10-702 Statistical Machine Learning: Assignment 5

Due Friday, April 11

Do 5 of the following 6 questions. If you do all 6, you will get bonus points.

(1) Let $\mathcal{D} = \{\psi_1, \psi_2, \ldots\}$ be an orthonormal dictionary. Recall that

$$L_{w,p}(C) = \left\{ f = \sum_j \beta_j \psi_j : \ \ |\beta_{(j)}| \leq \frac{C}{j^{1/p}}, \ \ j = 1, 2, \ldots, \right\}.$$

Any f in $L_{w,p}(C)$ can be written as $f = \sum_j \beta_j \psi_j$ where $\beta_j = \langle f, \psi_j \rangle$. Define

$$\sigma_N(f) = \inf_{|\Lambda| \leq N} \inf_{g \in \text{Span}(\Lambda)} ||f - g||.$$

(a) Show that the best N-atom approximation to f is $f_N = \sum_{j \in J_N} \beta_j \psi_j$ where J_N are the indices of the the N largest values of of $|\beta_j|$.

(b) Show that OGA recovers f_N exactly.

(c) Show that, if $0 < p < 2$ then

$$\sigma_N = O\left(\frac{1}{N^s}\right)$$

where $s = (1/p) - (1/2)$.

(2) Let $X_1, \ldots, X_n$ be iid and suppose that $|X_i| \leq c$ and $\mathbb{E}(X_i) = 0$. Show that, for every $\delta > 0$,

$$\mathbb{P}\left(|\overline{X}_n| > \sigma \sqrt{\frac{2\delta}{n}} + \frac{2c\delta}{3n} \right) \leq e^{-\delta}.$$

(3) We introduced covering numbers in class. A related idea is bracketing. Let $\mathcal{F}$ be a set of functions. If ℓ and u are two functions, we define the **bracket**

$$[\ell, u] = \left\{ f : \ \ell(x) \leq f(x) \leq u(x) \ \text{ for all } x \right\}.$$

We say that $[\ell, u]$ is an ϵ bracket in $L_r(P)$ if

$$\int (u(x) - \ell(x))^r \leq \epsilon^r.$$

The bracketing number $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the smallest number of ϵ brackets needed to cover $\mathcal{F}$.

(a) Suppose that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$. Also suppose that $||f||_\infty \leq B < \infty$ for every $f \in \mathcal{F}$. Show that

$$\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \xrightarrow{p} 0.$$

(b) Let $\mathcal{F}$ be the set of all indicator functions of the form $f_t = I_{(-\infty, t]}$ for $t \in \mathbb{R}$. Show that

$$N_{[]}(\epsilon, \mathcal{F}, L_1(P)) \leq \frac{C}{\epsilon}$$

for some $C > 0$.

Comment: From (a) and (b) you have proved that

$$\sup_t |\widehat{F}_n(t) - F(t)| \xrightarrow{p} 0$$

where $F(t) = \mathbb{P}(X \leq t)$ is the cdf and $\widehat{F}_n(t) = n^{-1} \sum_{i=1}^n I(X_i \leq t)$ is the empirical cdf.

(4) Find the VC dimension of all rectangles in $\mathbb{R}^d$.

(5) Download the control chart data from:

`http://archive.ics.uci.edu/ml/datasets/Synthetic+Control+Chart+Time+Series`

The data consist of $n = 600$ "control charts." Each control chart is a time series of length 60. You can think of these as 600 vectors each of length 60. These are synthetic data and there are actually 6 clusters each of size 100. Of course, you should not use knowledge of the true clusters in your analysis.

(a) Apply k-means clustering and hierarchical clustering and summarize your analysis. In particular, compare the clusters you found to the true clusters.

(b) In regression, we usually tried to reduce dimension by selecting a small set of important features. We can do the same thing for clustering. The idea is to choose a small set of the 60 columns and only use those for clustering. Your goal in this part of the question is to invent a feature selection method for clustering and apply it to the data.

(6) Generate $n = 200$ data points $X_1, \ldots, X_n$ in $\mathbb{R}^2$ as follows. First generate n Bernoulli random variables. $U_1, \ldots, U_n$. Next generate $Z_1, \ldots, Z_n$ as follows. If $U_i = 0$ take

$$Z_i(1) = -3, \quad Z_i(2) \sim \text{Uniform}(-1, 1)$$

and if $U_i = 1$ take

$$Z_i(1) \sim \text{Uniform}(2, 4), \quad Z_i(2) = 0.$$

Finally, set $X_i(s) = Z_i(s) + 0.1\,\epsilon_i(s)$, $s = 1, 2$ where $\epsilon_i(s) \sim N(0, 1)$. Plot the data. Write an R program to do k-lines clustering and apply it to your simulated data.