CSEP 590A, Summer 2006, Lecture 4

Based on earlier notes by C. Grant & M. Narasimhan

Introduction

Last lecture we began an examination of model based clustering. This lecture will be the technical background leading to the Expectation Maximization (EM) algorithm.

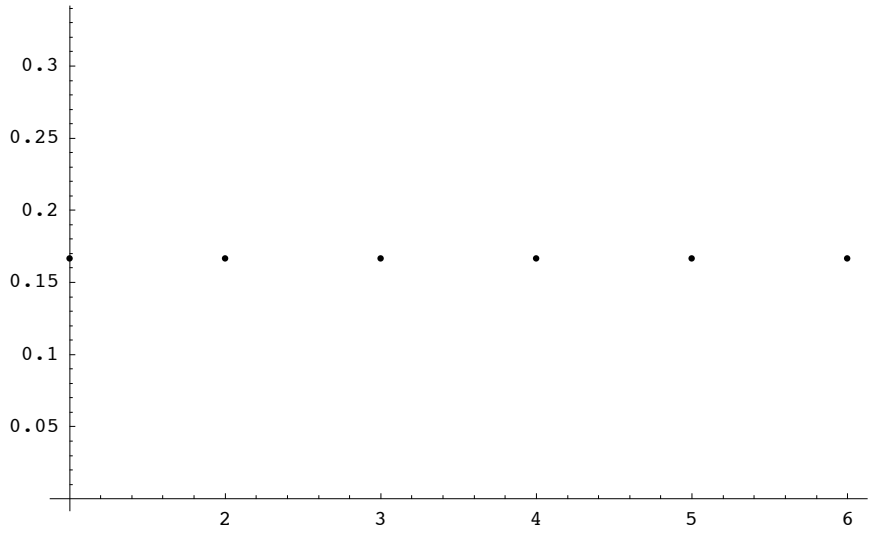
Do gene expression data fit a Gaussian model? The central limit theorem implies that the sum of a large number of independent identically distributed random variables can be well approximated by a Normal distribution. While it is far from clear that the expression data is a sum of independent variables, using the Normal distribution seems to work in practice. Besides, having a weak model is better than having no model at all.

Probability Basics

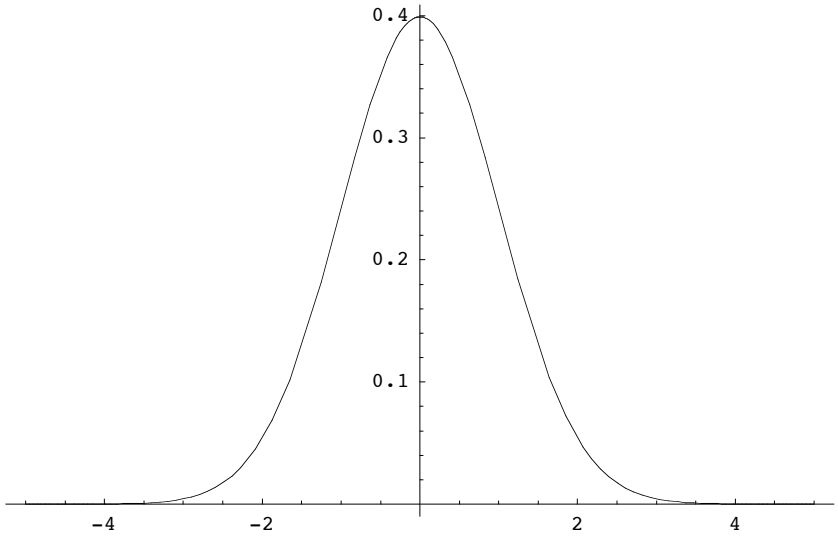
A random variable can be continuous or discrete (or both). A discrete random random variable corresponds to a probability distribution on a discrete sample space, such as the roll of a dice. A continuous random variable corresponds to a probability distribution on a continuous sample space such as $\mathbb{R}$. Shown in the table below are two examples of probability distributions, with the first representing a roll of an unbiased die, and the second representing a Normal distribution.

