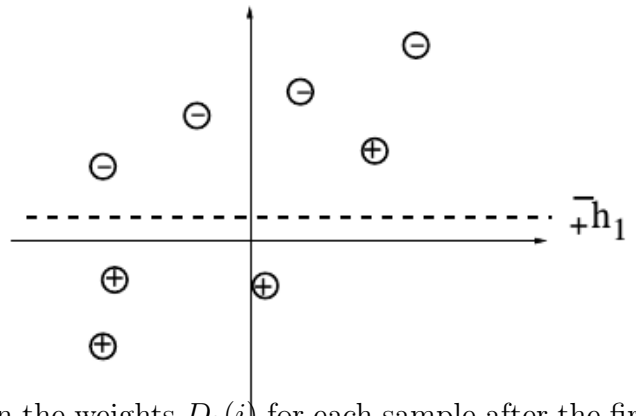
ISYE 6740 Midterm 2

Prof. Yao Xie

Due Nov. 25, 11:55pm
Total Point: 100, Bonus Points: 10.

1 Boosting algorithms [30 points]

In this problem, we test your understanding of AdaBoost algorithm. The figure shows a dataset of 8 points, equally divided among the two classes (positive and negative). The figure also shows a particular choice of decision stump h_1 picked by Adaboost in the first iteration.



(a) (10 points) Explain the weights $D_2(i)$ for each sample after the first iteration. You can explain by drawing figures similar to what we have in class.

(b) (10 points) Calculate the weight α_1 assigned to h_1 by Adaboost? (Note that initial weights of all the data points are equal, $D_1(i) = 1/8, \forall i$).

(c) **[True/False]** (5 points) The votes α_i assigned to weak classifiers in boosting generally goes down as the algorithm proceeds, because the weighted training error of the weak classifiers tend tends to go up.

(d) **[True/False]** (5 points) The votes α assigned to the classifiers assembled by Adaboost are always non-negative.

2 Random forrest for email spam classifier [30 points]

Your task for this question is to build a spam classifier. The UCR email spma dataset `https://archive.ics.uci.edu/ml/datasets/Spambase` came from the postmaster and individuals who had filed spam. The collection of non-spam e-mails came from filed work and personal e-mails. Hence the word 'george' and the area code '650' are indicators of non-spam. These are useful when constructing a personalized spam filter. One would either have to blind such non-spam indicators or get a very wide collection of non-spam to generate a general purpose spam filter.

1. (5 points) Load the data from spambase.data. If there are missing values, you can just fill in zero. Report: How many instances and how many features for each instance in the data set? How many instances of spam versus regular emails are there in the data?

2. (10 points) Partition the data to use the first 80% for training and the remaining 20% for testing. Build a classification tree model (also known as the CART model).

   In Python, you can use sklearn.tree. In MATLAB, this can be done with fitctree. In R, this can be done using library(rpart).

   Choose an appropriate depth of the tree, report the accuracy and the tree model fitted, similar to what is shown on Page 16 of "Random forest" lecture. In Python, this can be done export_graphviz which output a dot file. You can visualize it with dot -T png inputfile.dot -o outputfile.png in command line. In MATLAB, you can visualize the tree with view(tree). In R, this can be done using prp function in library(rpart).

3. (10 points) Build a random forest model.

   In Python, you can use sklearn.ensemble.RandomForestClassifier. In MATLAB, this can be done with TreeBagger. In R, this can be done using library(randomForest).

   Report the accuracy and classifier structure similar as decision tree classifier above.

4. (5 points) Compare and report the AUC for your classification tree and random forest models on testing data, respectively. In classification problem, we use AUC (Area Under The Curve) as a performance measure. It is one of the most important evaluation metrics for checking any classification model?s performance. ROC (Receiver Operating Characteristics) curve measures classification accuracy at various thresholds settings.

AUC measures the total area under the ROC curve. Higher the AUC, better the model is at distinguishing the two classes.

