140.754 Homework 2

Due date: Apr 26 (Thursday) 11:59pm

Q1. Show that $E[Y|X]$ is the minimum mean square error predictor of Y. That is, show that $g(X) = E[Y|X]$ minimizes $E[(Y - g(X))^2]$ among all functions $g(.)$ of X.

Q2. **EM in Gaussian mixture**. Suppose $x_1, \ldots, x_n$ are n observations independently sampled from a distribution with probability density function $f(x|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \theta_1\phi(x; \mu_1, \sigma_1^2) + \ldots + \theta_K\phi(x; \mu_K, \sigma_K^2)$. Here $\phi(x; \mu, \sigma^2)$ represents density function of a normal distribution with mean μ and variance σ^2. $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_k\}$, $\boldsymbol{\sigma^2} = \left\{\sigma_1^2, \ldots, \sigma_k^2\right\}$ are unknown parameters. $\sum_{k=1}^K \theta_k = 1$. Derive an EM algorithm for estimating $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$.

Q3. The data in hw1-1.txt are generated from the following random effects model.

$$\begin{aligned} y_{ij}|\mu_i &\overset{ind}{\sim} N(\mu_i, \sigma^2) \quad i = 1, \ldots, m; j = 1 \ldots, n_i \\ \mu_i &\overset{i.i.d}{\sim} N(\mu, \sigma_a^2) \end{aligned}$$

(1) Implement an EM algorithm for estimating μ, σ^2 and σ_a^2. Run your algorithm on the data in hw1-1.txt.
(2) Draw a plot to show that the observed data log-likelihood is non-decreasing when your EM iterates.
(3) Report your estimates for μ, σ^2 and σ_a^2
(4) Predict μ_i and report your predictions.

Q4. For the same data in Q3,
(1) Derive and implement an MCMC algorithm to infer unknowns.
(2) Draw plots to monitor convergence of your MCMC.
(3) Provide point and interval estimates for μ, σ^2 and σ_a^2.
(4) Provide predictions for μ_i.

Q5. **Gibbs Sampler for DNA motif discovery.**

1. Assume that DNA is a mixture of motif sites and background nucleotides. The motif sites are generated according to a probability matrix $\mathbf{\Theta}$ and the background nucleotides are generated according to a background probability vector $\boldsymbol{\theta}_0$. The length of the motif is W. W and $\boldsymbol{\theta}_0$ are known, but $\mathbf{\Theta}$ is unknown. Assume that there are N DNA sequences and each sequence has exactly one motif site. Let $\mathbf{A}$ be the location indicators of the motif sites. Derive a Gibbs sampler to find the motif sites after collapsing $\mathbf{\Theta}$ (i.e., provide an algorithm that samples $\mathbf{A}$ after integrating out $\mathbf{\Theta}$ analytically).

2. Implement your motif sampler. Download the data Seq.txt from the course website. The data contains 30 DNA sequences, each starting with a ">" and a sequence name. Run your motif sampler in (1) by assuming that the motif length is 18 bp. You can use the empirical frequencies of A, C, G and T in the homework data set as your $\boldsymbol{\theta}_0$. Based on the $\mathbf{A}$ obtained from the last iteration of your chain, estimate the motif probability matrix $\mathbf{\Theta}$. Collect the sequences covered by the motif site based on the last sample of $\mathbf{A}$, save them. Then go to the website http://weblogo.berkeley.edu/logo.cgi to create a sequence logo (a way to visualize motif) for the motif you found.