^{(t)}$ be the value of θ at time t. The algorithm above chooses a new value for θ such that we increase the value of $Q(\theta|\theta^{(t)})$. This proof will show that increasing $Q(\theta|\theta^{(t)})$ also implies an increase in $\log P(X|\theta)$.

Note that for random variables X and Z, conditioned on a third variable θ:

$$P(X, Z|\theta) = P(X|\theta) * P(Z|X, \theta)$$

And as long as our choice of Z has nonzero probability:

$$P(X|\theta) = \frac{P(X, Z|\theta)}{P(Z|X, \theta)}$$

Log both sides:

$$\log P(X|\theta) = \log P(X, Z|\theta) - \log P(Z|X, \theta)$$

Take the expectation of both sides over all values of Z, where the distribution of Z is conditioned on X and a different value of θ, which we denote $\theta^{(t)}$.

$$\mathbf{E}_{Z|X,\theta^{(t)}} \log P(X|\theta) = \mathbf{E}_{Z|X,\theta^{(t)}} \log P(X,Z|\theta) - \mathbf{E}_{Z|X,\theta^{(t)}} \log P(Z|X,\theta)$$

.

The left side does not depend on Z, so it's the expectation of a constant. For the right side, compute expectation as a weighted sum over values of Z:

$$\log P(X|\theta) = \sum_{z \in range(Z)} P(z|X,\theta^{(t)}) \log P(X,z|\theta) - \sum_{z \in range(Z)} P(z|X,\theta^{(t)}) \log P(z|X,\theta)$$

Note that $Q(\theta|\theta^{(t)})$ is exactly the first term. Let $H(\theta|\theta^{(t)})$ be the (negated) second term.

$$= Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)})$$

The above equation is true for all values of θ, including $\theta^{(t)}$. So:

$$\log P(X|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)})$$

Subtract the above two equations:

$$\log P(X|\theta) - \log P(X|\theta^{(t)}) = Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})$$

Because of Gibbs' inequality rule, we know that:

$$H(\theta|\theta^{(t)}) \geq H(\theta^{(t)}|\theta^{(t)})$$

So $H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \geq 0$ and we can write:

$$\log P(X|\theta) - \log P(X|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$$

Put in words, any change in θ that improves Q will improve $\log P(X|\theta)$ by at least as much. This completes the proof.

4 Guassian mixture model

This is one application of the EM algorithm.

4.1 Input

We are given observed input data, including n observations each containing d features. We represent each observation as a vector $x_i \in \mathbf{R}^d$, for $i = \{1, ..., n\}$. In addition, we are given K, the number of clusters.

4.2 Model

The goal is to assign each input vector to a cluster.

Each cluster is assumed to be a normal distribution, such that the likelihood of a given value of x_i, assumed to be in cluster k, is:

$$\Phi(x_i; \mu_k, \Sigma_k) \sim \mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left[-\tfrac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k)\right]$$

where μ_k is the mean and Σ_k the covariance matrix of cluster k.

To model which cluster each x_i belongs to, we need three more variables:

- $z_i \in \{1..K\}$ such that x_i belongs to cluster z_i.
- $\gamma_{i,k}^{(t)} = P(z_i = k | x_i, \theta^{(t)})$ as estimated at iteration t.
- $w_k^{(t)} = P(z_i = k | \theta^{(t)})$ as estimated at iteration t.

The following letters denote collections of variables:

- $\theta^{(t)} = \{w_k^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)}\}$ for $k = \{1..K\}$, as estimated at time t.
- $Z = \{z_i\}$ for $i = \{1..n\}$.

4.3 Initialization

Choose initial values for θ. Many choices work, but we set, for all $k \in \{1..K\}$:

- $\mu_k^{(0)} :=$ a unique random x_i
- $\Sigma_k^{(0)} := \frac{1}{n} X X^T$
- $w_k^{(0)} := \frac{1}{K}$

We assume that, at each step, the cluster assignments z_i are independent of all values in X except x_i. And that the likelihood of each x_i is independent of all cluster assignments except z_i.

