10-702 Statistical Machine Learning: Assignment 5

Due Friday, April 2

Hand in to Sharon Cavlovich, GHC (Gates Hillman Center) 8215 by 3:00, and email your code to `xichen@cs` and `mladenk@cs`. Use R for all numerical computations.

Problem 1. *Simulation and variational approximations*

The objective of this problem is to derive and experiment with a variational approximation for a simple Bayesian mixture model and compare it to MCMC.

Let $X_1, \ldots, X_n \sim g(x \,|\, p)$ where

$$g(x \,|\, p) = pf_0(x) + (1-p)f_1(x),$$

the densities f_i are Gaussian, and p is an unknown mixing weight. The problem is to find the posterior over p given $X^n = (X_1, \ldots, X_n)$ under a Beta(α, β) prior. The posterior is thus

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

where the likelihood is

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

and the prior is

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

(a) Derive the Gibbs sampling algorithm discussed in class for the posterior on p. Also derive a random walk MCMC algorithm. (For the latter, 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.)

(b) Implement the Gibbs sampling algorithm and MCMC random walk algorithm in R, and run it for the following four models. In each case use a Beta$(0.8, 0.8)$ prior distribution over p. Generate $p_1, \ldots, p_N$ from the posterior with $N = 10,000$ in each simulation.

1:	$n = 25$	$p = 0.18$	$f_0 = \mathcal{N}(0, 0.9)$	$f_1 = \mathcal{N}(0.8, 0.9)$
2:	$n = 75$	$p = 0.15$	$f_0 = \mathcal{N}(0, 0.9)$	$f_1 = \mathcal{N}(0.8, 0.4)$
3:	$n = 150$	$p = 0.40$	$f_0 = \mathcal{N}(0, 1.2)$	$f_1 = \mathcal{N}(0, 0.4)$
4:	$n = 500$	$p = 0.30$	$f_0 = \mathcal{N}(3.2, 3.2)$	$f_1 = \mathcal{N}(1.4, 1.4)$

For each case, you should generate your own data, and plot the estimated posterior distribution obtained by the Gibbs sampler. Also, make a trace plot for each simulation.

(c) Derive the mean field variational approximation of the posterior. Give the algorithm and its mathematical derivation in your solutions.

(d) Now run the variational approximation for the same models used in part (b). Produce plots that show the approximate posterior against the true posterior. What are your conclusions about the accuracy of the variational approximation for this problem?

Problem 2. *Bayesian inference and Simulation*

Get the fusion time data from

`http://lib.stat.cmu.edu/DASL/Datafiles/FusionTime.html`

The experiment is described as follows:

"Results of an experiment on the effect of prior information on the time to fuse random dot stereograms. One group (NV) was given either no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (e.g., a drawing of the object)."

There are two variables:

1. Time: Time to produce a fused image of the random dot stereogram.
2. Group: Treatment group (NV or VV).

Model the data from the two groups as:

$X_1, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$

$Y_1, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$

Here, the X_i's and the Y_i's are the *log times*. We are interested in $\delta = \mu_1 - \mu_2$.

(a) Draw boxplots of the data.

(b) Compute a 95 percent confidence interval for δ.

(c) Now perform a Bayesian analysis. Use the prior

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2) \propto \frac{1}{\sigma_1 \sigma_2}.$$

Write down an expression for the posterior density for $(\mu_1, \mu_2, \sigma_1, \sigma_2)$.

(d) Simulate from the posterior using direct sampling. Hint: Let D denote the data and write

$$p(\mu_1, \mu_2, \sigma_1, \sigma_2 \mid D) = p(\mu_1 \mid \sigma_1, D)p(\sigma_1 \mid D)p(\mu_2 \mid \sigma_2, D)p(\sigma_2 \mid D).$$

Get the 95 percent posterior interval for δ and compare to the confidence interval.

(e) Simulate from the posterior using Gibbs sampling. Get the 95 percent posterior interval for δ and compare to the confidence interval.

(f) Simulate from the posterior using Metropolis-Hastings sampling. Get the 95 percent posterior interval for δ and compare to the confidence interval.

Problem 3. *Nonparametric density estimation*

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. You data set will consist of 1000 points drawn from p.

Use the following hierarchical model to estimate the true density:

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

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 $(K_1, \ldots, K_m)$ 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) Explain why the Dirichlet process can be approximated with a finite mixture and give some guidance for the choice of N.

(b) Explain the intuition behind the choice of the hyperparameters and priors. What is the role of θ? What does the choice of $A = 1000$ correspond to? What is the role of σ_m? What is the effect of the parameters ν_1 and ν_2 on the estimation? Explain the role of α in the approximate Dirichlet process. What does the choice of $\eta_1 = \eta_2 = 2$ encourage?

(c) Implement the block Gibbs sampler using the form of conditional probabilities on page 10 and 11 of Ishwaran and James (2002).

(d) Set $N = 50$ and $B = 100$. Plot 30 predictive densities.

(e) Compare to the kernel density estimator in Chapter 26.