Statistics 745

Assignment 2

1. Let X have two variables, $x_1,\ x_2$. Define

$$W_{1_{ij}} = K_\lambda(x_{1_i}, x_{1_j}) \qquad W_{2_{ij}} = K_\lambda(x_{2_i}, x_{2_j})$$

From this let $\Delta_1,\ \Delta_2$ be the corresponding combinatorial Laplacian matrices. We wish to fit the following model:

$$\min_{\alpha, f_1, f_2} \sum_{i=1}^{n} \log\left(1 + e^{-(2y_i-1)\eta_i}\right) + f_1^T \Delta_1 f_1 + f_2^T \Delta_2 f_2$$

with $\eta = \alpha + f_1 + f$. Derive the z-scoring algorithm for this, make sure it has the proper form (similar to logistic linear regression). Using the simulated data, fit the z-scoring algorithm directly. This algorithm is called *local scoring*.

2. Implement the following AdaBoost algorithm.

   (a) Initialize observations weights $w_i = 1/n$, $i = 1, \cdots, n$.

   (b) For i in 1 to M repeat.

      i. Fit a classifier $h_m(x)$.

      ii. Compute

$$err_m = \frac{\sum_{i=1}^{n} w_i I(y_i \neq h_m(x_i))}{\sum_i w_i}$$

      iii. Compute $\beta_m = \log\left(\frac{1-err_m}{err_m}\right)$

      iv. Set $w_i = w_i \exp(\beta_m I(y_i \neq h_m(x_i)))$.

   (c) Output $f(x) = \text{sign}\left(\sum_{m=1}^{M} \beta_m h_m(x)\right)$.

   Let $X_1, \cdots, X_{10}$ be standard independent Gaussian and define the response as:

$$Y = \begin{cases} 1 & \sum_{j=1}^{10} X_j^2 > 9.34 \\ -1 & o.w. \end{cases}$$

   Using stumps fit this data with 2000 cases (1000 for each class) and a testing set of 5000 observations. Compute the iteration by error plot for the first 400 iterations of AdaBoost. Compare the performance to a 6 split tree.