Homework 5
Statistical Machine Learning
10/36-702
Due Friday April 20

1. 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

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

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. Generate $n = 400$ data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ as follows. Take $X_1, \ldots, X_n \sim$ Uniform$(-1, 1)$. Take

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i$$

where $\epsilon_1, \ldots, \epsilon_n \sim N(0, 1)$,

$$m(x) = \begin{cases} (x+2)^2/2 & -1 \le x < -0.5 \\ x/2 + 0.875 & -0.5 \le x < 0 \\ -5(x-0.2)^2 + 1.075 & 0 \le x < 0.5 \\ x + 0.125 & 0.5 \le x < 1 \end{cases}$$

and

$$\sigma(x) = 0.2 - 0.1\cos(2\pi x).$$

Randomly split the data into two sets of $n = 200$ observations each. The first half is the training data and the second is the testing data.

(a) Estimate m using kernel regression. Use a Gaussian kernel. Choose the bandwidth by cross-validation (using the test data). Plot the true function, the data and the estimated function. Plot the residuals. Plot the cross-vaidation function as a function of h.

(b) Now estimate m using RKHS methods. Specifically, choose $\widehat{m}$ to minimize

$$\sum_{i=1}(Y_i - m(X_i))^2 + \lambda ||m||_K^2$$

where the kernel K is $K(x, y) = e^{-(x-y)^2/\sigma^2}$. There are two tuning parameters, λ and σ. Choose both by cross-validation (using the test data). Make the same plots as in (a). Comment on the differences/similarities between the two estimates.

3. Let $\mathcal{F}$ denote all real-valued functions on $[0, 1]$ with m continous derivatives. Define the kernel

$$K(x, y) = \sum_{s=0}^{m-1} \frac{x^s}{s!}\frac{y^s}{s!} + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du$$

and inner predict

$$\langle f, g \rangle = \sum_{s=0}^{m-1} f^{(s)}(0)g^{(s)}(0) + \int_0^1 f^{(m)}(x)g^{(m)}(x)dx.$$

Verify that this kernel has the reproducing property: $\langle K_x, f \rangle = f(x)$.

Hint: By Taylor's theorem with remainder, we can write

$$f(x) = \sum_{s=0}^{m-1} \frac{x^s}{s!} f^{(s)}(0) + \int_0^1 \frac{(x-u)_+^{(m-1)}}{(m-1)!} f^{(m)}(u) du.$$

4. Let $Y_1, \ldots, Y_n \sim p$ where $Y_i \in \mathbb{R}$. Let

$$\widehat{p}_h(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{||y - Y_i||}{h}\right).$$

Let $L = \{p > \lambda\}$ and $\widehat{L} = \{\widehat{p} > \lambda\}$. Assume that p is smooth. Also assume that there exist positive constants δ, c, C such that: if $p(x) \in [\lambda - \delta, \lambda + \delta]$ then

$$c < |p'(x)| < C.$$

Prove that

$$H(L, \widehat{L}) \xrightarrow{P} 0$$

where H is Hausdorff distance.