	Discrete	Continuous
Sample Space	$\{1, 2, \ldots 6\}$	$\mathbb{R}$
Distribution	$p_1, p_2, \ldots p_6 \geq 0$, $\sum_{i=1}^{6} p_i = 1$ $p_1 = p_2 = \ldots = p_6 = \frac{1}{6}$	$f(x) \geq 0$, $\int_{\mathbb{R}} f(x)\, dx = 1$ $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$

Discrete Probability Distribution



Continuous Probability Distribution



Parameter Estimation

Many distributions are parametrized. Typically, we have data $x_1, x_2, ..., x_n$ that is sampled from a parametric distribution $f(x \mid \theta)$. Often, the goal is to estimate the parameter θ. The mean μ and variance σ^2 are often used as such parameters. Estimates of these quantities derived from the sampled data are often called the sample statistics, while the (true) parameter based on the entire sample space is called the population statistic. The following table illustrates these two concepts.

	Discrete	Continuous
Population Mean	$\mu = \sum_i i\, p_i$	$\mu = \int x\, f(x)\, dx$
Population Variance	$\sigma^2 = \sum_i (i - \mu)^2\, p_i$	$\sigma^2 = \int (x - \mu)^2\, f(x)\, dx$
Sample Mean	$\bar{x} = \sum_{i=1}^n x_i / n$	$\bar{x} = \sum_{i=1}^n x_i / n$
Sample Variance	$\bar{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$	$\bar{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$

While the sample statistics can be used as estimates of these parameters, this is often not the prefered way of estimating these quantities. For example, the sample variance $\overline{\sigma}^2 = \sum_{i=1}^n (x_i - \overline{x})^2 / n$ is a *biased* estimate of the true variance because it underestimates the quantity (an unbiased estimate of the variance is given by $\overline{\sigma}^2 = \sum_{i=1}^n (x_i - \overline{x})^2 / (n - 1)$). Maximum Likelihood Estimation is one of many parameter estimation techniques (note that the MLE is not guaranteed to be unbiased either).

Assuming the data are independent, the likelihood of the data $x_1, x_2, \ldots, x_n$ given the parameter θ is

$$L(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$$

where f is the probability density function of the presumed distribution (which of course dcepends on θ). Note that the x_i are known constants, not variables; they are the values we observed. On the other hand, θ is unknown. We treat the likelihood L as a function of θ and ask what value of θ maximizes it. The typical approach is to solve for

$$\frac{\partial}{\partial \theta} L(x_1, x_2, \ldots, x_n \mid \theta) = 0$$

Since the likelihood function is always positive (and we may assume it to be strictly positive), the log likelihood

$$\ln L(x_1, x_2, \ldots, x_n \mid \theta) = \ln \prod_{i=1}^n f(x_i \mid \theta) = \sum_{i=1} \ln f(x_i \mid \theta)$$

is well defined, and by the monotonicity of the logarithm, the log likelihood is maximized exactly when the likelihood is maximized. Hence we can solve for

$$\frac{\partial}{\partial \theta} \ln L(x_1, x_2, \ldots, x_n \mid \theta) = 0$$

Note that in general, these conditions are statisfied by maxima, minima and stationary points of the log-likelihood function. (A "stationary point" is a temporary flat spot on a curve that otherwise tends upward or downward.) Further, if θ is restricted to be in some bounded range, then maxima might occur at the boundary which does not satisfy this condition. Therefore, we need to check the boundaries separately. Here is an example which illustrates this procedure.

Example 1. Let $x_1, x_2, \ldots, x_n$ be coin flips, and let θ be the probability of getting heads. Suppose we observe n_0 tails and n_1 heads ($n_0 + n_1 = n$). Then the likelihood function is given by

$$L(x_1, x_2, \ldots, x_n \mid \theta) = (1-\theta)^{n_0}\, \theta^{n_1}$$

Hence the log – likelihood function is

$$\ln L(x_1, x_2, \ldots, x_n \mid \theta) = n_0 \ln(1-\theta) + n_1 \ln \theta$$

To find a value of θ that maximizes this function, we solve for

$$\frac{\partial}{\partial \theta} \ln L(x_1, x_2, \ldots, x_n \mid \theta) = \frac{-n_0}{1-\theta} + \frac{n_1}{\theta} = 0$$