5 E step

Distribution of z values (soft cluster mappings)

$$\gamma_{i,k}^{(t)} := P(z_i = k | x_i, \theta^{(t)})$$

Using Bayes' theorem:

$$= \frac{P(z_i = k | \theta^{(t)}) P(x_i | z_i = k, \theta^{(t)})}{P(x_i | \theta^t)}$$

$$= \frac{w_k^{(t)} \Phi(x_i; \mu_j^{(t)}, \Sigma_j^{(t)})}{\sum_{l=1}^{K} w_l^{(t)} \Phi(x_i; \mu_l^{(t)}, \Sigma_l^{(t)})}$$

The log likelihood function:

$$\lambda(\theta; X) = \ln P(X, Z|\theta)$$

By Bayes' rule:

$$= \ln P(X|Z, \theta) P(Z|\theta) = \ln P(X|Z, \theta) + \ln P(Z|\theta)$$

$$= \ln \left(\prod_{i=1}^{n} P(x_i|Z, \theta) \right) + \ln \left(\prod_{i=1}^{n} P(z_i|\theta) \right)$$

$$= \sum_{i=1}^{n} \ln P(x_i|Z, \theta) + \ln P(z_i|\theta)$$

Since x_i is independent of everything in Z except z_i:

$$= \sum_{i=1}^{n} \ln P(x_i|z_i, \theta) + \ln P(z_i|\theta)$$

Let $Q(\theta|\theta^{(t)})$ be the expected value of λ, over the distribution of Z conditional on the observed values x_i, and the previous value of θ, which we denote $\theta^{(t)}$.

$$Q(\theta|\theta^{(t)}) = \mathbf{E}_{Z|X,\theta^{(t)}} \lambda(\theta; X)$$

$$= \mathbf{E}_{Z|X,\theta^{(t)}} \left[\sum_{i=1}^{n} \ln P(x_i|z_i, \theta) + \ln P(z_i|\theta) \right]$$

Linearity of expectation:

$$= \sum_{i=1}^{n} \mathbf{E}_{z_i|X,\theta^{(t)}} \left[\ln P(x_i|z_i, \theta) + \ln P(z_i|\theta) \right]$$

To compute the expected value, we partition the sample space into all K possible values of z_i. Note that these events are disjoint, and their union fills the whole space. The expectation is the probability of a given assignment $z_i = k$ times the value of the quantity, given that $z_i = k$, summed over all possible values k.

In other words: consider the quantity in the square brackets a function of z_i. The expected value of a function is simply the sum over all possible inputs of the probability of that input times the value of the function with that input:

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k|x_i, \theta^{(t)}) \left[\ln P(x_i|z_i = k, \theta) + \ln P(z_i = k|\theta)\right]$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k}^{(t)} \left[\ln \Phi(x_i; \mu_k, \Sigma_k) + \ln w_k\right]$$

6 M step

6.1 choice of w_k

$$w_k^{(t+1)} = \arg\max_{w_k} Q(\theta|\theta^{(t)}) = \arg\max_{w_k} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k}^{(t)} \ln w_k$$

$$= \arg\max_{w_k} \sum_{k=1}^{K} \ln w_k \sum_{i=1}^{n} \gamma_{i,k}^{(t)}$$

Subject to the constraint that:

$$\sum_{k=1}^{K} w_k = 1$$

Use Lagrange multipliers. let

$$f(w_1, ..., w_K) = \sum_{k=1}^{K} \ln w_k \sum_{i=1}^{n} \gamma_{i,k}^{(t)}$$

Let

$$g(w_1, ..., w_K) = \sum_{k=1}^{K} w_k$$

Any critical points, including local and global maxima, will occur where $f()$ and $g()$ are tangential. Wherever they are tangential, they will have parallel gradient vectors. At those points we have:

$$\nabla g = \lambda \nabla f$$

for some scalar λ. The gradients are:

$$f_{w_k} = \frac{1}{w_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \qquad g_{w_k} = 1$$

This gives us a system of $K + 1$ equations:

$$\frac{1}{w_k}\sum_{i=1}^{n}\gamma_{i,k}^{(t)} = \lambda \text{ for } k = \{1..K\} \quad \text{and the constraint equation} \quad \sum_{k=1}^{K} w_k = 1$$

