**Intelligent Data Analysis**

**Problem Set 7** **Due 6/6/2016**

1. Answer the following questions.
   (a) For a two-class problem, show that the Gini index and entropy are maximized at $p$=0.5. See Chapter 12, pp 10 for details.
   (b) Suppose a certain node of a tree contains the following data points: 2 $A$'s, 3 $B$'s, and 5 $C$'s. Use the Gini index to compute the impurity of the node. Suppose this node is further split into two nodes. What is the best increase in purity you can get? Show the impurity gain. See Chapter 12, pp 11.

2. This problem is to prove the learning rule in Chapter 13, pp 25. Consider the following error function

$$E(w_0 \mid \chi) = \frac{1}{2}\sum_{t=1}^{N}(r^t - y^t)^2 ,$$

where $y^t = \dfrac{1}{1+\exp(-(\mathbf{w}^T\mathbf{x}+w_0))}$. Show that

$$\Delta\mathbf{w} = -\eta\frac{\partial E}{\partial \mathbf{w}} = \eta\sum_{t=1}^{N}(r^t - y^t)y^t(1-y^t)\mathbf{x}^t$$

$$\Delta w_0 = -\eta\frac{\partial E}{\partial w_0} = \eta\sum_{t=1}^{N}(r^t - y^t)y^t(1-y^t).$$

3. Local linear regression uses a linear function near query point $\mathbf{x}$ of the form $g(\mathbf{x}) = w_0 + \mathbf{w}^T\mathbf{x}$. One may further simplify the local linear function by a constant near point $\mathbf{x}$ as $g(\mathbf{x}) = w_0$. For query point $\mathbf{x}$, the problem is to find $w_0$ that minimizes

$$E(w_0) = \sum_{t=1}^{N} K_\lambda\left(\mathbf{x},\mathbf{x}^t\right)\left[r^t - w_0\right]^2 ,$$

where $K_\lambda\left(\mathbf{x},\mathbf{x}^t\right)$ is the Epanechnikov quadratic kernel, with parameter $\lambda$,

defined by $K_\lambda\left(\mathbf{x},\mathbf{x}^t\right) = D\left(\dfrac{\left\|\mathbf{x}-\mathbf{x}^t\right\|}{\lambda}\right)$ with $D(u) = \begin{cases}\frac{3}{4}(1-u^2) & \text{if } |u|\le 1\\ 0 & \text{otherwise}\end{cases}$.

   (a) Find an optimal estimate $g(\mathbf{x}) = w_0$ that minimizes the above performance index.
   (b) Under what condition the optimal estimate $g(\mathbf{x})$ cannot be obtained, i.e., the model fails in computing the estimate?

4. Program Assignment In this assignment, you are asked to do classification using Fisher's iris data set, which contains 50 specimens from each of three different species of iris—setosa, versicolor, and virginica—on the following dimensions: sepal length, sepal width, petal length, petal width. Use the first 100 cases as the training and the rest 50 cases as the testing data.

   You are asked to implement the logistic discriminant method (see Chapter 13, pp 19-21) with the training data. The logistic discriminant uses all the four input variables and returns the predicted class label. The softmax functions are used to represent posterior probabilities. Initialize the mode parameters with small random numbers. Apply stochastic gradient descent (see Chapter 10, pp 8) to update the model parameters. A complete pass over all the data points in the training set is called an epoch. Plot the curve of the cross entropy (the error function, see Chapter 13, pp 20) verse the epoch during the training process. Try different values of the learning rate $\eta$ if necessary. Apply the 50 testing cases to the classifier, report their classification matrix and compute the accuracy.