Machine Learning & Data Mining
Homework 3

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

This set is due 5pm, January 26^{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

This question will give them some data and ask them to go through the decision tree training logic by hand and show all the steps. We will tell them which stopping and splitting criteria to use.

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

Train a decision tree by hand using top-down greedy induction. Your splitting criterion should be *information gain* and since the data can be classified without error, the stopping criterion will be no impurity in the leaves.

i. Calculate the entropy at every level (including root) of your decision tree.

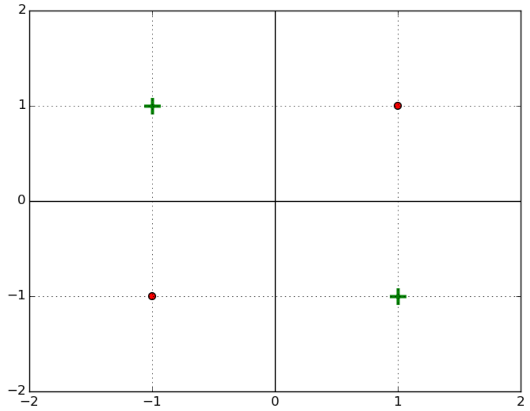
ii. Calculate the information gain at every level (including root) of each split in your decision tree.

iii. Draw your tree.

iv. Now using the same data, train a tree using the Gini index as your splitting criterion, and the same stopping criterion (no impurity in leaves). Show your calculations and draw the tree.

Question B: In this question, we compare decision trees to linear classifiers. Compared to a linear classifier, is the decision tree always preferred for classification problems? If not, could you please give a simple 2-D example (i.e., draw some training points) when the linear classifier can classify the data easily while the the required decision tree is overly complex?

Question C: Consider the following 2D data set.



i. Suppose we train a decision tree top-down using the Gini Index as the impurity measure. We define our stopping condition if no split of a node results in any reduction in impurity. What does the resulting tree look like, and what is the classification error?

ii. Suppose we instead define classification error as the impurity measure, and we stop when no split results in a reduction in classification error. What does the resulting tree look like, and what is the classification error?

iii. Suppose there are 100 data points instead of 4. Without worrying about overfitting, how many unique thresholds (i.e., internal nodes) do you need in the worst case in order to achieve zero classification training error? Please justify your answer.

Question D : Suppose we want to split a leaf node that contains N data points using D features/attributes (all of which are continuous). What is the worst-case complexity (big-O) of the number of possible splits we must consider (with respect to N and D)? Please justify your answer.

2 Overfitting Decision Trees

In this part of the problem, you will use the dataset 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/ wdbc.names gives the detailed explanation of the usage and the attributes information. wdbc.data contains all the data you need (ignore first column which is just 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 in Python or Weka2 in Java. Please attach the code to your submission (i.e., how you called Scikit-Learn or Weka).

Question A : Train your decision tree model using Gini impurity as metric 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 your decision tree model using Gini impurity as metric 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 uppper bound on the loss term at each iteration.

Question A: Let $h_t(x)$ be the weak classifier obtained at step t, and let α_t be its weight. Recall that the final classifier is

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

Show that the training set error of the final classifier can be bounded from above by an exponential loss function E

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

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

Question B: Show that

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

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

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

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: It is hard to directly minimize the training set error. Instead, let us try to 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:

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

Hint: Consider a special class of weak classifiers $h_t(x)$, that return exactly $+1$, if h_t classifies example x as positive, and -1 if h_t classifies x as negative. Then show that for this class of classifiers the normalizer Z_t can be written as

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

Now minimize Z_t with respect to α_t and then h_t.