# ECE 364: Assignment #1

1. (10 pts) The table below shows a training dataset containing four data instances. The two descriptive features, $d[1]$ and $d[2]$, and the target feature, $t$, are all binary.

| ID | $d[1]$ | $d[2]$ | $t$ |
|---|---|---|---|
| 1 | 0 | 0 | 1 |
| 2 | 0 | 1 | 1 |
| 3 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 |

   (a) (2 pts) What is the entropy of $t$, i.e., $H(t)$?

   (b) (2 pts) What is the information gain for $d[1]$?

   (c) (3 pts) Using the above training dataset, suppose a decision tree is built that contains only three nodes: root node $d[1]$ and its two children leaf nodes. Each leaf node is assigned the majority class of its associated set of data instances (breaking ties in favor of $t = 0$). What is the classification accuracy of this decision tree on the training dataset, i.e., what fraction of the data instances it classifies correctly?

   (d) (3 pts) Given a binary descriptive feature, $A$, that splits a set of data instances, $E$, into two nonempty subsets $E_1$ and $E_2$ such that $E_1$ has $p_1 > 0$ positive instances and $n_1 > 0$ negative instances for a binary target feature, $C$, and $E_2$ has $p_2 > 0$ positive instances and $n_2 > 0$ negative instances, is it possible for information gain for $A$ to be 0? Explain.

2. (10 pts) Suppose a dataset of size $n$ has been given to us. We induce a decision tree with it. It has a total error of $\epsilon$ in the set of predictions it makes for the instances in the dataset. A weighted dataset is then created for boosting. Prove that the sum of all the weights in this weighted dataset is 1.0.

3. (10 pts) Suppose we want to use bagging to generate a model. Our initial dataset contains 10 data instances, with ID: 1, 2, $\cdots$, 10. Let us break the dataset into subsets: $A$ with instances 1, 2, 3, 4, 5, 6 and $B$ with instances 7, 8, 9, 10. In order to generate a bootstrap sample, suppose each member of $A$ should be picked with twice the probability of each member of $B$. What is the probability that data instance with ID=2 will be included in the bootstrap sample?

4. (10 pts) Fig. 1 shows the decision tree for the task of predicting whether customers will buy the newly released smartphone BlueBerry $X$. There are two binary descriptive features: whether the customer thinks the phone's appearance is good and whether the price is affordable. The target binary feature is whether the customer will buy the phone.

   Below is a validation set collected from BlueBerry Princeton store through a customer survey. Using this validation set, apply **reduced error pruning** to the decision tree. For each iteration of the algorithm, assume that the algorithm is applied in a bottom-up, left-to-right fashion. For each iteration, explain why the algorithm chooses to prune the subtrees or not. Draw the final tree after pruning.

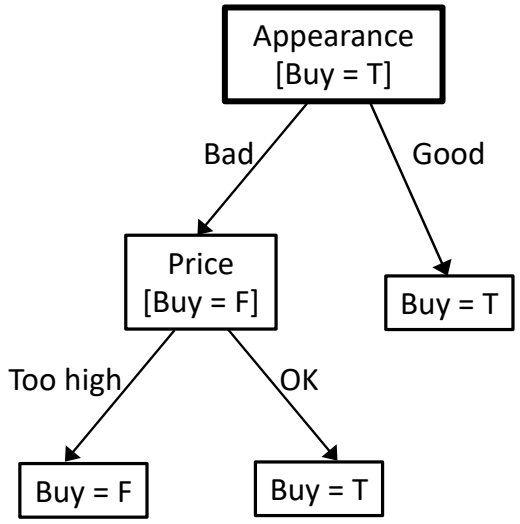


Figure 1: BlueBerry $X$ customer prediction (T: True; F: False).

| Customer ID | Appearance | Price | Buy |
|---|---|---|---|
| 1 | Good | OK | T |
| 2 | Bad | Too high | F |
| 3 | Good | Too high | T |
| 4 | Bad | OK | F |
| 5 | Bad | OK | F |

5. (20 pts) **Coding project**

   In this project, you will train classifiers based on decision trees to identify types of glass. The dataset consists of 214 data instances of glass with the following nine numerical descriptive features per instance: refractive index, sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron.

   The target class of each glass instance is one of six glass types: float processed building window, non-float processed building window, float processed vehicle window, container, tableware, and headlamp.

   The data will be divided into a training set and a validation set. You will start with training the default decision tree classifier in the Scikit-learn library and then train a decision tree classifier that is pre-pruned based on depth. You will then explore post-pruning. You will also train an ensemble of decision tree classifiers using bagging and boosting.

   Once you have built these classifiers, you will evaluate them to find the one with the best performance.

   Please see the Jupyter notebook for more details.

   GitHub repository for ECE364 coding projects: https://github.com/JHA-Lab/ece364_2024