CSE 527 Notes for Oct. 29 2003 taken by Charles E. Grant

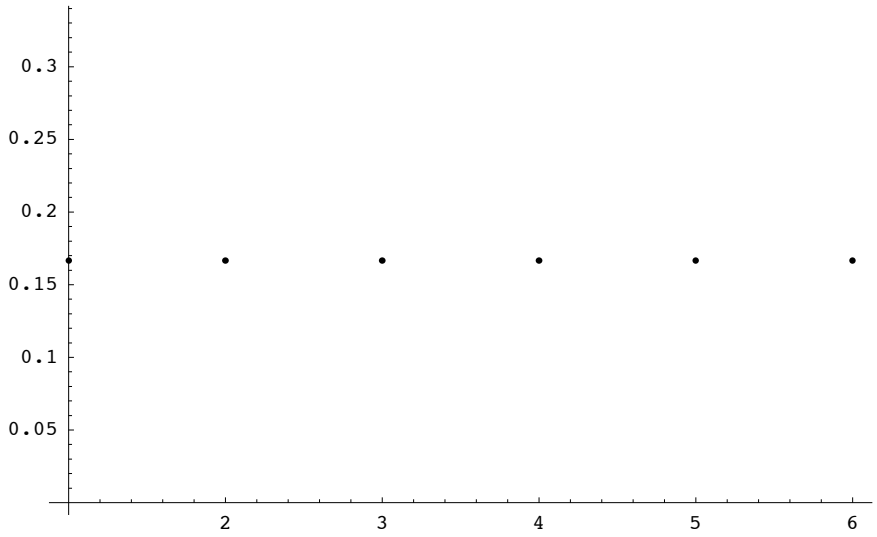
Last lecture we began an examination of model based clustering. This lecture will be the technical background leading to the Expectation Maximization (EM) algorithm.

Do gene expression data fit a Gaussian model? The central limit theorem implies that a variable which is the sum of lots of random variables will have a Normal distribution, but a cell is not random Nonetheless it seems to work and a weak model is better than no model.

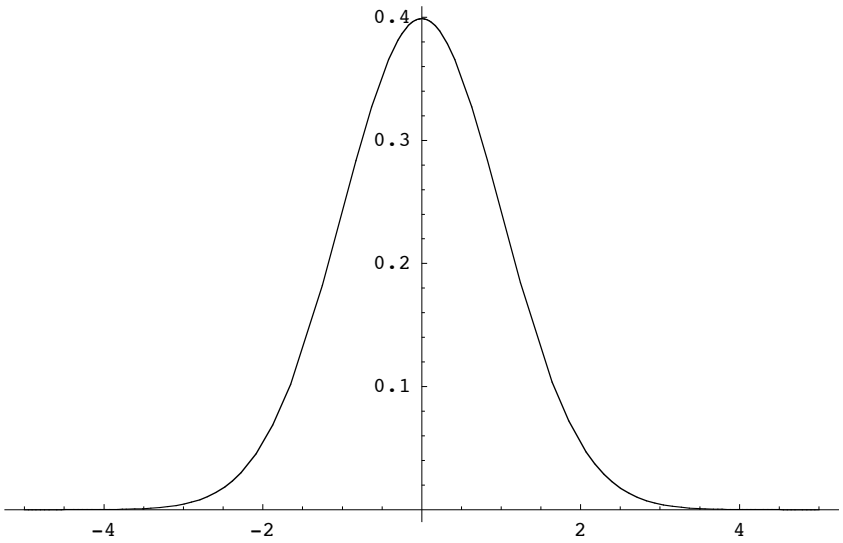
Probability Basics:

	Discrete Example	Continuous Example
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



Population vs Sample

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

Parameter Estimation

Assume that a data $x_1, x_2, \ldots, x_n$ are sampled from a parametric distribution $f(x \mid \theta)$. How do we estimate θ? For example the distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2} = f(x, \mu, \sigma)$$

is parameterized by μ, σ

The Maximum Likelihood Estimation is one of many parameter estimation techniques.

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

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

Treating the likelihood L as a function of θ, we ask what value of θ maximizes the likelihood. The typical approach is to solve

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

or

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

The properties of the logarithm make things easier to work with.

A likelihood is not a probability.

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 ($n_0 + n_1 = n$).

