Homework 6

Boosting, Learning Theory, Markov Decision Process

CMU 10-601: Machine Learning (Fall 2015)
http://www.cs.cmu.edu/~10601b/
OUT: Nov 30, 2015
DUE: Dec 10, 2015, 10:30 AM

START HERE: Instructions

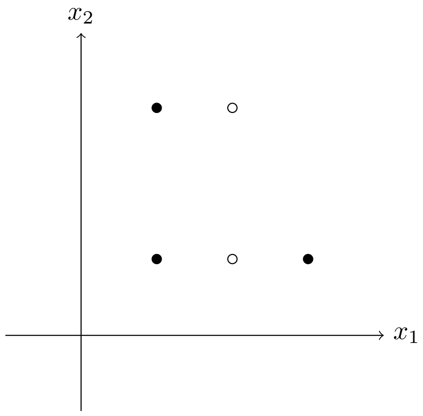
- The homework is due at 10:30 am on Thursday December 10, 2015. Each student will given two late days that can be spent on any homeworks but not on projects. Once you have used up your late days for the term, late homework submissions will receive 50% of the grade if they are one day late, and 0% if they are late by more than one day.
- ALL answers will be submitted electronically through the submission website: https://autolab.cs.cmu.edu/10601-f15. You can sign in using your Andrew credentials. You should make sure to edit your account information and choose a nickname/handle. This handle will be used to display your results for any competition questions (such as the class project) on the class leaderboard.
- There are no autograded questions on this homework.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). When you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution. You should also state your collaborations in your short-answer writeup. Specifically, please write down the following:

  1. Did you receive any help whatsoever from anyone in solving this assignment? Yes / No. If you answered yes, give full details: (e.g., "Jane explained to me what is asked in Question 3.4").
  2. Did you give any help whatsoever to anyone in solving this assignment? Yes / No. If you answered yes, give full details: (e.g., "I pointed Joe to section 2.3 to help him with Question 2").

  Collaboration without full disclosure will be handled severely, in compliance with CMU's Policy on Cheating and Plagiarism.

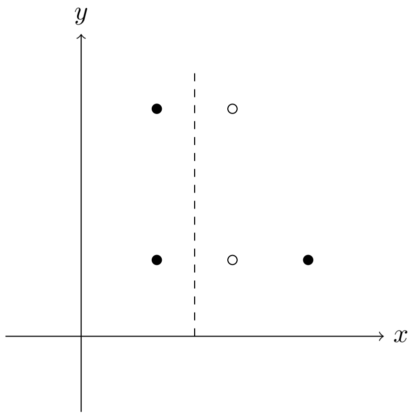
1 Boosting [Pengcheng Xu; 17 points]

Consider training a boosting classifier on the following data set:



We use a decision stump as each weak hypothesis h_i in Adaptive Boosting. A decision stump classifier is a one-level decision tree with a single test on a single input variable. Learning a decision stump involves choosing $x \in \{x_1, x_2\}$ and a constant c to define the test $x > c$.

1. What is the initial weight that is assigned to each data point [3pts]?
2. In the first iteration ($t = 1$) we choose the following decision stump. What are the updated weights of each data point? Show your calculations. [5pts] Circle (or describe) the point(s) whose weight increases in the next boosting iteration [3pts].



Decision stump for the first itertaion: classify the left as black point, the right as white point.

3. Is it possible to achieve zero training error with Adaptive Boosting [2 pt]? Explain your answer with the decision stumps in the following iterations and the resulting final decision boundary [4pt]?

2 PAC Learning [30 points]

Consider a PAC learning setting for learning a classifier for predicting a binary output variable Y from 20 binary input features $\mathbf{X} = \{x_1, \ldots, x_{20}\}$. Assume noiseless class labels in training data. Consider two hypothesis spaces H_1= $\{h_1 : \mathbf{X} \to Y\}$, where H_1 is identical to the set of all possible target concepts C, and H_2= $\{h_2 : (x_a, x_b) \to Y, \text{where } x_a, x_b \in \mathbf{X}, a \neq b\}$.

1. Is a learner using H_1 guaranteed to learn a consistent hypothesis? Explain in one sentence. [4pts]
2. What is the size of H_1, i.e., $|H_1|$? [4pts]
3. How many training examples m suffice to assure that with probability 95% any consistent learner using H_1 will output a hypothesis with true error at most 0.01? [5pts]
4. Is a learner using H_2 guaranteed to learn a consistent hypothesis? Explain in one sentence. [4pts]
5. What is the size of H_2, i.e., $|H_2|$? [4pts]
6. How many training examples m suffice to assure that with probability 95% any consistent learner using H_2 will output a hypothesis with true error at most 0.01? [5pts]
7. Which of the two hypothesis sets has a greater bias? [4pts]

3 Bias/Variance Trade-off [Pengcheng Xu; 24 points]

When choosing a model, one needs to find a model with the right balance between bias and variance. This ensures the model is not overfitting but still has the expressive power to represent the underlying pattern in data. What is the effect of the tuning parameters in the following examples on bias and variance?

