Probabilistic Graphical Models

HWK 3 (Part A)

Assigned Sunday, March 8, 2020

Due: Tuesday, March 31, 2020

Problem 1 Markov Chain Monte Carlo (15 Points)

1. A simple sampling method adopted by many of the standard math libraries is the *inverse probability transform*: draw $u \sim \text{Unif}\,(0,1)$, then draw $x \sim F^{-1}(u)$, where F^{-1} is the inverse of the cdf. Show that x generated by this procedure follows distribution F. What is the drawback of this method?

2. We have learned transition matrix in class which only applies to discrete state Markov chains. A more general concept is the transition kernel, which can apply to both discrete and continuous state spaces. The transition kernel $k(x, x')$ is defined as $k(x, x') = P(\theta^{(t+1)} = x' | \theta^{(t)} = x)$, where $\theta^{(t)}$ is the state at time t. Show that if both transition kernels K_1 and K_2 have $p(\cdot)$ as stationary density, so do $K_1 K_2$ and $\lambda K_1 + (1-\lambda)K_2$ for any $\lambda \in [0,1]$. In practice, the former corresponds to sampling from K_1 and K_2 *cyclically* and the latter draws either K_1 with probability λ or K_2 otherwise. Note that in the continuous case, the cyclic kernel can be defined as composition of functions:

$$(K_1 \circ K_2)(x, z) = \int K_2(x, y) K_1(y, z)\, \mathrm{d}y. \tag{1}$$

3. Recall MH sampling for target distribution $p(x)$ using proposal $q(x|y)$: at state s, first draw $t \sim q(t|s)$, then accept t with probability

$$A = \min\left(1, \frac{\widetilde{p}(t) q(s|t)}{\widetilde{p}(s) q(t|s)}\right), \tag{2}$$

   where $\widetilde{p}(x)$ is the unnormalized target distribution. Show that $p(x)$ is the stationary distribution of the Markov chain defined by this procedure. Consider both continuous and discrete cases.

4. Recall Gibbs sampling for target distribution $p(\mathbf{x}) = p(x_1, \ldots, x_d)$: for each $j \in \{1, \ldots, d\}$, draw $t \sim p(x_j | \text{rest})$ and set $x_j = t$. Show that $p(\mathbf{x})$ is the stationary distribution of the Markov chain defined by this procedure.

Problem 2 Box-Muller method (10 points)

Let $x_1 \sim U\left(x_1 | [0,1]\right)$, $x_2 \sim U\left(x_2 | [0,1]\right)$ *and*

$$y_1 = \sqrt{-2\log x_1}\cos 2\pi x_2, \qquad y_2 = \sqrt{-2\log x_1}\sin 2\pi x_2$$

Show that

$$p(y_1, y_2) = \int p(y_1|x_1, x_2)p(y_2|x_1, x_2)p(x_1)p(x_2)dx_1dx_2 = \mathcal{N}\left(y_1|0,1\right)\mathcal{N}\left(y_2|0,1\right)$$

and suggest an algorithm to sample from a univariate normal distribution. Hint: Use the change of variable result for vectors $y = (y_1, y_2)$ and $x = (x_1, x_2)$.

Problem 3 (6 points)

Consider an Ising model on an $M \times M$ square lattice with nearest neighbour interactions:

$$p(x) \propto \exp \beta \sum_{i \sim j} \mathbb{I}\left[x_i = x_j\right]$$

Now consider the $M \times M$ grid as a checkerboard, and give each white square a label w_i, and each black square a label b_j, so that each square is associated with a particular variable. Show that

$$p(b_1, b_2, \ldots, |w_1, w_2, \ldots) = p(b_1|w_1, w_2, \ldots)p(b_2|w_1, w_2, \ldots) \ldots$$

That is, conditioned on the white variables, the black variables are independent. The converse is also true, that conditioned on the black variables, the white variables are independent. Explain how this can be exploited by a Gibbs sampling procedure. This procedure is known as checkerboard *or* black and white sampling.

Problem 4 **Approximate Inference via Sampling (10 points)**

Consider a setting in which there are **D** diseases and a patient can either have or not have a particular disease $\mathbf{d_i} \in \{0, 1\}$, for $\mathbf{i = 1,...,D}$. Here $\mathbf{d_i = 1}$ means that the patient has disease '**i**'; $\mathbf{d_i = 0}$ means the patient does not have disease '**i**'. A patient may have more than one disease. There are a set of **S** symptoms the hospital can measure; if $\mathbf{s_j = 1}$, for $\mathbf{j = 1,...,S}$ then the patient has symptom '**j**'; otherwise $\mathbf{s_j = 0}$ means that the patient does not have symptom '**j**'. A simple disease-symptom network is given by

$$p(s_1, \ldots, s_S, d_1, \ldots, d_D) = \prod_{j=1}^{S} p(s_j|\mathbf{d}) \prod_{i=1}^{D} p(d_i)$$

where $\mathbf{d} = (d_1, \ldots, d_D)^T$ *and*

$$p(s_j = 1|\mathbf{d}) = \sigma\left(\mathbf{w}_j^T \mathbf{d} + b_j\right)$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

