Homework 5
10-702/36-702 Statistical Machine Learning

Due: Friday April 1 3:00

Hand in to: Michelle Martin GHC 8001.

1. **Simulation, E.M. and Variational Approximations.** Let $X_1, \ldots, X_n \sim g(x; p)$ where

$$g(x; p) = pf_0(x) + (1-p)f_1(x).$$

For simplicity, we will assume that f_0 and f_1 are one-dimensional Gaussian distributions with **known means and variances**. The only unknown is p. The problem is to estimate p.

(a) Derive the explicit steps for the EM algorithm for finding the MLE of p.

(b) Suppose we take a Bayesian approach with a Beta(α, β) prior for p. The posterior for p given $X^n = (X_1, \ldots, X_n)$ is

$$\pi(p \,|\, X_1, \ldots, X_n) \propto \mathcal{L}_n(p)\pi(p)$$

where the likelihood is

$$\mathcal{L}_n(p) = \prod_{i=1}^{n} (pf_0(x_i) + (1-p)f_1(x_i))$$

and the prior is

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}.$$

Derive the steps for the Gibbs sampling algorithm (by introducing latent variables).

(c) Derive a random walk MCMC algorithm. (You will need to work with a transformation of p such as $\psi = h(p) = \log(p/(1-p))$; otherwise the boundaries of the unit interval will cause problems.)

(d) Implement the algorithms from parts (a), (b) and (c). Simulate $n = 25$ observations from the model

$$\frac{1}{3}N(0,1) + \frac{2}{3}N(3,1).$$

Use a Beta$(4,4)$ prior distribution over p. For the mle, compare the EM estimate with the exact MLE (which you can compute numerically). For the Bayesian analysis, show trace plots and compare the approximate posterior with the exact posterior (obtained numerically).

(e) Derive the mean field variational approximation of the posterior. Run the variational approximation for the same data and compare with the exact answer.

2. **Nonparametric Density Estimation Using Bayesian Simulation Methods.** In this problem, you will estimate an unknown density using a mixture of Gaussians, as described in Ishwaran and James (2002), *'Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information'*.

   Consider the Bart Simpson density

$$p(x) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{10}\sum_{j=0}^{4} \phi(x; (j/2) - 1, 1/10),$$

   where $\phi(x; \mu, \sigma)$ denotes a Gaussian density with mean μ and standard deviation σ. Draw $n = 1000$ observations from p.

   Use the following hierarchical model to estimate the true density:

$$\begin{aligned} F &\sim F_0 \\ (\mu_i, \sigma_i) \,|\, F &\sim F \\ X_i \,|\, \mu_i, \sigma_i &\sim \mathcal{N}(\mu_i, \sigma_i), \end{aligned}$$

   where $F_0 = \sum_{k=1}^{N} w_k \delta_{(m_k, s_k)}$ is a random probability measure and $\boldsymbol{w} = (w_1, \ldots, w_N)$ are random weights chosen using the stick-breaking construction

$$\begin{aligned} w_1 &= V_1 \\ w_k &= V_k \prod_{i=1}^{k-1} (1 - V_i) \qquad k = 2, \ldots, N, \end{aligned}$$

   where $V_1, V_2, \ldots, V_{N-1} \sim \text{Beta}(1, \alpha)$ and $V_k = 1$ to ensure that $\sum_k w_k = 1$. The priors on $\{(m_k, s_k)\}_{k=1,\ldots,N}$ and α are set as follows

$$\begin{aligned} \theta &\sim \mathcal{N}(0, A) \\ m_k \,|\, \theta, \sigma_m &\sim \mathcal{N}(\theta, \sigma_m) \\ (s_k^2)^{-1} \,|\, \nu_1, \nu_2 &\sim \text{Gamma}(\nu_1, \nu_2) \\ \alpha \,|\, \eta_1, \eta_2 &\sim \text{Gamma}(\eta_1, \eta_2), \end{aligned}$$

   where $A, \sigma_m, \nu_1, \nu_2, \eta_1, \eta_2$ are hyperparameters. We are going to set the hyperparameters as follows: $A = 1000$, $\nu_1 = \nu_2 = 2$, σ_m is set equal to 4 standard deviations of the data and $\eta_1 = \eta_2 = 2$.

   Denote by $(K_1, \ldots, K_m)$ the classification variables that map the realization of (μ_i, σ_i) to a particular cluster (m_k, s_k). Note that $K_i \in \{1, \ldots, N\}$ and

$$K_i \,|\, \boldsymbol{w} \sim \sum_{k=1}^{N} w_k \delta_k.$$

You will implement the blocked Gibbs sampler to explore the posterior $\mathcal{P}_N \,|\, \boldsymbol{X}$. The blocked Gibbs sampler is implemented by iteratively drawing values from the following conditional distributions:

$$\begin{aligned} \boldsymbol{m} &\mid \boldsymbol{s}, \boldsymbol{K}, \theta, \boldsymbol{X} \\ \boldsymbol{s} &\mid \boldsymbol{m}, \boldsymbol{K}, \boldsymbol{X} \\ \boldsymbol{K} &\mid \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{X} \\ \boldsymbol{w} &\mid \boldsymbol{K}, \alpha \\ \alpha &\mid \boldsymbol{w} \\ \theta &\mid \boldsymbol{m}. \end{aligned}$$

The method eventually produces a draw from the distribution $(\boldsymbol{m}_,, \boldsymbol{K}, \boldsymbol{w}, \alpha, \theta \,|\, \boldsymbol{X})$. These values produce a random probability measure

$$\mathcal{P}_N^*(\cdot) = \sum_{k=1}^{N} w_k \delta_{(m_k, s_k)}(\cdot),$$

which is a draw from the posterior $\mathcal{P}_N \,|\, \boldsymbol{X}$.

The predictive density $f(x \,|\, \boldsymbol{X})$ can be approximated as

$$f(x \,|\, \boldsymbol{X}) \approx \frac{1}{B} \sum_{b=1}^{B} \sum_{k=1}^{N} w_k^{(b)} \phi(x; m_k^{(b)}, s_k^{(b)}),$$

where $(\boldsymbol{m}^{(b)}, \boldsymbol{s}^{(b)}, \boldsymbol{w}^{(b)})$ are different realizations of $\mathcal{P}_N \,|\, \boldsymbol{X}$.

(a) Implement the Gibbs sampler using the form of conditional probabilities on page 10 and 11 of Ishwaran and James (2002). Use $N = 50$ and $B = 100$. Plot the predictive density.

(b) Compare to the kernel density estimator.