1. Given $\mathbf{x}^i = \{x_1, \ldots, x_n\}$ for n input features and real-valued output y^i for the ith sample, we would like to fit a linear regression with L_1 penalty to select input features relevant to predicting the output.

$$J(\theta) = \frac{1}{2n}\left[\sum_{i=1}^{m}(y^i - h_\theta(\mathbf{x}^i))^2 + \lambda \sum_{j=1}^{n}|\theta_j|\right],$$

   How does a larger λ affect the estimated model in terms of bias and variance? [4pts]

2. In neural networks, use a higher number of hidden units. [4pts]
3. In decision trees, use a higher upper limit on the number of nodes. [4pts]
4. Use a larger K, when reducing the dimensionality of input data (with original dimension $> K$) to K dimension with PCA, and using the PCA projections as the inputs for classification task. [4pts]
5. Use a hidden Markov model with fewer states. [4pts]
6. In Boosting, use a higher number of iterations. [4pts]

4 Bayes Error [Zhenzhen; 15 points]

		X_1	
		0	1
X_2	0	0.2	0.1
	1	0.4	0.2
	2	0	0.1

$P(X_1, X_2|Y = 0)$

		X_1	
		0	1
X_2	0	0.6	0.1
	1	0.1	0.1
	2	0.1	0

$P(X_1, X_2|Y = 1)$

		X_1	
		0	1
X_2	0	0.1	0.4
	1	0.3	0
	2	0.2	0

$P(X_1, X_2|Y = 2)$

Assume that the true underlying probability distribution over two input features $\mathbf{X} = (X_1, X_2)$ and class label Y has the marginal distribution $P(Y)$ given by $P(Y = 0) = 0.4$, $P(Y = 1) = 0.3$, and $P(Y = 2) = 0.3$, and class-conditional distributions $P(\mathbf{X}|Y)$ as specified in Table 4.

- Give the classifier based on the Bayes optimal decision rule. In other words, for each of the 6 possible values of $\mathbf{X}$, what is the prediction for Y? [5pt]
- Compute the Bayes error. [5pt]
- Is it possible to build a classifier for this problem that achieves 0 training error? Explain your answer in one sentence. [5pt]

5 Markov Decision Process [Zhenzhen; 14 points]

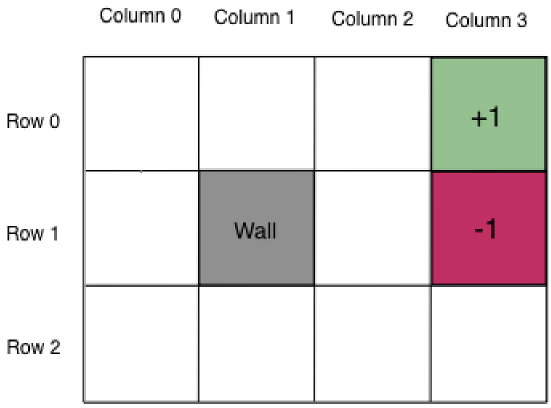
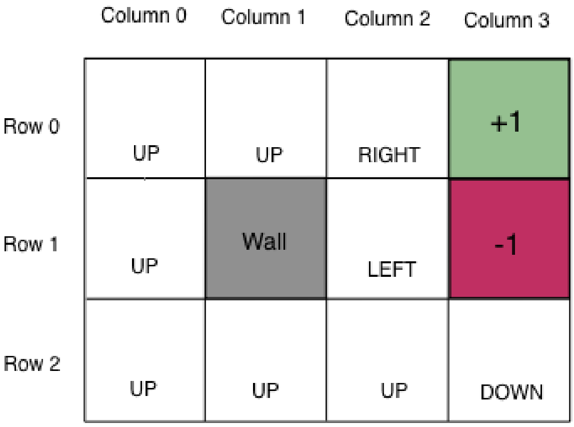


Figure 1

Recall the “Robot in a Room” example from lecture. The robot lives in a grid as in Figure 1, and it wants to figure out which action to take in a given position. A robot has four actions: UP, DOWN, LEFT, and RIGHT. When it tries to move forward, it has a 80% chance of succeeding, a 10% chance of ending up in the block to its left, and a 10% chance of ending up in the block to its right. If the robot runs into the wall or the edge, its locations does not change. The robot gets reward +1 if it moves to state (Row 0, Column 3), and it gets reward -1 if it moves to state (Row 1, Column 2). Every other move has reward 0. When the robot moves to (Row 0, Column 2) or (Row 1, Column 2), it terminates.

Perform the value iteration algorithm for 2 iterations with discount factor 0.9. For each iteration, calculate the best policy for each state and its corresponding value. If multiple policies have the same value, break ties by giving priority in the following order:“UP > DOWN > LEFT > RIGHT”. For example, after the first iteration, the optimal policy is shown in Figure 2.



	Column 0	Column 1	Column 2	Column 3
Row 0	UP	UP	RIGHT	+1
Row 1	UP	Wall	LEFT	-1
Row 2	UP	UP	UP	DOWN

Figure 2

- What is the corresponding value for each policy in Figure 2? [7pts]
- What are the optimal policies after iteration 2? What are their corresponding values? [7pts]