B365 Homework 5

Read Ch. 4 (Classification) of *Principles of Data Mining*

1. Suppose we have the two-dimensional points (1,1), (2,2), (3,1), (4,2). Write $R(1,1)$ for the region in 2-space that is closer to the point $(1,1)$ than any of the other points, with similar definitions for the other three points.

   (a) Assume we use Euclidean distance as our distance measure. Sketch the regions $R(1,1), R(2,2), R(3,1), R(4,2)$ clearly drawing the boundaries between the regions.

   (b) Recall that the Manhattan distance between two points $x, y \in \Re^2$ (points in the plane) is given by

   $$d(x,y) = |x_1 - y_1| + |x_2 - y_2|$$

   Sketch the analogous regions using Manhattan distance.

2. Consider the following way of generating labeled triples with two possible classes. We generate (x,y,z) triples using "runif" for each coordinate, setting the associated class variable to 1 if $z^2 < x^2 + y^2$, and 0 otherwise.

   (a) In R generate 300 class-labeled (x,y,z,c) instances according to the above model.

   (b) Using the "Leave-one-out" scheme discussed in class, estimate the error-rate of the nearest neighbor classifier, based on Euclidean distance.

   (c) Generate 10,000 new (x,y,z,c) points according to the same model, classifying each as you did in the previous part. That is, use the first 300 instances to classify the new labeled points. Compute the error rate on these new points. The two error rates should be about the same. That is, the estimate of the "generalization error" we get from part b) should be accurate.

3. This problem deals with the "Ketchup"' data that can be found on our Canvas site. Each instance in this data set gives the price of 4 different kinds of ketchup: Heinz, Hunts, Del Monte, and a generic brand. In each instance the consumer makes a choice of which kind to purchase based on these prices and other experience. Your classifier must be based only on these 4 variables.

   (a) Use a nearest neighbor classifier to estimate the choice of ketchup each consumer will make, using the "leave one out" scheme.

   (b) What would you estimate your error would be for new data not seen in this data set?

   (c) Estimate the prior distribution over the four different classes (Heinz, Hunts Del Monte, and Generic).

   (d) Considering the performance of your classifier, does it appear that the price data and the consumer's purchase are independent?

   (e) Reasoning from what you know about purchase habits, provide a possible explanation of why the consumer's choice may be independent of the four given prices?

   (f) What would be the best possible classifier (in terms of error rate) if the consumer's choice is independent of the price data?

4. Consider the Fisher iris data, as plotted in the program "iris.r" studied in class.

   (a) Reasoning from this plot, choose a variable and a split point for the variable so that the resulting two regions have one that contains only Setosa flowers, while the ther mixes the other two classes.

   (b) For the "mixed" region resulting from the first part, choose a variable and split point that separates the Versicolor and Virginica flowers as well as possible.

   (c) Sketch the resulting three regions over the scatterplot of the two relevant variables, clearly labeling each region with the resulting class.