Machine Learning & Data Mining — Caltech CS/CNS/EE 155
Homework 3 — February 3rd, 2015

Overview

- This homework is due February 17th, 2015 at 2pm via Moodle.
- This homework is intended to take about 6-8 hours to complete.

1 Decision tree basic questions [Shenghan Yao]

1. (*13 points*) Considering the information about 4 different foods below:
The "Package Type", "Unit Price $>$ \$5" and "Contain>5 grams of fat" are input variables and the "Healthy?" is the variable you want to predict.

No.	Package type	Unit price $>$ \$5	Contain $>$ 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

Please train a depth-2 decision tree using top-down greedy induction by hand. Please use information gain as splitting criteria. Since the data can be classified with no error, the stopping condition is when the leave nodes have no impurity.

   (a) (*3 points*) Please calculate the entropy at each split point (as well as at the root).

   (b) (*3 points*) Calculate the information gain at each split.

   (c) (*2 points*) Draw the tree.

   (d) (*5 points*) Using the same data set above, train your decision tree using Gini index as splitting criteria. Since the data can be classified with no error, the stopping condition is when the leave nodes have no impurity. Show your calculation and draw the tree.

2. (*4 points*) 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?

3. (*9 points*) Consider following 2D dataset.

X_1	X_2	Sign
0	0	Negative
0	1	Positive
1	0	Positive
1	1	Negative

(a) (*3 points*) 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?

(b) (*3 points*) 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?

(c) (*3 points*) Suppose there are 100 data points instead of 4. Without worrying about over-fitting, 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?

4. (*4 points*) 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)?

2 Implementation of Decision Tree [Minfa Wang]

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[1] in Python or Weka[2] in Java. Please attach the code to your submission (i.e., how you called Scikit-Learn or Weka).

1. (*10 points*) 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.

2. (*10 points*) 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.

3. (*10 points*) What effects does early stopping have on the performance of decision tree model? Please justify your answer based on the two plots you derived. (Hint: this is a SHORT answer.)

[1] http://scikit-learn.org/stable/
[2] http://www.cs.waikato.ac.nz/ml/weka/

3 The AdaBoost algorithm [Masoud Farivar]

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. To review the notations and details of AdaBoost you may refer to this document.

(a) (*10 points*) An exponential upper bound
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

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

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

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

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

$$\mathbb{1}\left(H(x_i) \neq y_i\right) = \begin{cases} 1 & H(x_i) \neq y_i \\ 0 & H(x_i) = y_i \end{cases}$$

(b) (*10 points*) Show that

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

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

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

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.

(c) (*10 points*) Minimizing E
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:

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

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.