Machine Learning & Data Mining
Homework 3

Caltech CS/CNS/EE 155
January 19^{th}, 2016

This set is due 9pm, January 27^{th}, via Moodle. You are free to collaborate on all of the problems, subject to the collaboration policy stated in the syllabus.

1 Decision Trees

Question A: Consider the following data. "Package Type", "Unit Price >$5" and "Contains >5 grams of fat " are input variables and "Healthy?" is the variable you want to predict.

No.	Package Type	Unit Price > $5	Contains > 5 grams of fat	Healthy?
1	Canned	Yes	Yes	No
2	Bagged	Yes	No	Yes
3	Bagged	No	Yes	Yes
4	Canned	No	No	Yes

i. Train a decision tree by hand using top-down greedy induction. Use *entropy* as the impurity measure. Since the data can be classified without error, the stopping criterion will be no impurity in the leaves.

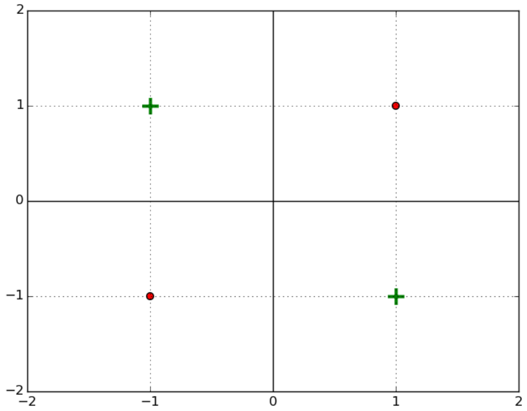
Submit a drawing of your tree showing the impurity reduction yielded by each split (including root) in your decision tree.

ii. Train another decision tree on these data, as in **i**, this time using the *Gini index* as your impurity measure.

Submit a drawing of your tree showing the impurity reduction yielded by each split (including root) in your decision tree.

Question B: Compared to a linear classifier, is a decision tree always preferred for classification problems? If not, draw a simple 2-D dataset that can be perfectly classified by a simple linear classifier but which requires an overly complex decision tree to perfectly classify.

Question C: Consider the following 2D data set:



i. Suppose we train a decision tree on this dataset using top-down greedy induction, with the Gini index as the impurity measure. We define our stopping condition to be if no split of a node results in any reduction in impurity. Submit a drawing of the resulting tree. What is its classification error ((number of misclassified points) / (number of total points))?

ii. Submit a drawing of a two-level decision tree that classifies the above dataset with zero classification error. (You don't need to use any particular training algorithm to produce the tree.)

Is there any impurity measure (i.e. any function that maps the data points under a particular node in a tree to a real number) that would have led top-down greedy induction to produce the tree you drew? If so, give an example of one, and briefly describe its pros and cons as an impurity measure for training decision trees in general (on arbitrary datasets).

iii. Suppose there are 100 data points in some 2-D dataset. What is the largest number of unique thresholds (i.e., internal nodes) you might need in order to achieve zero classification training error (on the training set)? Please justify your answer.

Question D: Suppose in top-down greedy induction we want to split a leaf node that contains N data points composed of D continuous features. What is the worst-case complexity (big-O in terms of N and D) of the number of possible splits we must consider in order to find the one that most reduces impurity? Please justify your answer.

Note: Recall that at each node-splitting step in training a DT, you must consider all possible splits that you can make. While there are an infinite number of possible decision boundaries since we are using continuous features, there are not an infinite number of boundaries that result in unique child sets (which is what we mean by "split").

2 Overfitting Decision Trees

In this problem, you will use the Breast Cancer Wisconsin (Diagnostic) Data Set. Please download files wdbc.data and wdbc.names from the link below:

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
The file wdbc.names gives a detailed explanation of the dataset, including what the features are. The file wdbc.data contains all the data you need.

In the following question, your goal is to predict the diagnosis, which is the second column in wdbc.data. Ignore the first column, which is an ID number. Use the first 400 rows as training data, and the last 169 rows as test data. Please feel free to use additional packages such as Scikit-Learn. Include your code in your submission.

Question A: Train a decision tree classifier using Gini as the impurity measure and minimal leaf node size as early stopping criterion. Try different node sizes from 1 to 25 in increments of 1. Then, on a single plot, plot both training and test error versus leaf node size.

Question B: Train a decision tree classifier using Gini as the impurity measure and maximal tree depth as early stopping criterion. Try different tree depth from 2 to 20 in increments of 1. Then, on a single plot, plot both training and test error versus tree depth.

Question C: What effects does early stopping have on the performance of decision tree model? Please justify your answer based on the two plots you derived.

3 The AdaBoost Algorithm

In this problem, you will show that the choice of the α_t parameter in the AdaBoost algorithm corresponds to greedily minimizing an exponential upper bound on the loss term at each iteration.

Question A: Let $h_t : \mathbb{R}^m \to \{-1, 1\}$ be the weak classifier obtained at step t, and let α_t be its weight. Recall that the final classifier is

$$H(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{i=1}^{T} \alpha_t h_t(x)\right).$$

Suppose $\{(x_1, y_1), ..., (x_N, y_N)\} \subset \mathbb{R}^m \times \{-1, 1\}$ is our training dataset. Show that the training set error of the final classifier can be bounded from above if an an exponential loss function is used:

$$E = \frac{1}{N}\sum_{i=1}^{N} \exp(-y_i f(x_i)) \geq \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(H(x_i) \neq y_i),$$

where $\mathbb{1}$ is the indicator function.

Question B: Show that

$$E = \prod_{t=1}^{T} Z_t,$$

where Z_t is the normalization factor for distribution D_{t+1}:

$$Z_t = \sum_{i=1}^{T} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$$

Hint: Express the data weights at each iteration in terms of the initial data weights, and then use the fact that the weights at iteration $t+1$ sum to 1.

Question C: The reason for deriving all this is that it is hard to directly minimize the training set error, but we can greedily minimize the upper bound E on this error. Show that choosing α_t and h_t greedily to minimize Z_t at each iteration leads to the choices in AdaBoost:

$$\alpha_t^* = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$$

where ϵ_t is the training set error of weak classifier h_t for weighted dataset:

$$\epsilon_t = \sum_{i=1}^{N} D_t(i) \mathbb{1}(h_t(x_i) \neq y_i)$$

Hint: Show that the normalizer Z_t can be written as

$$Z_t = (1-\epsilon_t)\exp(-\alpha_t) + \epsilon_t \exp(\alpha_t),$$

and minimize Z_t with respect to α_t and then h_t.