For a given w_k we have:

$$w_k = \frac{1}{\lambda}\sum_{i=1}^{n}\gamma_{i,k}^{(t)}$$

Plug into second:

$$1 = \sum_{k=1}^{K}\frac{1}{\lambda}\sum_{i=1}^{n}\gamma_{i,k}^{(t)}$$

$$\lambda = \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{i,k}^{(t)}$$

Over all k, the values $\gamma_{i,k}^{(t)}$ are a probability distribution, so their sum is 1.

$$\lambda = \sum_{i=1}^{n} 1 = n$$

Plug into the first:

$$w_k = \frac{1}{n}\sum_{i=1}^{n}\gamma_{i,k}^{(t)}$$

6.2 Choice of μ

$$\mu_k^{(t+1)} = \arg\max_{\mu_k} Q(\theta|\theta^{(t)}) = \arg\max_{\mu_k} \sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{i,k}^{(t)}[-\tfrac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)]$$

Take the vector derivative with respect to a particular μ_k. Use the fact that $\frac{\delta}{\delta v}v^TAv = 2Av$ and use the chain rule..

$$\frac{\delta}{\delta\mu_k} = \sum_{i=1}^{n}\gamma_{i,k}^{(t)}\Sigma^{-1}(x_i-\mu_k)$$

Set this to zero.

$$0 = \Sigma^{-1}\sum_{i=1}^{n}\gamma_{i,k}^{(t)}(x_i-\mu_k)$$

Σ^{-1} is obviously invertible, so we need the vector to be zero.

$$0 = \sum_{i=1}^{n} \gamma_{i,k}^{(t)}(x_i - \mu_k)$$

$$\sum_{i=1}^{n} \gamma_{i,k}^{(t)}\mu_k = \sum_{i=1}^{n} \gamma_{i,k}^{(t)} x_i$$

$$\mu_k^{(t+1)} = \frac{\sum\limits_{i=1}^{n} \gamma_{i,k}^{(t)} x_i}{\sum\limits_{i=1}^{n} \gamma_{i,k}^{(t)}} = \frac{\sum\limits_{i=1}^{n} \gamma_{i,k}^{(t)} x_i}{n w_k^{(t)}}$$

6.3 Choice of Σ_k

$$\Sigma_k^{(t+1)} = \arg\max_{\Sigma_k} Q(\theta|\theta^{(t)}) = \arg\max_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \left[\ln \Phi(x_i; \mu_k, \Sigma_k) + \ln w_k\right]$$

$$\arg\max_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \left[-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k) + \ln w_k\right]$$

$$\arg\min_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|\Sigma_k| + \sum_{i=1}^{n} \gamma_{i,k}^{(t)}(x_i - \mu_k)^T \Sigma_k^{-1}(x_i - \mu_k)$$

Consider the quadratic form to be the trace of a 1×1 matrix. Use the cyclic property of trace:

$$\arg\min_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|\Sigma_k| + \sum_{i=1}^{n} \gamma_{i,k}^{(t)} tr\left((x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}\right)$$

$$\arg\min_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|\Sigma_k| + \sum_{i=1}^{n} tr\left(\gamma_{i,k}^{(t)}(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}\right)$$

$$\arg\min_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|\Sigma_k| + tr\left(\sum_{i=1}^{n} \gamma_{i,k}^{(t)}(x_i - \mu_k)(x_i - \mu_k)^T \Sigma_k^{-1}\right)$$

Let $S = \sum\limits_{i=1}^{n} \gamma_{i,k}^{(t)}(x_i - \mu_k)(x_i - \mu_k)^T$

$$\arg\min_{\Sigma_k} \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|\Sigma_k| + tr\left(S\Sigma_k^{-1}\right)$$

Let $B = S\Sigma_k^{-1}$. B is $d \times d$. Aside:

$$|B| = |S|\frac{1}{|\Sigma_k|}$$

$$\ln|B| = \ln|S| - \ln|\Sigma_k|$$

$$\ln|\Sigma_k| = \ln|S| - \ln|B|$$

Back to minimizing:

$$\arg\min_B \sum_{i=1}^{n} \gamma_{i,k}^{(t)} \left(\ln|S| - \ln|B|\right) + tr\left(B\right)$$

$$\arg\min_B \left(-\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln|B| + tr\left(B\right) \right)$$