This yields

$$\frac{-n_0}{1-\theta} + \frac{n_1}{\theta} = 0$$

$$n_1(1-\theta) = n_0\,\theta$$

$$n_1 = (n_0 + n_1)\,\theta$$

$$\frac{n_1}{(n_0+n_1)} = \theta$$

$$\frac{n_1}{n} = \theta$$

(The sign of 2nd derivative can then be checked to guarantee that this is a maximum not a minimum. Likewise, you can easily verify that the maximum is not attained at the boundaries of the parameter space, i.e. at θ=0 or θ=1.) This estimate for the parameter of the distribution matches our intuition.

Example 2. Suppose $x_i \sim N(\mu, \sigma)$, $\sigma^2 = 1$ and μ unknown. Then

$$L(x_1, x_2, \ ..., \ x_n \mid \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2}$$

$$\ln L(x_1, x_2, \ ..., \ x_n \mid \theta) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi \ - \frac{(x_i-\theta)^2}{2} \right)$$

$$\frac{\partial}{\partial \theta} \ln L(x_1, x_2, \ ..., \ x_n \mid \theta) = \sum_{i=1}^{n} (x_i - \theta) = \sum_{i=1}^{n} x_i - n\,\theta = 0$$

So the value of θ that maximizes the likelihood is

$$\theta = \sum_{i=1}^{n} x_i / n$$

Again matching our intuition: the sample mean is the maximum likelihood estimator (MLE) for the population
mean.

Example 3. Suppose $x_i \sim N(\mu, \sigma)$, σ^2 and μ unknown. Then

$$L(x_1, x_2, \ ..., \ x_n \mid \theta_1, \ \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_i-\theta_1)^2/2\,\theta_2}$$

$$\ln L(x_1, x_2, \ ..., \ x_n \mid \theta_1, \ \theta_2) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln 2\pi\,\theta_2 \ - \frac{(x_i-\theta_1)^2}{2\,\theta_2} \right)$$

$$\frac{\partial}{\partial \theta_1} \ln L(x_1, x_2, \ ..., \ x_n \mid \theta_1, \theta_2) = \sum_{i=1}^{n} \frac{(x_i-\theta_1)}{\theta_2} = 0 \Longrightarrow \sum_{i=1}^{n} x_i / n = \ \theta_1$$

$$\frac{\partial}{\partial \theta_2} \ln L(x_1, x_2, \ ..., \ x_n \mid \theta_1, \theta_2) =$$

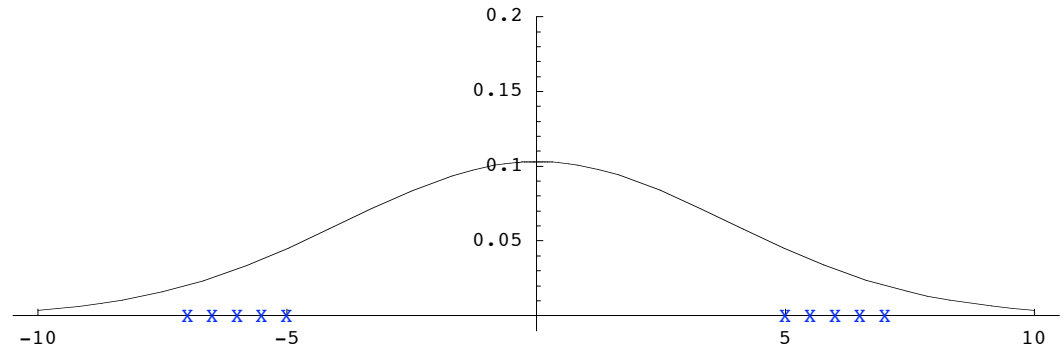
$$\sum_{i=1}^{n} \left(-\frac{1}{2} \ \frac{2\pi}{2\pi\theta_2} + \frac{(x_i-\theta_1)^2}{2\,{\theta_2}^2} \right) = \sum_{i=1}^{n} \left(-\frac{1}{2\ \theta_2} + \frac{(x_i-\theta_1)^2}{2\,{\theta_2}^2} \right) = 0 \Longrightarrow \sum_{i=1}^{n} (x_i - \theta_1)^2 / n = \theta_2$$

The MLE for the population variance is the sample variance. This is a *biased* estimator. It systematically underestimates the population variance, but is none the less the MLE. The MLE doesn't promise an unbiased estimator but it is a reasonable approach.

