Homework 4
Statistical Machine Learning
10/36-702
Due Friday March 30

1. (Nonparametric Bayes) Let $X_1, \ldots, X_n \sim F$ where $X_i \in \mathbb{R}$ and $F(x) = \mathbb{P}(X \leq x)$ is the unknown cdf. Let the prior π for F be $\mathrm{DP}(\alpha, F_0)$.

   (a) Show that $\mathbb{E}_\pi(F(x)) = F_0(x)$.

   (b) Let $F_0 = N(0,1)$ and $\alpha = 1$. Draw 100 random distribution functions from the prior. Plot them. Verify that the mean of these 100 cdf's is close to F_0.

   (c) Suppose that the true distribution is $F = N(10, 4)$. Draw $X_1, \ldots, X_{10} \sim F$. Find the Bayes estimator of F. Plot it and compare it to the true F. Draw 1000 random distribution functions from the posterior. Use these to find a 95 percent posterior band for F. Plot the true F and the 95 percent posterior posterior band.

   (d) Let $\overline{F}_n(x)$ be the posterior Bayes estimator of $F(x)$. Here x is some arbitrary, fixed value. Find the bias and variance of $\overline{F}_n(x)$. Let $F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$ be the empirical cdf. When is the the mean squared error of $\overline{F}_n(x)$ smaller than the mean squared error of $F_n(x)$?

   (e) Use Hoeffding's inequality to get bounds on

$$\mathbb{P}(|F_n(x) - F(x)| > \epsilon)$$

   and

$$\mathbb{P}(|\overline{F}_n(x) - F(x)| > \epsilon).$$

2. (Minimax Theory) For $0 \leq x \leq 1$ and $0 < \theta \leq 1$ define the density

$$p_\theta(x) = (1-\theta)I(x \leq 1/2) + (1+\theta)I(x > 1/2).$$

We want to estimate θ. The loss function is $d(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$.

(a) Use LeCam's lemma to show that the minimax risk is bounded below by $C_1/\sqrt{n}$ for some $C_1 > 0$.

(b) Find an estimator whose risk is bounded above by $C_2/\sqrt{n}$ for some $C_2 > 0$.

3. (Undirected Graphs)

(a) Let $X = (X_1, \ldots, X_d)$ have a multivariate distribution with mean μ and covariance matrix Σ. Let $\Omega = \Sigma^{-1}$. Show that $X_j \amalg X_k | \mathsf{rest}$ if and only if $\Omega_{jk} = 0$, where rest refers to all the other variables.

(b) Let $X = (X_1, X_2, X_3)$ where X_1, X_2 and X_3 are binary random variables. Let $p(x) = \mathbb{P}(X = x) = \mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)$. Consider the loglinear model

$$\log p(x) = \beta_0 + \beta_1 x_1 + \beta_{12} x_1 x_2.$$

(i)Is this model hierarchical?
(ii)Is this model graphical?
(iii) Draw the undirected graph corresponding to this model.
(iv) Let $\beta_1 = \beta_{12} = 2$. What is the value of β_0?
(v) Simulate 100 observations from the model in part (iv). Fit the saturated model

$$\log p(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$$

by maximum likelihood. Test each coefficient and remove all non-significant terms. What is the resulting graph?

Hint: In R, create a vector of counts (of length $8 = 2^3$). Also, create vectors $x1$, $x2$ and $x3$ each of length 8. Then use the commands:

```
out = glm(count ~ x1+x2+x3+ x1*x2 + x2*x3 + x1*x3 + x1*x2*x3,family="poisson")
summary(out)
```