Lecture 4

Tae Kim

1/12/2017

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

Mixture Models and EM algorithm

Modeling with distributions

There are many distributions that can be used for modeling data.

Univariate Distributions	Multivariate Distributions
Normal	Multivariate Normal
Poisson	Multinomial
Exponential	Dirichlet
Binomial	Wishart
Geometric	
Weibull	
Laplace	
Bernoulli	
Gamma	
Beta	

In the elephant tusk example, we've been modeling independent bernoulli variables as a vector.

Let's say we have a count data for gene 1, gene 2, .., gene 10000 from RNA sequencing. Then, the most sensible distribution to model this data would be multinomial.

Notation:

p_i : parameter for frequency of gene i.

x_i : data for count observed for gene i.

N : number of total observations $= \sum_{i=1}^{10000} x_i$

Then, the likelihood of the data given the parameters is

$$p(x_1, .., x_{10,000}; N = 10^7, p_1, .., p_{10,000}) = \frac{N!}{x_1!x_2!..x_{10,000}!} p_1^{x_1}..p_{10,000}^{x_{10,000}}$$

$$= \frac{N!}{\prod_i x_i!} \prod_{i=1}^{10,000} p_i^{x_i}$$

This is equivalent to saying

$$X \sim Mult(N, \mathbf{p})$$

where $\mathbf{p}$ is a vector of probabilities $p_1, .., p_{10,000}$

Note that this model assumes independence of each observation.

Inference with likelihood

One common approach to inference is to estimate parameters through maximizing teh likelihood. Often, we use log-likelihood instead of likelihood due to both mathematical and practical convenience.

Given parameter θ,

$$L(\theta) = p(x|\theta)$$
$$\ell(\theta) = logp(x|\theta)$$
$$\text{MLE } \hat{p} = argmax_p \ell(p)$$

MLE's are often analytic (having a closed form expression), making inference easier.

Mixture Models

Possible examples

(a) Heights of a sample of university population have a heterogeneous distribution because men and women have different heights

(b) Forest elephants and savanna elephants have different allele frequencies

(c) diseased individuals and non-diseased individuals have different protein distributions.

Let's expand on example (b),

Population contains fraction 0.2 ($= \pi_S$) forest elephants and 0.8 ($= \pi_F = 1 - \pi_S$) savanna elephants. Let $\mathbf{x}$ denote genetic data for a single tusk sampled from population. Let f_{F_j} be allele frequency of forest elephant in allele j, while f_{S_j} be allele frequency of savanna elephant in allele j. Then,

$$p(\mathbf{x}) = 0.2p(\mathbf{x}|forest) + 0.8p(\mathbf{x}|savanna)$$
$$= \pi_F(\prod_j f_{F_j}^{x_j}(1-f_{F_j})^{1-x_j}) + \pi_S(\prod_j f_{S_j}^{x_j}(1-f_{S_j})^{1-x_j})$$
$$= p(\mathbf{x}|\pi, \mathbf{f_F}, \mathbf{f_S})$$

The first expression of $p(\mathbf{x})$ is from the law of total probability that

$$p(A) = p(A|B)p(B) + p(A|B^c)p(B^c)$$

or

$$p(A) = \sum_{k=1}^{n} p(A|B_k)p(B_k)$$

if $B_1, ..B_n$ are mutually exclusive and cover all possibility.

When we have more than one sample, we probably need an assumption that each individual is independent. Then, the likelihood becomes

$$p(x_1, .., x_n|\pi, f_F, f_S) = \prod_{i=1}^{n}(\pi_F(\prod_j f_{F_j}^{x_j}(1-f_{F_j})^{1-x_j}) + \pi_S(\prod_j f_{S_j}^{x_j}(1-f_{S_j})^{1-x_j}))$$
$$= L(\pi_F, f_F, f_S; x)$$

EM algorithm

EM algorithm is an iterative method to find parameter that maximizes the likelihood.

Pro : Easy to code

Con : If you're estimating just one parameter, it's probably inefficient, and it's probably better to put it into a one dimensional optimizer. Also, depending on the starting value, it might not converge to the global maximum and instead to a local maximum. (However, since log likehood is a concave function, it's always going to converge to the global maximum.)

EM is also called data augmentation problem because we introduce an additional variable to the algorithm : Z.

Where X_i is the data, Z_i denotes a mixture component that X_i came from. In other words, in the elephant example, $Z_i = Bern(\pi_F)$, saying $Z_i = F$ with probability π_F and $Z_i = S$ with probability π_S.

Then,

$$p(X_i|Z_i = k) \sim \prod_j (f_j^k)^{x_{ij}} (1 - f_j^k)^{1-x_{ij}}$$

The algorithm looks like following.

(a) form the complete data log-likelihood $= log(p(x, z|\theta))$ where θ is a set of parameters we're trying to estimate.

(b) take expectation $E_{z|x,\theta_0} logp(x, z|\theta) = Q(\theta, \theta_0)$ where θ_0 is the initial guess and θ is the true parameters.

(c) maximize the log likelihood by $\theta_1 = argmax_\theta Q(\theta, \theta_0)$. Then, do the iteration again with θ_1