# ST510: Foundations of Machine Learning
# Exercise Sheet

1. This problem is about the construction of classification trees. Each split is typically made in order to minimise the cost function

$$C(T) = \sum_{m=1}^{|T|} N_m Q_m,$$

where $T$ is the tree object, $|T|$ is total number of terminal nodes, $N_m$ is the number of training data points in the region corresponding to the $m$-th terminal node, and $Q_m$ is the node impurity measure of the same region. In class, we have introduced two common measures including "misclassification error" and "Gini index". Consider a two-class classification problem with two inputs and a training dataset illustrated in Figure 1. There are infinitely many possible splits we could make, but we restrict ourselves to split at integer values in the region where the training data is located.

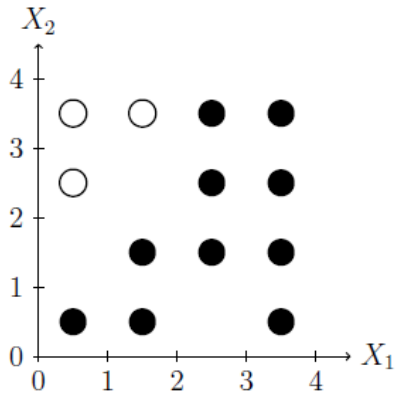


Figure 1: Training dataset with two inputs $X_1, X_2$ and two classes for $Y$ : "solid circle" and "hollow circle".

   (a) Give an optimal classification tree with two split points for minimizing the training misclassification loss.

   (b) If we use the misclassification error as the node impurity measure, what is the resulting classification tree with two split points obtained by recursive binary splitting? Explain the suboptimality of the solution.

(c) Consider the first split point. If we use the Gini index as the node impurity measure, compute the cost $C(T)$ for the following two candidate splits

i. $R_1 = \{X : X_1 \leq 1\}$ and $R_2 = \{X : X_1 > 1\}$,

ii. $R_1 = \{X : X_1 \leq 2\}$ and $R_2 = \{X : X_1 > 2\}$.

Among the above two candidate splits in i. and ii., where would we make the first split?