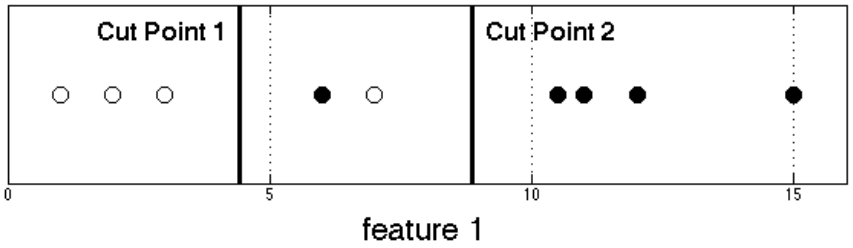
Data Mining Quiz 1

1 Problems

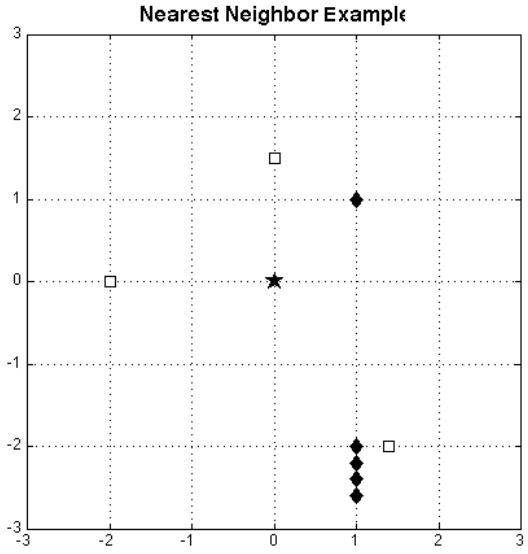
1. Suppose we have the data set with 10 instances, one feature, and one label shown below. The feature is given in the first column while the label $y \in \{1, 2\}$ is given in the second column. If we create new bootstrap samples of size $N = 10$ from this dataset, what is the probability that a bootstrap sample *does not contain* an instance with a label of '2'?

$$[X, y] = \begin{bmatrix} 7.9 & 1 \\ 3.4 & 1 \\ 1.0 & 1 \\ 3.1 & 1 \\ 5.4 & 1 \\ 3.8 & 1 \\ 1.0 & 1 \\ 1.0 & 1 \\ 1.0 & 1 \\ 10.0 & 2 \end{bmatrix} \tag{1}$$

2. Examine the labelled data shown in the figure. The circles represent data instances and the color (white, black) represents the class label. We want to grow a classification tree from this data. The solid dark lines in the figure represent two potential cutpoints. Calculate the node impurities that would result for each choice of cutpoint. For measures of node impurity use both node error $Q_{err} = 1 - p(k)$ (k index of most represented class in node) and the Gini Diversity Index $Q_{gdi} = 1 - \sum_i p^2(i)$. Which cutpoint would you choose if you wanted to minimize the overall impurity with the objective $J = \sum_m N_m Q(m)$, where N_m is the number of instances in child node m? Would the answer change if you did not weight by the number of instances in the child nodes?



3. The figure below (Nearest Neighbor Example) shows a typical binary classification problem. The diamonds represent one class while the squares represent another class. We are attempting to classify the origin (five-pointed star) using the nearest neighbor method. Determine the classification with k-nearest neighbor using both cityblock and euclidean distance measures and $k \in (1, 3, 5)$ (in total there are 6 possibilities). Also answer the following: how many *features* are shown in this dataset? (Recall that the cityblock distance between two-vectors in $\mathbb{R}^n$ is $d(w, v) = \sum_i^n |w_i - v_i|$)



The same data in matrix form is given below (data X and label y).

$$X = \begin{bmatrix} 1 & 1 \\ 0 & 1.5 \\ -2 & 0 \\ 1 & -2 \\ 1 & -2.2 \\ 1 & -2.4 \\ 1 & -2.6 \\ 1.4 & -2.0 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \end{bmatrix} \tag{2}$$

4. Using the same data as in problem 3 we attempt to perform a *linear discriminant analysis*. We compute in-class scatter matrices for each class. From the matrices below find the corresponding in-class covariance matrices and find which of these matrices definitely *cannot* be a covariance matrix (why?).

$$S_a = \begin{bmatrix} 0 & 0 \\ 0 & 2.228 \end{bmatrix}, S_b = \begin{bmatrix} 2.92 & -1.45 \\ -1.45 & 3.0833 \end{bmatrix}, S_c = \begin{bmatrix} 0 & 2.228 \\ 0 & 0 \end{bmatrix} \tag{3}$$

   Once we know the in-class scatter matrices how do we proceed (symbolically) with linear discriminant analysis? How many separate coefficients does our linear predictor consist of?

5. Explain the distinction between *supervised* and *unsupervised* machine learning problems and classify the following algorithms as one or the other: clustering, regression, k-nearest neighbor, k-means, linear discriminant analysis, principal component analysis, classification and decision trees.