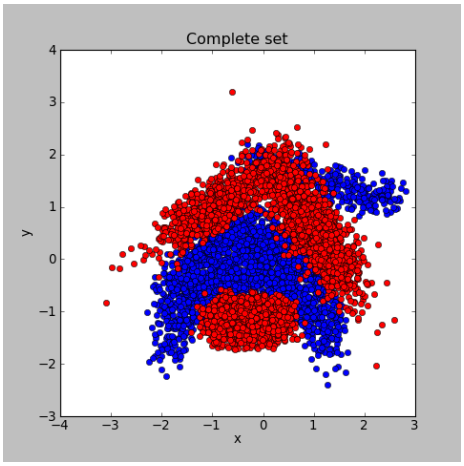
DS-GA 1003: Machine Learning and Computational Statistics
Homework 5: Trees and Boosting

Due: Monday, April 4, 2016, at 6pm (Submit via NYU Classes)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single file, either HTML or PDF. 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 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:



(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.

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

2 Decision Trees

2.1 Trees on the Banana Dataset

The official `sklearn` documentation provides code that constructs a decision tree and visualizes the decision boundary on the "Iris[2] 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.

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

3 AdaBoost

3.1 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 2.1. 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.

[2] `https://archive.ics.uci.edu/ml/datasets/Iris`

4 Gradient Boosting Machines

Recall the general gradient boosting algorithm[3], 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
   $$\nu_m = \arg\min_{\nu > 0} \sum_{i=1}^{n} \ell\left(y_i, f_{m-1}(x_i) + \nu h_m(x_i)\right) .$$
   (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
   $$\ell(\hat{y}, y) = \frac{1}{2} \left(\hat{y} - y\right)^2 ,$$
   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
   $$h_m = \arg\min_{h \in \mathcal{F}} \sum_{i=1}^{n} \left[\left(y_i - f_{m-1}(x_i)\right) - h(x_i) \right]^2 .$$
   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.]

[3] 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`.

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 $\mathcal{F}$.

5 From Margins to Conditional Probabilities[4]

Let's consider the classification setting, in which $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ are sampled i.i.d. from some unknown distribution. For a prediction function $f : \mathcal{X} \to \mathbf{R}$, we define the **margin** on an example (x, y) to be $m = yf(x)$. Since our class predictions are given by $\text{sign}(f(x))$, we see that a prediction is correct iff $m(x) > 0$. We have said we can interpret the magnitude of the margin $|m(x)|$ as a measure of confidence. However, it is not clear what the "units" of the margin are, so it is hard to interpret the magnitudes beyond saying one prediction is more or less confident than another. In this problem, we investigate how we can translate the margin into a conditional probability, which is much easier to interpret. In other words, we are looking for a mapping $m(x) \mapsto p(y = 1 \mid x)$.

In this problem we will consider margin losses. A loss function is a **margin loss** if it can be written in terms of the margin $m = yf(x)$. We are interested in how we can go from an empirical risk minimizer of a margin loss, $\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell\left(y_i f(x_i)\right)$, to a conditional probability estimator $\hat{\pi}(x) \approx p(y = 1 \mid x)$. Our approach will be to find the mapping in terms of the Bayes prediction function[5], and then apply the same mapping to the empirical risk minimizer. While there is plenty that can go wrong with this "plug-in" approach (primarily, the empirical risk minimizer from a hypothesis space $\mathcal{F}$ may be a poor estimate for the Bayes prediction function), it is at least well-motivated, and it can work well in practice. And please note that we can do better than just hoping for success: if you have enough validation data, you can directly assess how well "calibrated" the predicted probabilities are. This blog post has some discussion of calibration plots: https://jmetzen.github.io/2015-04-14/calibration.html.

