140.778 Homework 1

October 28, 2010

1. Derive the Newton's algorithm for binary logistic regression. Implement the algorithm using R. Download hwdata1-1.txt from the course website http://www.biostat.jhsph.edu/~hji/courses/computing. Use your R function to find the MLE and standard error of the model parameter. Compare your result with the fitted parameter obtained using R's glm function.

2. $Y_1, Y_2, ..., Y_n$ are i.i.d random variables with probability density function $f(Y_i) \sim \pi_0 * \phi(0, \tau^2 + \sigma_i^2) + (1 - \pi_0) * \phi(\mu, \tau^2 + \sigma_i^2)$. Here, $\phi(\mu, \sigma^2)$ is the probability density for a normal distribution with mean μ and variance σ^2. Assume that σ_i^2's are known, but π_0, μ and τ^2 are unknown. Design an EM algorithm to find the MLE of unknown parameters. Provide details of the E-step and M-step. Implement your algorithm using R, then download hwdata1-2.txt from the course website and use your R function to find the MLE of π_0, μ and τ^2. (Hint: $X \sim N(\mu, \tau^2 + \sigma^2)$ can be viewed as a random variable generated hierarchically, i.e., $X|\xi \sim N(\xi, \sigma^2)$, and $\xi \sim N(\mu, \tau^2)$).

3. Convergence rate of EM.
Suppose (y_1, y_2, y_3, y_4) follow a multinomial distribution $M(\sum_i y_i; \theta)$, where $\theta = ((1/2 + \pi/4), (1 - \pi)/4, (1 - \pi)/4, \pi/4)$. You observe $(y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$.

(1) Derive an EM algorithm to estimate π. The algorithm creates a sequence $\pi^{(n+1)} = f(\pi^{(n)})$. Derive the function $f(.)$ using the observed data. (Hint: You can treat $y_1 = x_0 + x_1$, and let $(x_0, x_1, y_2, y_3, y_4)$ follow a multinomial distribution with parameter $(1/2, \pi/4, (1 - \pi)/4, (1 - \pi)/4, \pi/4)$.

(2) Find the fixed point of $f(.)$, i.e., the π^* that solves $\pi = f(\pi)$. This gives the mode that your algorithm will converge to.

(3) Using the results in (1) and (2), find the convergence rate

$lim_{n\to\infty}\frac{\pi^{(n+1)}-\pi^*}{\pi^{(n)}-\pi^*}$.

(4) Based on the Q-function of your algorithm, $Q(\pi'|\pi)$, compute $D^{20}Q(\pi'|\pi)$. Similarly, compute $D^{20}H(\pi'|\pi)$ which relates to the missing information. Finally, compute $D^{20}H(\pi^*|\pi^*)[D^{20}Q(\pi^*|\pi^*)]^{-1}$. Compare this value with the convergence rate you obtained from (3).