Both determinant and trace can be written in terms of eigenvalues. Let λ_j be the jth eigenvalue of B.

$$\arg\min_B \left(-\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \ln \prod_{j=1}^{d} \lambda_j + \sum_{j=1}^{d} \lambda_j \right)$$

$$\arg\min_B \left(-\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \sum_{j=1}^{d} \ln \lambda_j + \sum_{j=1}^{d} \lambda_j \right)$$

Take the derivative with respect to one particular λ_j, and set it to zero.

$$0 = -\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \frac{1}{\lambda_j} + 1$$

$$\sum_{i=1}^{n} \gamma_{i,k}^{(t)} = \lambda_j$$

Since every eigenvalue is the same, we choose:

$$B = \left(\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \right) I$$

$$S\Sigma_k^{-1} = \left(\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \right) I$$

$$\Sigma_k^{-1} = S^{-1} \left(\sum_{i=1}^{n} \gamma_{i,k}^{(t)} \right)$$

$$\Sigma_k = \frac{S}{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}} = \frac{\sum_{i=1}^{n} \gamma_{i,k}^{(t)} (x_u - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{n} \gamma_{i,k}^{(t)}}$$

7 Whitening

Suppose we have a random vector $\vec{z} = [z_1 ... z_n]^T$, whose elements are jointly Gaussian, i.e. $\vec{y} \sim N\{\mu_y, \Sigma_y\}$. What is the joint distribution of a related random vector, $\vec{z} = \Sigma_y^{-1/2}(\vec{y} - \mu_y)$. This common preprocessing step in machine-learning algorithms is called "whitening" - can you explain why this might be a reasonable name?

7.1 Solution

By definition of mean, the mean of z is its expected value. Here is shown that the mean of $\vec{z}$ is 0:

$$\mu_z = \mathbf{E}[\vec{z}] = \mathbf{E}[\Sigma_y^{-1/2}(\vec{y} - \mu_y)] = \Sigma_y^{-1/2}\left(\mathbf{E}[\vec{y}] - \mu_y\right) = \Sigma_y^{-1/2}(\mu_y - \mu_y) = 0$$

The definition of covariance matrix is the expected value of the outer product of normalized vectors:

$$cov(\vec{z}) = \mathbf{E}\left[(\vec{z} - \mu_z)(\vec{z} - \mu_z)^T\right] = \mathbf{E}[\vec{z}\vec{z}^T]$$

Substitute, remembering that $\Sigma_y^{-1/2}$ is symmetric:

$$= \mathbf{E}\left[\left(\Sigma_y^{-1/2}(\vec{y} - \mu_y)\right)\left(\Sigma_y^{-1/2}(\vec{y} - \mu_y)\right)^T\right] = \mathbf{E}\left[\Sigma_y^{-1/2}(\vec{y} - \mu_y)(\vec{y} - \mu_y)^T\Sigma_y^{-1/2}\right]$$

$\Sigma_y^{-1/2}$ is constant:

$$= \Sigma_y^{-1/2}\mathbf{E}\left[(\vec{y} - \mu_y)(\vec{y} - \mu_y)^T\right]\Sigma_y^{-1/2}$$

The expectation in the middle is exactly the definition of Σ_y:

$$= \Sigma_y^{-1/2}\Sigma_y\Sigma_y^{-1/2} = I$$

The covariance matrix is the identity. In other words, there is no correlation between variables, and the variance of each variable is 1.

This called "whitening" because variables are consider "white" if they are mutually independent.

8 Acknowledgements

This was originally an assignment from Gopal Nataraj, and portions of this were adapted from his lectures. I freely consulted, and used the wikipedia articles on the topic:

- `http://en.wikipedia.org/wiki/Expectation_maximization`

- http://en.wikipedia.org/wiki/Estimation_of_covariance_matrices
- http://en.wikipedia.org/wiki/Gibbs%27_inequality

For LaGrange multipliers, I consulted the MIT opencourseware lecture on the topic by Professor Denis Auroux:

http://ocw.mit.edu/courses/mathematics/18-02-multivariable-calculus-fall-2007/video-lec