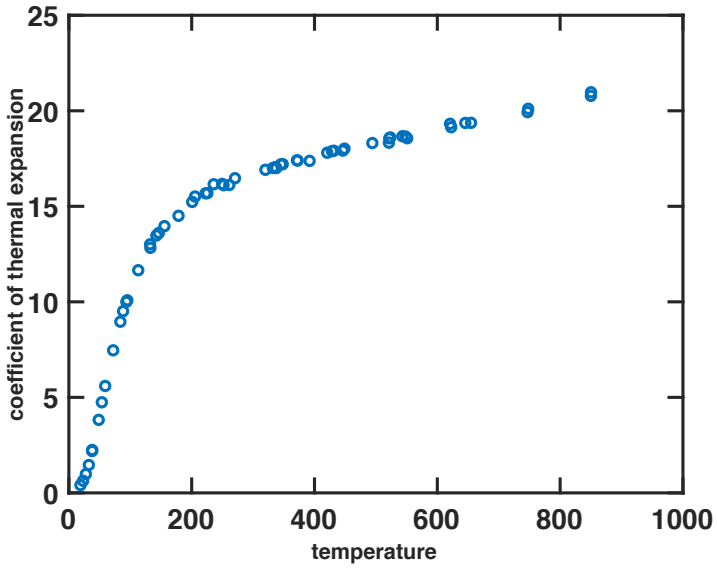
If you want to read a bit more about AUC curve, check out this link `https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5`

For instance, in R, this can be done using library(ROCR) and performance(MyModel, "auc")@y.values and you will have to figure out the details. In MATLAB, a similar function is perfcurve. In Python, you can try sklearn.metrics.roc_auc_score

Try different tree sizes. Plot the curve of AUC versus Tree Size, similar to Page 15 of the Lecture Slides on "Random Forest".

3 Nonlinear regression and cross-validation [40 points]

The coefficient of thermal expansion y changes with temperature x. An experiment to relate y to x was done. Temperature was measured in degrees Kelvin. (The Kelvin temperature is the Celcius temperature plus 273.15). The raw data file is copper-new.txt.



1. (10 points) Perform linear regression on the data. Report the fitted model and the fitting error.

2. (10 points) Perform nonlinear regression with polynomial regression function up to degree $n = 10$ and use ridge regression (see Lecture Slides for "Bias-Variance Trade-off"). Write down your formulation and strategy for doing this, the form of the ridge regression.

3. (10 points) Use 5 fold cross validation to select the optimal regularization parameter λ. Plot the cross validation curve and report the optimal λ.

4. (10 points) Predict the coefficient at 400 degree Kelvin using both models. Comment on how would you compare the accuracy of predictions.

4 (Bonus 10 points) Regression, bias-variance tradeoff

Consider a dataset with n data points (x_i, y_i), $x_i \in \mathbb{R}^p$, drawn from the following linear model:

$$y = x^T \beta^* + \epsilon,$$

where ϵ is a Gaussian noise and the star sign is used to differentiate the true parameter from the estimators that will be introduced later. Consider the regularized linear regression as follows:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2 \right\},$$

where $\lambda \geq 0$ is the regularized parameter. Let $X \in \mathbb{R}^{n \times p}$ denote the matrix obtained by stacking x_i^T in each row.

1. Find the closed form solution for $\hat{\beta}(\lambda)$ and its distribution.

2. Calculate the bias $\mathbb{E}[x^T \hat{\beta}(\lambda)] - x^T \beta^*$ as a function of λ and some fixed test point x.

3. Calculate the variance term $\mathbb{E}\left[\left(x^T \hat{\beta}(\lambda) - \mathbb{E}[x^T \hat{\beta}(\lambda)] \right)^2 \right]$.

4. Use the results from parts (b) and (c) and the bias-variance decomposition to analyze the impact of λ in the squared error. Specifically, which term dominates when λ is small, and large, respectively?

(Hint.) Properties of an affine transformation of a Gaussian random variable will be useful throughout this problem.