CS57300 Spring 2018 Homework	— Homework 2 —	Last update: March 16, 2018
https://www.cs.purdue.edu/homes/ribeirob/courses/Spring2018/index.html		Due: March 24th 6:00am, 2018

Instructions and Policy: Each student should write up their own solutions independently, no copying of any form is allowed. You MUST to indicate the names of the people you discussed a problem with; ideally you should discuss with no more than two other people.
YOU MUST INCLUDE YOUR NAME IN THE HOMEWORK
You need to submit your answer in PDF. LaTeX is typesetting is encouraged but not required. Please write clearly and concisely - clarity and brevity will be rewarded. Refer to known facts as necessary.

Q0 (0pts correct answer, -1,000pts incorrect answer: (0,-1,000) pts): A correct answer to the following questions is worth 0pts. An incorrect answer is worth -1,000pts, which carries over to other homeworks and exams, and can result in an F grade in the course.

(1) Student interaction with other students / individuals:

   (a) I have copied part of my homework from another student or another person (plagiarism).

   (b) Yes, I discussed the homework with another person but came up with my own answers. Their name(s) is (are) _____

   (c) No, I did not discuss the homework with anyone

(2) On using online resources:

   (a) I have copied one of my answers directly from a website (plagiarism).

   (b) I have used online resources to help me answer this question, but I came up with my own answers (you are allowed to use online resources as long as the answer is your own). Here is a list of the websites I have used in this homework:
   _____

   (c) I have not used any online resources except the ones provided in the course website.

Q1 (1.5 pts): Ensembles

1. (0.5pt) Consider a hypothetical case in which you are using bagging for classification. You have two models, both use bagging. The only difference is one uses decision stumps as the weak classifier (M_1) and other decision trees without any depth limit or pruning (M_2). Which model do you expect to have better performance in practice? Why?

2. (1pt) Consider the scenario with training data $\{x_i, y_i\}_{i=1}^N$, where x_i is the set of features of the i-th example and y_i is the value we want to be able to predict. We are interested in bagging K models, where $f_k(x)$ is the prediction of the k-th individual model, $k = 1, \ldots, K$. Assume the squared error of the k-th model is $\epsilon_{k,i} = (y_i - f_k(x_i))^2$, and $\{\epsilon_{k,i}\}_{k,i}$ are independent and identically distributed random variables. Let $e_{\text{single,k}}$ be the expected error of individual model k,

$$e_{\text{single,k}} = E\left[\sum_i \epsilon_{k,i}\right].$$

Let

$$f^\star(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x)$$

be the prediction of the combined model and

$$e_{\text{bagging}} = E\left[\sum_i \epsilon'_{k,i}\right],$$

where $\epsilon'_{k,i} = (y_i - f^\star(x_i))^2$.

Derive the relationship between $e_{\text{single,k}}$ and e_{bagging}.

Q2 (2.5 pts): Classification Tasks

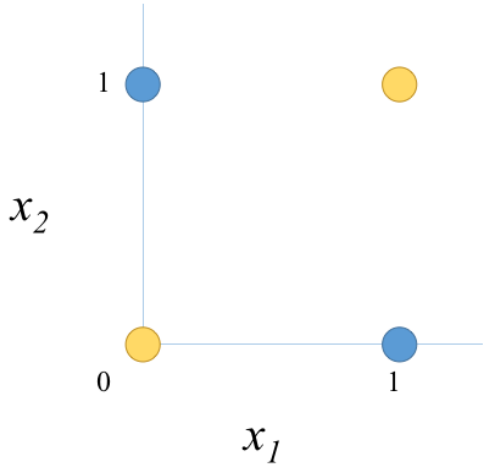


Figure 1: Training data

Consider a classification task in which each data point is represented by two boolean attributes $x_1, x_2 \in \{0, 1\}$. The training data with two class labels {Yellow, Blue} is depicted in Figure 1.

Answer the following questions:

(a) (1pts) Suppose we use the above dataset for training a decision tree of arbitrary depth. Would the tree be able to correctly classify all training data correctly. If yes, write the decision tree. If no, explain why and what you would do to solve the issue.

(b) (1pts) Suppose we use the above dataset for training a logistic regression classifier (over the original feature space). Would the logistic classifier be able to correctly classify all points? If yes, write the logistic classifier. If no, explain why and what you would do to solve the issue.

(c) (0.5pts) Suppose we use the above dataset for training a Naive Bayes classifier (NBC). Would the NBC be able to correctly classify all points? If yes, give the NBC parameters (conditional probability distributions and prior probabilities). If no, explain why and what you would do to solve the issue.