Expectation Maximization

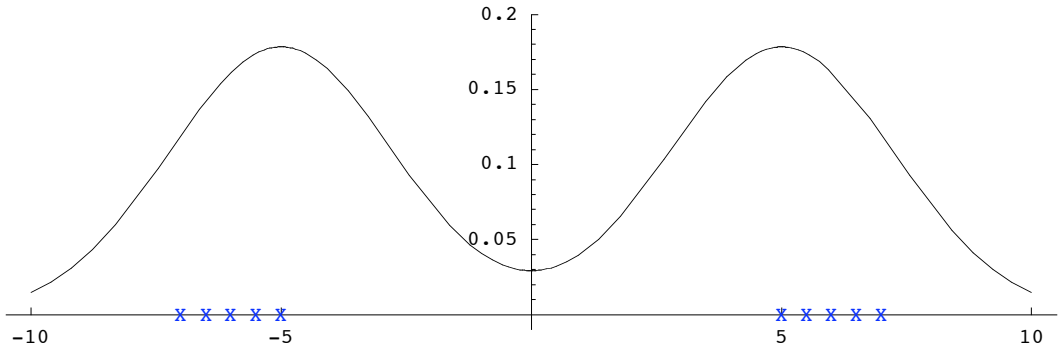
The MLE approach works well when we have relatively simple parametrized distributions. However, when we have more complicated situations, we may not be able to solve for the ML estimate because the complexity of the likelihood function precludes both analytical and numerical optimization. The EM algorithm can be thought of as an algorithm that provides a tractable approximation to the ML estimate.

Consider the following example. We have data corresponding to heights of individuals, as shown in the figures below. Is this distribution likely to be Normally distributed as shown below?



- Graphics -

Or is there some hidden variable, like gender, so the distribution should be more like this:



- Graphics -

The clustering problem can is essentially a parameter estimation problem : Try to find if there are hidden parameters that cause the data to fall into two distributions $f_1(x)$, $f_2(x)$. These distributions depend on some parameter θ: $f_1(x, \theta)$, $f_2(x, \theta)$, and there are also mixing parameters τ_1 and τ_2, $\tau_1 + \tau_2 = 1$, which describe the probability of sampling from each group. Can we estimate the parameters for the this more complex model? Let's suppose that the two groups are normal but with different, unknown, parameters.

The likelihood is now given by

$$L(x_1, x_2, \ \ldots, \ x_n \mid \tau_1, \tau_2, \ \mu_1, \ \mu_2, \ \sigma_1, \ \sigma_2) = \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j \, f_j(x_i \mid \theta_j)$$

If we try to work with this in our existing framework it becomes messy and algebraically intractable, due to the product-of-sums form, and remains so even if we take the log of the likelihood.

This leads us to introduce the Expectation Maximization (EM) algorithm as a heuristic for finding the MLE. It is particularly useful for problems containing a hidden variable. It uses a hill-climbing strategy to find a local maximum of the likelihood.

Introduce new variables

$$z_{ij} = \begin{cases} 1 & \text{iff } x_i \text{ was sampled from distribution } j \\ 0 & \text{otherwise} \end{cases}$$

These variables are introduced for mathematical convenience. They let us avoid a sum over j in the expression for the likelihood. The full data table becomes

$x_1 \quad z_{11} \quad z_{12}$
$x_2 \quad z_{21} \quad z_{22}$
$x_n \quad z_{n1} \quad z_{n2}$

If the z were known estimating τ_1, τ_2 would be easy, and estimation of the parameters would become easy again. If we knew the parameters estimation of the z would be easy. The EM algorithm iterates over these alternatives. It can be proved that the likelihood will be monotonically increasing, and so will converge to a (local) maximum. [There is a polynomial time algorithm for estimating Gaussian mixtures under the assumption that the components are "well-separated," but the method is not used much in practice. I don't know whether the complexity of the general problem is known; plausibly it's NP-hard. So, the EM algorithm is probably the method of choice.]

