ELE 364: Assignment #1

1. (10 pts) Consider a training dataset of size $a + b$ whose target feature has two levels: true and false; true for a data instances and false for b data instances. Prove that the entropy of the dataset with respect to the target feature is given by: $\frac{1}{a+b} \log_2 \frac{(a+b)^{a+b}}{a^a b^b}$ if $a \neq 0$ and $b \neq 0$.

2. (10 pts) Consider the following dataset:

ID	AGE	EDUCATION	OCCUPATION	INCOME
1	28	MS	teacher	25-50
2	30	BS	professional	75-100
3	42	PhD	professional	100-150
4	28	BS	farmer	75-100
5	32	BS	teacher	25-50
6	64	PhD	professional	100-150
7	50	PhD	professional	75-100
8	37	MS	farmer	25-50
9	42	BS	teacher	75-100
10	70	MS	professional	100-150

   (a) Calculate the entropy and Gini index of this dataset using the income target feature.

   (b) Find the threshold value that would maximize the information gain for splitting the data based on the AGE feature (when building a decision tree).

   (c) Calculate the information gain for the education and occupation features.

   (d) Calculate the information gain ratio for the education and occupation features using entropy.

   (e) Calculate information gain for the education and occupation features using the Gini index.

3. (10 pts) Compute the probability of a model ensemble, which uses simple majority voting, making a correct prediction in the following scenario: the ensemble contains five independent models, all of which have an error rate of 0.2.

4. (10 pts) We would like to create an ensemble based on boosted decision trees. A training dataset with 10 data instances is given. It is used to obtain the first decision tree of the ensemble. This decision tree has a 10% error in its predictions made on the instances in the training dataset.

   - What are the new weights of the instances misclassified by this decision tree?
   - What are the new weights of the instances correctly classified by this decision tree?
   - What is the confidence factor for this decision tree?

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

   For this project, you will train classifiers based on decision trees to determine whether a patient has heart disease.

   The dataset consists of the following set of descriptive features:

(a) numeric: age, resting blood pressure, serum cholesterol, max. heart rate achieved, level of exercise-induced ST segment depression, number of major vessels colored by flouroscopy.

(b) categorical: sex, chest pain level, whether the fasting blood sugar > 120 (mg/dl), resting ECG type, whether the patient suffers from exercise-induced angina, slope type of the peak exercise ST segment, type of thalassemia.

We will only use the numerical descriptive features due to implementation constraints.

The target feature is a binary variable, where 1 indicates the presence of heart disease.

You will train decision trees using different impurity measures and also pre-prune trees based on depth. You will also train ensemble classifiers using bagging and boosting.

See the Jupyter notebook for more details.