It turns out it is straightforward to find the Bayes prediction function f^* for margin losses, at least in terms of the data-generating distribution: For any given $x \in \mathcal{X}$, we'll find the best possible prediction $\hat{y}$. This will be the $\hat{y}$ that minimizes

$$\mathbb{E}_y\left[\ell\left(y\hat{y}\right) \mid x\right].$$

If we can calculate this $\hat{y}$ for all $x \in \mathcal{X}$, then we will have determined $f^*(x)$. We will simply take

$$f^*(x) = \arg\min_{\hat{y}} \mathbb{E}_y\left[\ell\left(y\hat{y}\right) \mid x\right].$$

[4] This problem is based on Section 7.5.3 of Schapire and Freund's book *Boosting: Foundations and Algorithms*.

[5] In this context, the Bayes prediction function is often referred to as the "population minimizer." The term "population" makes most sense in a context where we are using a sample to approximate some statistic of an entire population. In our case, "population" referes to the fact that we are minimizing with respect to the true distribution, rather than a sample.

It may be intuitively obvious that this f^* is the risk minimizer. Here's a mathematical proof:

$$\begin{aligned}
\min_f \mathbb{E}_{x,y}\ell\left(yf(x)\right) &= \min_f \mathbb{E}_x\left[\mathbb{E}_y\left[\ell\left(yf(x)\right) \mid x\right]\right] \\
&\geq \mathbb{E}_x\left[\min_{\hat{y}} \mathbb{E}_y\left[\ell\left(y\hat{y}\right) \mid x\right]\right] \\
&= \mathbb{E}_x\left[\mathbb{E}_y\left[\ell\left(yf^*(x)\right) \mid x\right]\right] \\
&= \mathbb{E}_{x,y}\ell\left(yf^*(x)\right).
\end{aligned}$$

But of course we must also have $\min_f \mathbb{E}_{x,y}\ell\left(yf(x)\right) \leq \mathbb{E}_{x,y}\ell\left(yf^*(x)\right)$. So the inequality must actually be an equality, and thus the minimum is attained at f^*.

Below we'll calculate f^* for several loss functions. It will be convenient to let $\pi(x) = \mathbb{P}\left(y = 1 \mid x\right)$ in the work below.

1. Write $\mathbb{E}_y\left[\ell\left(yf(x)\right) \mid x\right]$ in terms of $\pi(x)$ and $\ell\left(f(x)\right)$. [Hint: Use the fact that $y \in \{-1, 1\}$.]

2. Show that the Bayes prediction function $f^*(x)$ for the exponential loss function $\ell\left(y, f(x)\right) = e^{-yf(x)}$ is given by

   $$f^*(x) = \frac{1}{2}\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

   and, given the Bayes prediction function f^*, we can recover the conditional probabilities by

   $$\pi(x) = \frac{1}{1+e^{-2f^*(x)}}.$$

   [Hint: Differentiate the expression in the previous problem with respect to $f(x)$. If this is confusing, you may find it more comforting to change variables a bit: Fix an $x \in \mathcal{X}$. Then write $p = \pi(x)$ and $\hat{y} = f(x)$. After substituting these into the expression you had for the previous problem, you'll want to find $\hat{y}$ that minimizes the expression. Use differential calculus. Once you've done it for a single x, it's easy to write the solution as a function of x.]

3. Show that the Bayes prediction function $f^*(x)$ for the logistic loss function $\ell\left(y, f(x)\right) = \ln\left(1 + e^{-yf(x)}\right)$ is given by

   $$f^*(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

   and the conditional probabilities are given by

   $$\pi(x) = \frac{1}{1+e^{-f^*(x)}}.$$

   Again, we may assume that $\pi(x) \in (0, 1)$.

4. [Optional] Show that the Bayes prediction function $f^*(x)$ for the hinge loss function $\ell\left(y, f(x)\right) = \max\left(0, 1 - yf(x)\right)$ is given by

   $$f^*(x) = \operatorname{sign}\left(\pi(x) - \frac{1}{2}\right).$$

   Note that it is impossible to recover $\pi(x)$ from $f^*(x)$ in this scenario. However, in practice we work with an empirical risk minimizer, from which we may still be able to recover a reasonable estimate for $\pi(x)$. An early approach to this problem is known as "Platt scaling": https://en.wikipedia.org/wiki/Platt_scaling.