Statistics 745

Assignment 1, Due: October 1st

Assignments are to be written in Latex or another word processing software package, and are to be professionally prepared to receive credit. Each student is required to do there own assignment.

1. Let $K_\lambda(x_i, x_j)$ be a kernel function. Define the matrix,

$$W_{ij} = K_\lambda(x_i, x_j).$$

Also, define the combinatorial Laplacian operator as $\Delta = D - W$ where D_{ij} =diag$(W \times \mathbf{1})$ (the diagonal row sum matrix). For example, if W was set as

$$W = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & 1 \end{pmatrix} \quad D = \begin{pmatrix} 1\frac{1}{2} & 0 & 0 \\ 0 & 1\frac{1}{4} & 0 \\ 0 & 0 & 1\frac{3}{4} \end{pmatrix} \quad \Delta = \begin{pmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

   (a) Show that the solution to:

$$\min_f (y - f)^T W (y - f) + f^T \Delta f \tag{1}$$

   is equivalent to local regression of an intercept.

   (b) This result useful in that it can apply to observed graphs. Consider the cora text data (from the first group assignment). Let $W = A$ (the adjacency matrix for the network) and apply the result from (1) to fit a function f over the text network. Compute the training accuracy with response as the paper's topic. Compare this to linear regression for the title data only. Determine, which result yields better performance.

2. Derivation of smoothing splines (page 183). Suppose that $N \geq 2$ and that g is the natural cubic spline interplant to the pairs $\{x_i, z_i\}_{i=1}^n$ with $a < x_1 < \cdots < x_N < b$. This is a natural spline with a knot at every x_i; being an N-dimensional space of functions, we can determine the coefficients such that it interpolates the sequence z_i exactly. Let $\widetilde{g}$ be any differentiable function on $[a, b]$ that interpolates the N paris.

(a) Let $h(x) = \widetilde{g}(x) - g(x)$. Use integration by parts and the fact that g is a natural cubic spline to show that:

$$\int_a^b g''(x)h''(x)dx = -\sum_{j=1}^{N-1} g'''(x_j^+)\{h(x_{j+1}) - h(x_j)\} = 0$$

(b) Hence show that

$$\int_a^b \widetilde{g}''(t)^2 dt \geq \int_a^b g''(t)^2 dt$$

and that the equality can only hold if h is zero in $[a, b]$.

(c) Consider the penalized least squares problem

$$\min_f \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int_a^b f''(t)^2 dt$$

Use (b) to argue that the minimizer must be a cubic spline with knots at each point x_i.

3. Smoothing Spline characteristics:

(a) Determine the form of the smoother S_λ such that $\hat{f} = S_\lambda y$ minimizes:

$$\min_f (y - f)^T (y - f) + \lambda f^T R f. \tag{2}$$

(b) Define $\mathbf{N}$ as the $N \times N$ natural spline function with each knot chosen at each data point. It is known that $\hat{f}$ is the solution to:

$$\min_\theta (y - \mathbf{N}\theta)^T (y - \mathbf{N}\theta) + \lambda \theta^T \Omega_{\mathbf{N}} \theta$$

From this, we established that $\hat{f} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\Omega_{\mathbf{N}})^{-1}\mathbf{N}^T y$. Identify the matrix R independent of λ so that S_λ can be written as the minimizer to (2).