Machine Learning and Computational Statistics, Spring 2015
Homework 5: Trees and Ensemble Methods

Due: Wednesday, March 25, 2015, at 4pm (Submit via NYU Classes)

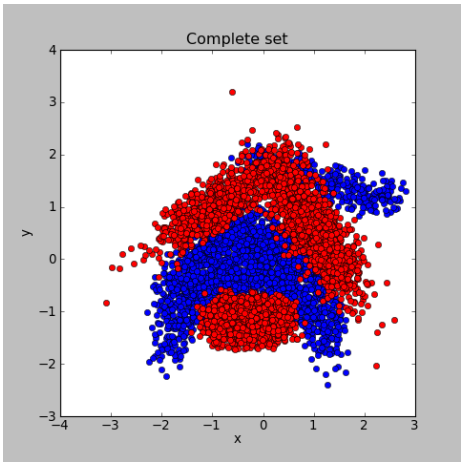
Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. You may include your code inline or submit it as a separate file. You may either scan hand-written work or, preferably, write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython).

1 Introduction

In this problem set, you will work with decision trees and ensemble methods. You will be using the decision tree implementation from `sklearn`, and implementing `AdaBoost` from scratch. You'll also work on some simple theoretical problems that highlight interesting properties of decision trees and ensemble methods.

2 Dataset description

You will be working with a simple two-feature binary dataset, known as the Banana dataset[1], which can be visualized as follows:



[1] `http://mldata.org/repository/data/viewslug/banana-ida/`

(Source: `http://adessowiki.fee.unicamp.br/adesso/wiki/courseIA368Q1S2012/eri_test_2/view/`)

The data consists of 5,300 instances, which have been split into 3,500 training points and 1,800 test points for this assignment. The csv files are included in the `data` directory. Each row corresponds to a data point - the first entry of the row gives the class label, and the next two entries give the values of the attributes.

3 Decision Trees

3.1 Building Trees by Hand[2]

In this problem we're going to be build a a small decision tree by hand for predicting whether or not a mushroom is poisonous. The training dataset is given below:

Poisonous	Size	Spots	Color
N	5	N	White
N	2	Y	White
N	2	N	Brown
N	3	Y	Brown
N	4	N	White
N	1	N	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

We're going to build a binary classification tree using the Gini index as the node impurity measure. The feature "Size" should be treated as numeric (i.e. we should find real-valued split points). For a given split, let R_1 and R_2 be the sets of data indices in each of the two regions of the split. Let $\hat{p}_1$ be the proportion of poisonous mushrooms in R_1, and let $\hat{p}_2$ be the proportion in R_2. Let N_1 and N_2 be the total number of training points in R_1 and R_2, respectively. Then the Gini index for the first region is $Q_1 = 2\hat{p}_1(1-\hat{p}_1)$ and $Q_2 = 2\hat{p}_2(1-\hat{p}_2)$ for the second region. When choosing our splitting variable and split point, we're looking to minimize the weighted impurity measure:

$$N_1 Q_1 + N_2 Q_2.$$

1. What is the first split for a binary classification tree on this data, using the Gini index? Work this out "by hand", and show your calculations. [Hint: This should only require calculating 6 weighted impurity measures.]

2. Compute the full decision tree by hand, building until all terminal nodes are either completely pure, or we cannot split any further.

[2]Based on Homework #4 from David Sontag's DS-GA 1003, Spring 2014.

3. Suppose we built the same type of tree described above (binary, Gini criterion, terminal nodes are either pure or cannot be split further) on the dataset given below. What would the training error be, given as a percentage? Why? [Hint: You can do this by inspection, without any significant calculations.]

Y	A	B	C
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	0
0	0	1	1
1	0	1	1
0	1	0	0
1	1	0	1
1	1	1	0
0	1	1	1
1	1	1	1

3.2 Investigating Impurity Measures[3]

1. Consider a data set with 400 data points from class C_1 and 400 data points from class C_2 . Suppose that a tree model A splits these into $(300, 100)$ at the first leaf node and $(100, 300)$ at the second leaf node, where (n, m) denotes that n points are assigned to C_1 and m points are assigned to C_2 . Similarly, suppose that a second tree model B splits them into $(200, 400)$ and $(200, 0)$. Show that the misclassification rates for the two trees are equal, but that the cross-entropy and Gini impurity measures are both lower for tree B than for tree A.

3.3 Trees on the Banana Dataset

The official `sklearn` documentation provides code that constructs a decision tree and visualizes the decision boundary on the "Iris[4] dataset" (`http://scikit-learn.org/stable/auto_examples/tree/plot_iris.html#example-tree-plot-iris-py`). Note that the `sklearn` implementation of decision trees is a bit different from that described in lecture: they just build to a certain depth, without a pruning step.

1. Modify the code referenced above to work on the Banana dataset. The default class labels are -1 and 1 in the given data files, but for the visualization code snippet to work, you will have to modify the class labels to 0 and 1. Note that the Iris dataset is a multiclass problem with 3 classes, while the Banana dataset is a binary dataset.

2. Run your code for different depths of decision trees, from 1 through 10, and briefly describe your observations of the decision surface visualization. [Use the default values for all other parameters.]

3. Plot the train and test errors as a function of the depth. Again, give a brief description of your observations.

[3] From Bishop's *Pattern Recognition and Machine Learning*, Problem 14.11
[4] `https://archive.ics.uci.edu/ml/datasets/Iris`

4. [Optional] Experiment with the other hyperparameters provided by `DecisionTreeClassifier` and find the combination giving the smallest test error. Summarize what you learn.

