10-702 Statistical Machine Learning: Assignment 6

Due Friday, May 9

(1) (Principal Curves)
(a) Generate data as follows:

```
sigma = 0.01
n = 100
y = seq(0,1,length=n) + rnorm(n,0,sigma)
x = cos(5*pi*y) + rnorm(n,0,sigma)
```

Plot y versus x. Fit a principal curve. (You can use the R function for this.) Try different smoothing parameters. Now repeat with $\sigma = 0.1$. Report the training error in each case.

(2) (Spectral Clustering)
Use the following code to generate data:

```
th = seq(0,2*pi,length=150)
x  = c(cos(th),1.5*cos(th),2*cos(th))
y  = c(sin(th),1.5*sin(th),2*sin(th))
X  = cbind(x,y)
plot(X[,1],X[,2])
```

Use spectral clustering to cluster the data. Choose a bandwidth h and an integer m. Construct the spectral eigenvectors $v_1, v_2, \ldots$ as we discussed in class. Define new data $Z_1, \ldots, Z_n$ where $Z_i = (v_1(i), \ldots, v_m(i))^T$. Apply k-means clustering to the Z_i's with $k = 3$. Experiment with different values of h and m. Summarize your results.

(3) Kernels
Let

$$\mathcal{F} = \left\{ f : \ ||f||_K \leq B \right\}$$

where K is a Mercer kernel. Recall that $K(x,y) = \langle \phi(x), \phi(y) \rangle$ where

$$\phi(x) = (\phi_1(x), \phi_2(x), \ldots) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots)$$

where $\psi_1, \psi_2, \ldots$ are the eigenfunctions of K.

(a) If $f \in \mathcal{F}$ we can write $f(x) = \sum_i \alpha_i K(x_i, x)$. We can also write $f = \sum_j \beta_j \phi_j$. Show that $\beta_\ell = \sum_i \alpha_i \phi_\ell(x_i)$.

(b) Show that $||f||_K \leq B$ implies that $||\beta|| \leq B$.

(c) Show that the Rademacher complexity satisfies

$$R_n(\mathcal{F}) \leq B\sqrt{\frac{\kappa}{n}}$$

where $\kappa = \sup_x K(x,x)$. Hint: Write $f = \sum_j \beta_j \phi_j$. Take the definition of $R_n(\mathcal{F})$ and apply the Cauchy-Schwarz inequality.

Comment: Part (c), toegether with McDiarmid's inequality shows that

$$\mathbb{P}\left(P(f) \leq P_n(f) + B\sqrt{\frac{\kappa}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}, \quad \text{for all } f \in \mathcal{F} \right) \geq 1 - \delta.$$

(d) Here we will compare "smoothing kernel regression" and "Mercer kernel" regression. Define

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

Let

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

Generate $X_1, \ldots, X_n$ uniformly on $[-1, 1]$ where $n = 200$. Then take

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon_i \sim N(0,1)$. Use a Gaussian kernel. Compare the best "smoothing kernel" fit and the best "Mercer kernel" fit. Choose the degree of regularization by GCV.

(4) Bayesian inference and MCMC

Consider the following Bayesian regression model:

$$y_i = \sum_{j=1}^{D} \beta_j x_{ij} + \epsilon_i$$

where $i = 1, \ldots, n$, the noise is

$$\epsilon_i \sim N(0, 1/\rho_\epsilon)$$

and the prior for the parameters is:

$$\beta_j \sim N(0, 1/\rho_\beta).$$

Assume the precisions have exponential priors:

$$\rho_\epsilon \sim \text{Expon}(\lambda) = \lambda \exp\{-\lambda \rho_\epsilon\}$$

$$\rho_\beta \sim \text{Expon}(\lambda) = \lambda \exp\{-\lambda \rho_\beta\}$$

To keep things simple, we'll assume $D = 1$, and $\lambda = 1$.

(a) Draw 1000 samples from the prior predictions for y at $x = 1$:

$$p(y_i | x_{i1} = 1, \lambda = 1),$$

display a histogram of the samples and compute the mean and variance of the samples.

(b) Consider a data set of $n = 3$, (x, y) pairs, $\mathcal{D} = \{(0, 0), (2, 2.1), (3, 3)\}$. Derive the Gibbs sampling updates for:

$$p(\rho_\epsilon, \rho_\beta, \beta_1 | \mathcal{D}, \lambda = 1)$$

In other words, derive the distribution of each of the above 3 variables being sampled, conditional on the other variables (you will need to use Gamma distributions for the ρs).

(c) Write code that will implement a Gibbs sampler to draw samples from $p(\rho_\epsilon, \rho_\beta, \beta_1 | \mathcal{D}, \lambda = 1)$. Plot traces of these variables as a function of sampling iteration. Discuss how many iterations are needed for "burn in" and "convergence". Display histograms of the three variables, and also of $p(y_i | x_{i1} = 1, \mathcal{D}, \lambda = 1)$. Are the results reasonable?