Q3 (2.0 pts): Decision Tree
Consider a dataset with n points, each having m real-valued attributes. We will construct a decision tree on these points by maximizing information gain as the splitting criterion.

(a) (0.5pts) Is it possible that while constructing the decision tree we need to use the same attribute for splitting more than once at different levels of the tree? If yes, give an example. If no, explain why.

(b) (0.5pts) Decide whether this statement is correct: "The information gain at the root node will always be greater than or equal to the information gain at any lower node in the tree". Explain why or why not.

(c) (0.5pts) **Random Forests**
Lets suppose we build a random forest consisting of P binary trees where each binary tree has Q internal nodes. Assume that for splitting we randomly select 1 feature at each node. Find the probability that a particular feature does not get considered for splitting even once.

(d) (0.5pts) Given the data, attribute X_1, X_2 can have values {0,1,2}, Label can be {0,1}

X_1	X_2	Label
1	1	0
2	2	1
0	2	1
0	1	0
2	0	1
2	1	?
1	0	?

Find the correct labels for the '?' in the table if we know the information gain for attribute $X_1 = 0.29169$ and for $X_2 = 0.4695$.

Q4 (1.5 pts): Naive Bayes Classifier
Consider a Naive Bayes classifier with the following conditional probability: Where $\mathbf{x} = (x_1, x_2, x_3)$ is a

$P(x_1 = 1 \mid y = 1)$	$1/2$	$P(x_1 = 0 \mid y = 1)$	$1/2$
$P(x_1 = 1 \mid y = 0)$	$1/3$	$P(x_1 = 0 \mid y = 0)$	$2/3$
$P(x_2 = 1 \mid y = 1)$	$3/10$	$P(x_2 = 0 \mid y = 1)$	$7/10$
$P(x_2 = 1 \mid y = 0)$	$1/4$	$P(x_2 = 0 \mid y = 0)$	$3/4$
$P(x_3 = 1 \mid y = 1)$	$1/5$	$P(x_3 = 0 \mid y = 1)$	$4/5$
$P(x_3 = 1 \mid y = 0)$	$2/3$	$P(x_3 = 0 \mid y = 0)$	$1/3$

document, and x_is are words. $x_i \in \{1, 0\}$. And y is the class of the documents with the following prior probabilities:

$P(y = 1)$	$P(y = 0)$
$4/10$	$6/10$

(a) (0.5pts) Consider a document with counts $\mathbf{x} = (1, 0, 1)$, which class has highest posterior probability?

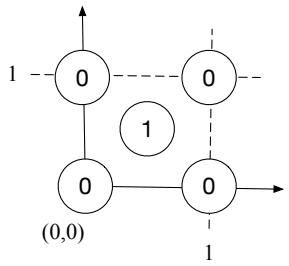
(b) (0.5pts) Describe the decision boundary for the above Naive Bayes classifier.

(c) (0.5pts) Suppose the modeling assumptions made by the Naive bayes classifier are true, and we have infinite training data. Will the learned Naive Bayes classifier have zero training error? If so, explain why; if not explain why not.

Q5 (2.5 pts): Boosting

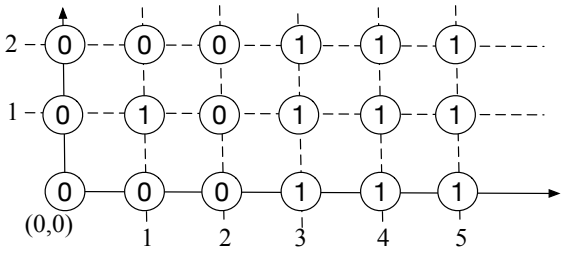
In this question, we explore how to build a AdaBoost model with decision stumps (trees of depth one).

(a) (1pt) Consider the following dataset $\{((0,0),0),((0,1),0),((1,0),0),((1,1),0),((0.5,0.5),1)\}$ whose elements are of the form $((x_{i,1},x_{i,2}),y_i)$, where $(x_{i,1},x_{i,2})$ is a vector of features of example i and y_i is its true class label. This data is illustrated in the following figure:



Answer the following question: **Which examples will have their weights increased at the end of the first iteration? Describe how the algorithm proceeds.**

(b) (1.5pts) Consider the training data depicted in the figure:
Hint: Note that one of the data points seems to be mislabeled.



Answer the following questions:

(i) (0.5pts) What is the minimum number of iterations to achieve zero training error with Adaboost and decision stumps over this data?

(ii) (0.5pts) What is the minimum number of iterations to achieve zero training error with Adaboost using decision trees with depth two over this data?

(iii) (0.5pts) What's the advantage of using decision stumps against a depth-two tree in this dataset?