HGEN 48600/STAT 35450: Fundamentals of Computational Biology

Lecture 4 — January 14

Lecturer: Matthew Stephens *Scribe: Justin Hsu*

Note: These lecture notes are still rough, and have only have been mildly proofread.

4.1 Likelihood Analysis

Consider the set of data of 100 tusks, 40 of which have the "1" allele, 60 with the "0" allele Then the data has the likelihood function

$$L(q) = P(Data|M_q) \tag{4.1}$$

. We can write this as

$$L(q) = q^{40}(1-q)^{60} \tag{4.2}$$

Now consider the log of the likelihood function:

$$\ell(q) = 40log(q) + 60log(1-q) \tag{4.3}$$

We can estimate q by finding the value of q that maximizes $L(q)$. This is known as the Maximum Likelihood estimator (mle), which we denote as $\hat{q}$. A useful feature is that the value that maximizes the likelihood function also maximizes the log likelihood function.

$$\begin{aligned} \hat{q} &= \operatorname*{argmax}_q L(q) \\ &= \operatorname*{argmax}_q \ell(q) \end{aligned} \tag{4.4}$$

This is useful because it is sometimes easier to find the maximum of $\ell(q)$.
Returning to the elephant tusk example, we find the maximum of the likelihood function by taking the derivative of the log likelihood.

$$\begin{aligned} \ell'(q) &= \frac{40}{q} - \frac{60}{1-q} \\ 0 &= \frac{40}{q} - \frac{60}{1-q} \\ \hat{q} &= \frac{40}{100} \end{aligned} \tag{4.5}$$

We can extend this generally so that given two populations n_1 and n_0, we have a likelihood function

$$L(q) = q^{n_1}(1-q)^{n_0} \tag{4.6}$$

and a log likelihood function

$$\ell(q) = n_1 log(q) - n_0 log(1-q) \tag{4.7}$$

Then the maximum likelihood estimate will have the form

$$\hat{q} = \frac{n_1}{n_1 + n_0} \tag{4.8}$$

This is the maximum likelihood of the Binomial Distribution.

4.2 Mixture Models

We now move on to mixture models, which our models that consist of a mixture of two or more distributions. As an example, consider the heights of all humans of these worlds. What would be the distribution of these heights. We could assume that they are normally distributed, but what if the male heights come from a different distribution than the female heights?
Suppose we have

$$\begin{aligned} maleheight &\sim N(\mu_m, \sigma_m^2) \\ femaleheight &\sim N(\mu_f, \sigma_f^2) \end{aligned} \tag{4.9}$$

and suppose the population is 50% male and 50% female.
Let X be the height of a randomly chosen person. What would be the density function for X?
If X was discrete, then

$$Pr(X = x) = Pr(X = x|male)Pr(male) + Pr(X = x|female)Pr(female) \tag{4.10}$$

The continuous analogue would be:

$$\begin{aligned} f_x(x) &= \frac{1}{2} f_{x|male}(x) + \frac{1}{2} f_{x|female}(x) \\ &= \frac{1}{2} N(X; \mu_m, \sigma_m^2) + \frac{1}{2} N(X; \mu_f, \sigma_f^2) \end{aligned} \tag{4.11}$$

We call the probabilities $Pr(male) = \frac{1}{2}$ and $Pr(female) = \frac{1}{2}$ the "mixture proportions". We call $f_{x|male}(x)$ and $f_{x|female}(x)$ the "component densities".
Returning to our elephant tusk example, suppose we have data $X = (X_1, X_2, ..., X_n)$ on n tusks, and that we know the allele frequencies.

Let the proportion of elephants that are Savannah be Π_S.
Let $Z_i \in \{S, F\}$ denote whether tusk i came from either a Savannah or Forest elephant. We call $\{S, F\}$ the "component labels".
Then we have the mixture model

$$\begin{aligned} P(X_i = x_i|\Pi_S) &= Pr(Z_i = S)Pr(X_i = x_i|Z_i = S) + Pr(Z_i = F)Pr(X_i = x_i|Z_i = F) \\ &= \Pi_S Pr(X_i = x_i|Z_i = S) + (1 - \Pi_S)Pr(X_i = x_i|Z_i = F) \end{aligned} \tag{4.12}$$

More generally

$$PrX_i = x_i = \sum_k \Pi_k Pr(X_i = x_i|Z_i = k) \tag{4.13}$$

where $\Pi_1, ...\Pi_k$ are nonnegative and sum to 1.
The likelihood function of this mixture model is

$$\begin{aligned} L(\Pi_S) &= P(X|\Pi_S) \\ &= \prod_{i=1}^{n} P(X_i = x_i|\Pi_S) \end{aligned} \tag{4.14}$$

When we take the log of this likelihood function, we get

$$\begin{aligned} \ell(\Pi_S) &= \sum_{i=1}^{n} log(Pr(X_i = x_i|\Pi_S)) \\ &= \sum_{i=1}^{n} log[\Pi_S P(X_i = x_i|Z_i = S) + (1 - \Pi_S)P...] \end{aligned} \tag{4.15}$$

Unlike the example with the binomial distribution, this log likelihood is difficult to differentiate, so to find the maximum, we must rely on numerical methods.

4.3 EM Algorithm

The Expectation Maximization (EM) Algorithm is a method for finding maximum likelihood estimates for a model. The key idea behind the EM algorithm is "data augmentation". It is data which we do not have but wish we would have. Suppose our data is X, then the augmented data would be (X, Z), where Z is the "missing data".
Let $L(\theta) = P(X|\theta)$ be the "marginal likelihood"/"observed likelihood". The "complete likelihood" is $L_{comp}(\theta) = P(X, Z|\theta)$.
The steps of the EM algorithm are as follows:

1. Choose some θ_0

2. E step: Form the "expected" complete log likelihood by taking the expectation over Z. In other words find
$$Q(\theta, \theta_0) = E_{Z|X,\theta_0}[\ell_{comp}(\theta; Z, X)] \tag{4.16}$$

3. M step: Choose the value of θ which maximizes $Q(\theta, theta_0)$.

4. The maximizes theta is your new θ_0. Repeat the E and M steps until $\ell(\theta)$ does not change very much.

The advantage of the EM algorith is that the likelihood will always increase with each iteration.
Sometimes the algorithm will converge to a local optimum rather than a global optimum. In practice the algorithm is run multiple times.
Returning to elephant tusk mixture model, which has a complete likelihood

$$\begin{aligned} L(\Pi_S) = P(X, Z|\Pi_S) = \prod_{i=1}^{n} P(X_i, Z_i|\Pi_S) \\ = \prod_{i=1}^{n} P(Z_i|\Pi_S) \propto \prod_{i=1}^{n} \Pi_S^{\mathbb{1}_{Z_i=S}} (1-\Pi_S)^{\mathbb{1}_{Z_i=F}} \end{aligned} \tag{4.17}$$

$\mathbb{1}$ stands for the indicator function, which is 1 for the given event and 0 otherwise.
Taking the log of this expression, we get

$$log(\prod_{i=1}^{n} P(Z_i|\Pi_S)) = \underset{constant}{C} + \sum_i \mathbb{1}(Z_i = S) log(\Pi_S) + \sum_i \mathbb{1}(Z_i = F) log(1-\Pi_S) \tag{4.18}$$

If we take the expectation of the sum of indicator functions, we find the probability of that event occurring. We find that the log likelihood above is maximizes at

$$\frac{\sum_i E(\mathbb{1}(Z_i = S)|X, \theta)}{\sum_i E(...) = n} \tag{4.19}$$