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

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

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

Setting this equal to 0 and solving we get

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

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

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

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

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

(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

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

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

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

So the value of θ that maximizes the likelihood is

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

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

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

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

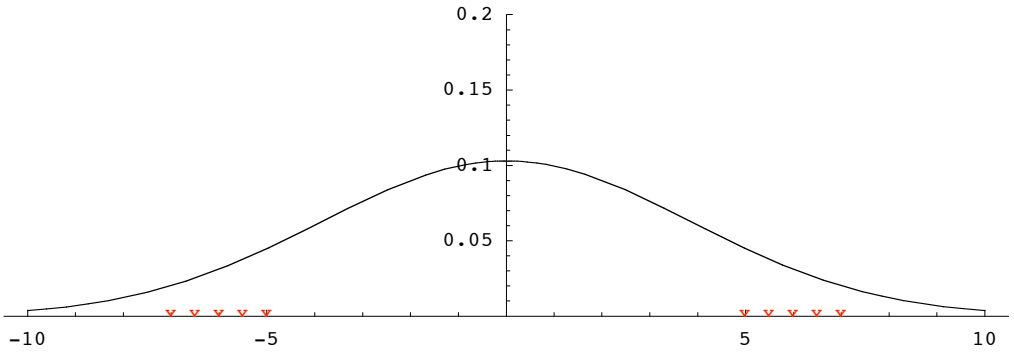
$$\frac{\partial}{\partial\theta_1}\ln L(x_1, x_2, \ ..., \ x_n \mid \theta_1, \theta_2) = \sum_{i=1}^{n} \frac{(x_i-\theta_1)}{\theta_2} = 0 \Longrightarrow \sum_{i=1}^{n} x_i / n = \ \theta_1$$

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

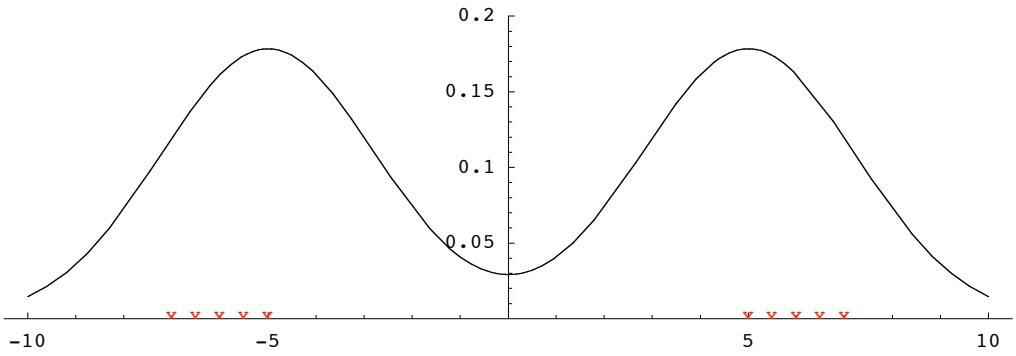
$$\sum_{i=1}^{n}\left(-\frac{1}{2}\ \frac{2\pi}{2\pi\theta_2} + \frac{(x_i-\theta_1)^2}{2{\theta_2}^2}\right) = \sum_{i=1}^{n}\left(-\frac{1}{2\ \theta_2} + \frac{(x_i-\theta_1)^2}{2{\theta_2}^2}\right) = 0 \Longrightarrow \sum_{i=1}^{n}(x_i - \theta_1)^2 / n = \theta_2$$

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.

Think of a more complex situation. Plot some data, say the height of some individuals. Is the distribution they come from this?



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



Clustering: 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 a 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

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

If we try to work with this in our existing framework it becomes messy and algebraically intractable, 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 a new variable

$$z_{\text{ij}} = \begin{cases} 0 \\ 1 \text{ iff } x_i \in \text{dist. } j \end{cases}$$

This variable is introduced for mathematical convenience. It lets us avoid a sum over j in the expression for the likelihood. The full data table becomes

x_1	z_{11}	z_{12}
x_2	z_{21}	z_{22}
x_n	z_{n1}	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 Ruzzo thinks the method

is not used much in practice. He doesn'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 maximzation 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 likelihood of the visible data $E(L(x_1, x_2, \ldots, x_n \mid \theta, \tau))$. The expectation is taken over the distribution of the hidden variables z_{ij}. Assuming ${\sigma_1}^2 = \sigma^2 = {\sigma_2}^2$

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

so

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

And we calculated $E(z_{\text{ij}})$ in the previous step. We can now solve for μ_j that maximizes the expectation.

We have yet to show that this converges.