In the above $\mathbf{w_j}$ is a vector of parameters relating symptom **j** to the diseases and $\mathbf{b_j}$ is related to the prevalence of the symptom. The hospital provides the collection of parameters **W** and **b**, the prior disease probabilities **p** (with $\mathbf{p(d_i = 1) = p_i}$) and a vector **s** of symptoms for the patient, see SymptomDiseasePars.mat.

Use Gibbs sampling (using a reasonable amount of burn-in and sub-sampling) to estimate the vector

$$[p(d_1 = 1|\mathbf{s}), \ldots, p(d_D = 1|\mathbf{s})]$$

Problem 5 **Approximate Inference via Sampling (10 points)**

For a disease-symptom network similar to the previous question and a collection of N patient records $D = (s^n, d^n),\ n = 1, ..., N$, derive the following for the prediction of the diseases for a new patient with symptoms s:.

$$p(\mathbf{d}|\mathbf{s}, \mathcal{D}) = \int_{\mathbf{W},\mathbf{b},\mathbf{p}} p(\mathbf{d}|\mathbf{s}, \mathbf{W}, \mathbf{b}, \mathbf{p}) p(\mathbf{W}, \mathbf{b}, \mathbf{p}|\mathcal{D})$$

where

$$p(\mathbf{W}, \mathbf{b}, \mathbf{p}|\mathcal{D}) \propto p(\mathbf{W}, \mathbf{b}, \mathbf{p}) \prod_{n=1}^{N} p(\mathbf{s}^n|\mathbf{d}^n, \mathbf{W}, \mathbf{b}) p(\mathbf{d}^n|\mathbf{p})$$

and explain how you could estimate

$$p(d_i = 1|\mathbf{s}, \mathcal{D})$$

using sampling.

Problem 6 MCMC Sampling (10 points) (CMU spring 2016-HWK3)

Assume we have a mixture of Gaussians model to generate $x \in R$ given parameters μ_1, μ_2:

$$p(x|\mu_1, \mu_2) \propto \frac{1}{2}\exp(-\frac{1}{2}(x-\mu_1)^2) + \frac{1}{2}\exp(-\frac{1}{2}(x-\mu_2)^2)$$

Assume we have N samples $x_1, \cdots, x_N$ drawn from the model. We also impose a Gaussian prior on μ_1 and μ_2:

$$p(\mu_1, \mu_2) \propto \exp(-\frac{1}{2}\frac{\mu_1^2}{100})\exp(-\frac{1}{2}\frac{\mu_2^2}{100})$$

In the below questions, you will be asked to sample from the posterior using Metropolis–Hastings sampling methods, although we can directly sample from it. We have picked the simplest possible model we could think of so that the mechanics of the samplers are incredibly basic. Of course, one would never actually use Metropolis–Hastings for this posterior, but we want to make the behavior of the sampler very easy to understand.

a) Generate 100 samples $x_1, \cdots, x_{100}$ from $0.5\mathcal{N}(-5, 1^2) + 0.5\mathcal{N}(5, 1^2)$ where $\mathcal{N}(\mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2. You can use any package to sample from a Gaussian. You do not need to visualize the samples, just submit your code to get full credit.

b) Use Metropolis Hastings Sampling to sample from $P(\mu_1, \mu_2 \mid x_1, \cdots, x_{100})$. Let the proposal distribution $q(\mu'_1, \mu'_2 \mid \mu_1, \mu_2)$ be $\mathcal{N}(\mu_1, \sigma^2)\mathcal{N}(\mu_2, \sigma^2)$. Experiment with $\sigma = 0.5$ and $\sigma = 5$ respectively, let the initial point be $(0, 0)$, discard the first 10000 samples and plot the next 1000 samples. Calculate the acceptance rate during the whole process. Report the estimated mean of μ_1 and μ_2. Repeat the experiment 6 times. What's the difference between $\sigma = 0.5$ and $\sigma = 5$?

c) Use Gibbs Sampling to sample from $P(\mu_1, \mu_2 \mid x_1, \cdots, x_{100})$. In order to make the sampling simple, introduce latent variables $z_1, \cdots, z_{100}$ for $x_1, \cdots, x_{100}$ indicating the Gaussian component where the data comes from. Discard the first 10000 samples and plot the next 1000 samples. Report the estimated mean of μ_1 and μ_2. Repeat the experiment 6 times.