4 Bagging[5] [Optional Problem]

Consider a regression problem where we wish to learn function $y(\mathbf{x})$. Suppose we learn M functions $\hat{y}_1(\mathbf{x}), \ldots, \hat{y}_M(\mathbf{x})$. The predictions of each of these functions can be expressed as the sum of the true prediction plus an error term

$$\hat{y}_m(\mathbf{x}) = y(\mathbf{x}) + \epsilon_m(\mathbf{x})$$

The expected squared-error of the function is then given by $\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$. The average squared-error of the models acting individually is therefore

$$E_{av} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})^2]$$

Bagging involves constructing the final function as an average over the M functions:

$$\hat{y}_{bag}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m(\mathbf{x})$$

The error of bagging is therefore

$$\epsilon_{bag}(\mathbf{x}) = \hat{y}_{bag}(\mathbf{x}) - y(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \epsilon_m(\mathbf{x})$$

The expected squared-error is

$$E_{bag} = \mathbb{E}_{\mathbf{x}}[\epsilon_{bag}(\mathbf{x})^2]$$

1. [Optional] Assuming that the individual errors $\epsilon_m(\mathbf{x})$ have mean zero and are uncorrelated, that is, $\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})] = 0$ and $\mathbb{E}_{\mathbf{x}}[\epsilon_m(\mathbf{x})\epsilon_l(\mathbf{x})] = 0$ for $m \neq l$, show that

$$E_{bag} = \frac{1}{M} E_{av}$$

2. [Optional] In practice, however, the errors may be highly correlated. Nevertheless, using Jensen's inequality for the special case of the convex function $f(x) = x^2$, show that the average expected squared-error E_{av} of the individual functions and the expected error of bagging E_{bag} satisfy $E_{bag} \leq E_{av}$, without any assumptions on $\epsilon_m(\mathbf{x})$.

   Jensen's inequality can be stated as follows: For a convex function $f(x)$, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

This can easily be generalized to arbitrary convex functions $E(y)$ of the error.

[5] Based on a problem from Bishop's *Pattern Recognition and Machine Learning*.

5 AdaBoost

5.1 Implementation

In this problem, you will implement AdaBoost, one of the most popular techniques in ensemble methods.

1. Implement AdaBoost for the Banana dataset with decision trees of depth 3 as the weak classifiers (also known as "base classifiers"). Use the decision tree implementation from `sklearn` as in 3.3. The `fit` function of `DecisionTreeClassifier` has a parameter `sample_weight`, which you can use to weigh training examples differently during various rounds of AdaBoost.

2. [Optional] Visualize the AdaBoost training procedure for different numbers of rounds from 1 through 10. Plot the decision surface, and the training examples, such that training samples with larger weights in any round are represented as larger points compared to those with smaller weights. Provide a brief description of your observations.

3. Plot the train and test errors as a function of the number of rounds from 1 through 10. Again, give a brief description of your observations.

6 Gradient Boosting Machines

Recall the general gradient boosting algorithm[6], for a given loss function ℓ and a hypothesis space $\mathcal{F}$ of regression functions (i.e. functions mapping from the input space to $\mathbf{R}$):

1. Initialize $f_0(x) = 0$.

2. For $m = 1$ to M:

   (a) Compute:

   $$\mathbf{g}_m = \left(\frac{\partial}{\partial f(x_i)} \sum_{i=1}^{n} \ell \left\{ y_i, f(x_i) \right\} \Big|_{f(x_i)=f_{m-1}(x_i)} \right)_{i=1}^{n}$$

   (b) Fit regression model to $-\mathbf{g}_m$:

   $$h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left((-\mathbf{g}_m)_i - h(x_i) \right)^2 .$$

   (c) Choose fixed step size $\nu_m = \nu \in (0, 1]$, or take

   $$\nu_m = \arg\min_{\nu > 0} \sum_{i=1}^{n} \ell \left\{ y_i, f_{m-1}(x_i) + \nu h_m(x_i) \right\} .$$

[6] Besides the lecture slides, you can find an accessible discussion of this approach in `http://www.saedsayad.com/docs/gbm2.pdf`, in one of the original references `http://statweb.stanford.edu/~jhf/ftp/trebst.pdf`, and in this review paper `http://web.stanford.edu/~hastie/Papers/buehlmann.pdf`.

(d) Take the step:

$$f_m(x) = f_{m-1}(x) + \nu_m h_m(x)$$

3. Return f_M.

In this problem we'll derive two special cases of the general gradient boosting framework: L_2-Boosting and BinomialBoost.

1. Consider the regression framework, where $\mathcal{Y} = \mathbf{R}$. Suppose our loss function is given by

$$\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2,$$

and at the beginning of the m'th round of gradient boosting, we have the function $f_{m-1}(x)$. Show that the h_m chosen as the next basis function is given by

$$h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left[(y_i - f_{m-1}(x_i)) - h(x_i)\right]^2.$$

In other words, at each stage we find the weak prediction function $h_m \in \mathcal{F}$ that is the best fit to the residuals from the previous stage. [Hint: Once you understand what's going on, this is a pretty easy problem.]

2. Now let's consider the classification framework, where $\mathcal{Y} = \{-1, 1\}$. In lecture, we noted that AdaBoost corresponds to forward stagewise additive modeling with the exponential loss, and that the exponential loss not very robust to outliers (i.e. outliers can have a large effect on the final prediction function). Instead, let's consider instead the logistic loss

$$\ell(m) = \ln\left(1 + e^{-m}\right),$$

where $m = yf(x)$ is the margin. Similar to what we did in the L_2-Boosting question, write an expression for h_m as an argmin over $\mathcal{F}$.