Expectation step

Assume fixed values for τ_j and θ_j. Let A be the event that x_i is drawn from the distribution f_1, let B be the event that x_i is drawn from f_2, and let D be the event that x_i is observed. We want $P(A \mid D)$, but it is easier to find $P(D \mid A)$. We use Bayes' rule:

$$P(A \mid D) = \frac{P(D|A)\,P(A)}{P(D)}$$

$$P(D) =$$
$$P(D \mid A)\,P(A) + P(D \mid B)\,P(B) = \tau_1\,P(D \mid A) + \tau_2\,P(D \mid B) = \tau_1\,f_1(x_i \mid \theta_1) + \tau_2\,f_2(x_i \mid \theta_2)$$

$P(A \mid D)$ is the expected value of z_{i1} given θ_1 and θ_2. This is the expectation step of the EM algorithm.

To be concrete, consider a sample of points taken from a mixture of Gaussian distributions with unknown parameters and unknown mixing coefficients. The EM algorithm will give estimates of the parameters that raise the likelihood of the data.

An easy heuristic to apply is

If $E(z_{i1}) \geq 1/2$ then set $z_{i1} = 1$
If $E(z_{i1}) < 1/2$ then set $z_{i1} = 0$

This gives rise to the so-called Classification EM algorithm (we *classify* each observation as coming from exactly one of the component distributions). The k-means clustering algorithm is an example. In this case, the maximization step is just like the simple Maximum Likelihood Estimation examples considered above. The more general M-step (below) accounts for the inherent uncertainty in these classifications, appropriately weighting the contributions of each observation to the parameter estimates for each mixture component.

Maximization step

The expression for the likelihood is

$$L(x_1,\ z_{11}, z_{12},\ x_2,\ z_{21},\ z_{22},\ \ldots \mid \theta,\ \tau)$$

The x_i are known. If the z_{ij} were known finding the MLE of θ, τ would be easy, but we don't. Instead we maximize the *expected* log likelihood of the visible data $E(\ln L(x_1, x_2, \ldots, x_n \mid \theta, \tau))$. The expectation is taken over the distribution of the hidden variables z_{ij}. Assuming ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$, and $\tau_1 = \tau_2 = \tau = (\frac{1}{2})$:

$$L(\boldsymbol{x}, \boldsymbol{z} \mid \theta, \tau) \quad = \quad \prod_{i=1}^{n} \tau \frac{1}{\sqrt{2\pi\sigma^2}} e^{-1/2\,\sigma^2 (\sum_{j=1}^{2} z_{ij}(x_i - \mu_j)^2)}$$

so

$$E(\ln L(\boldsymbol{x}, \boldsymbol{z} \mid \theta, \tau)) = E(\textstyle\sum_{i=1}^{n} \ln \frac{1}{2} - \frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{2} z_{ij}(x_i - \mu_j)^2) =$$
$$\textstyle\sum_{i=1}^{n} \ln \frac{1}{2} - \frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{2} E(z_{ij})\,(x_i - \mu_j)^2)$$

The last step above depends on the important fact that expectation is linear: if c and d are constants and X and Y are random variables, then E(cX+dY) =c E(X) + d E(Y). We calculated $E(z_{ij})$ in the previous step. We can now solve for the μ_j that maximize the expectation by the methods given earlier: set derivatives to zero, etc. With a little more algebra you will see that the MLE for μ_j is the *weighted* average of the x_i's, where the weights are the $E(z_{ij})$'s,which makes sense intuitively: if a given point x_i has a high probability of having been sampled from distribution 1, then it will contribute strongly to our estimate of μ_1 and weakly to our estimate of μ_2.

It can be shown that this procedure increases the likelihood at every iteration, hence is guaranteed to converge to a local maximum. Unfortunately, it is not guaranteed to be the global maximum, but empirically it works well in many situations.