15-780: Graduate AI
Homework Assignment #4A

Out: April 11, 2015
Due: April 22, 2015 5 PM

Collaboration Policy: You may discuss the problems with others, but you must write all code and your writeup independently.

Turning In: Please email your assignment by the due date to shayand@cs.cmu.edu and vdperera@cs.cmu.edu. Your solutions should be submitted as a **single** pdf file. If your solutions are handwritten, **scan** them and make sure they are legible and clear. Please submit your code in separate files and provide instructions on how to run it.

1 Decision Trees

Your are trying to build a classifier to figure out which restaurant is best suited for a dinner with your friends. You gathered data about 11 different restaurants and in particular about the kind of restaurant (fast food, ethnic or casual dining), their prices (low, average or high), their locations (Oakland, Shadyside or Squirrel Hill), whether they can comply with dietary restrictions (none, vegetarian or gluten free) and whether you enjoyed them or not. The data is reported in the following table:

Restaurant	Type	Price	Neighborhood	Restriction	OK
R_1	Fast Food	$	Oakland	Vegetarian	0
R_2	Ethnic	$$	Squirrel Hill	Gluten Free	0
R_3	Casual Dining	$$	Squirrel Hill	None	0
R_4	Casual Dining	$$$	Shadyside	Vegetarian	0
R_5	Casual Dining	$	Oakland	Vegetarian	1
R_6	Fast Food	$$	Squirrel Hill	None	1
R_7	Ethnic	$	Squirrel Hill	None	1
R_8	Casual Dining	$	Shadyside	Gluten Free	0
R_9	Fast Food	$$$	Oakland	None	0
R_{10}	Ethnic	$$	Shadyside	Vegetarian	1
R_{11}	Casual Dining	$$	Shadyside	Gluten Free	1

a) Using this data build a decision tree to decide whether you would enjoy a particular restaurant or not, showing at each level how you decided which attribute to expand next.

b) What is the training set error $E_{train}(h)$ of your decision tree (i.e. the fraction of points in the training set that it misclassified)?

c) You are now given data from five more restaurants:

Restaurant	Type	Price	Neighborhood	Restriction
R_{12}	Fast Food	\$	Squirrel Hill	None
R_{13}	Ethnic	\$\$	Shadyside	None
R_{14}	Ethnic	\$	Oakland	Gluten Free
R_{15}	Casual Dining	\$	Shadyside	Vegetarian
R_{16}	Ethnic	\$	Squirrel Hill	Gluten Free

To which one would you go?

d) Out of curiosity, and to verify your decision tree accuracy, you decide to try them all. The results are:

Restaurant	OK
R_{12}	0
R_{13}	1
R_{14}	0
R_{15}	1
R_{16}	0

How good did your decision tree do? What is the test set error $E_{test}(h)$? What is the F_1 score? To what do you attribute the results of your decision tree?

Note: For this problem use the following definition of F_1:

$$F_1 = 2 \cdot \frac{\text{precision} + \text{recall}}{\text{precision} \cdot \text{recall}}$$

Where, given the number of true positives (TP), false positives (FP) and false negatives (FN)

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2 VC-Dimension

In this problem we ask you to determine the VC-dimension of various concept classes, as well as seeing how the theoretical bounds on the true error of a classifier using VC-dimension compares to the empirical error estimate obtained in Problem 1. Please justify all your answers.

1. What is the VC dimension for the concept class $\mathcal{H}$ of all circles in $\mathbb{R}^2$? Formally, $\mathcal{H} = \{h_{r,c_1,c_2} | r, c_1, c_2 \in \mathbb{R}\}$ where

$$h_{r,c_1,c_2}(\mathbf{x}) = \begin{cases} +1 & \text{if } (x_1 - c_1)^2 + (x_2 - c_2)^2 = r^2 \\ -1 & \text{otherwise} \end{cases}$$

2. What is the VC dimension for the concept class of all decision trees with at most 7 nodes in one dimension (i.e. with only one attribute)?

3. What is the VC dimension for the concept class of decision trees with an arbitrary number of nodes in one dimension?

4. Think of a reasonable concept class of decision trees that your algorithm considered when constructing the decision tree in Problem 1 taking into account the restrictions of the data. What is the VC-dimension of that concept class?

5. Recall that in class, we gave the following bound for the true error of a hypothesis $h \in \mathcal{H}$ that holds with probability $1 - \delta$:

$$E_{true}(h) < E_{train}(h) + \sqrt{\frac{VC(\mathcal{H})(\ln(\frac{2m}{VC(\mathcal{H})}) + 1) + \ln\frac{4}{\delta}}{m}}$$

   where m is the size of the training set. Use this equation to compute an upper bound on $E_{true}(h)$ for the decision tree h that you came up with in Problem 1 using $\delta = 0.9$ (i.e. the bound will hold with only 0.1 probability) and $\delta = 0.1$ (i.e. the bound will hold with 0.9 probability). How do these compare to $E_{test}(h)$ as computed in Problem 1?

3 k-NN

In Figure 1 we show a set of training points classified as being either black or white. Consider using the k-Nearest Neighbors algorithm to classify new points.

1. How is the point marked by "?" classified using Euclidean distance as the distance metric for k=1, 2, and 3?

2. Are there any points in the training set that would be misclassified using k=1? If so, identify them.

3. Come up with a simple distance metric that would properly classify all the points in the training set for k=1?

4. What happens when k=5 using your distance metric?

5. How does your distance metric classify the "?" for k=1, 2, and 3?

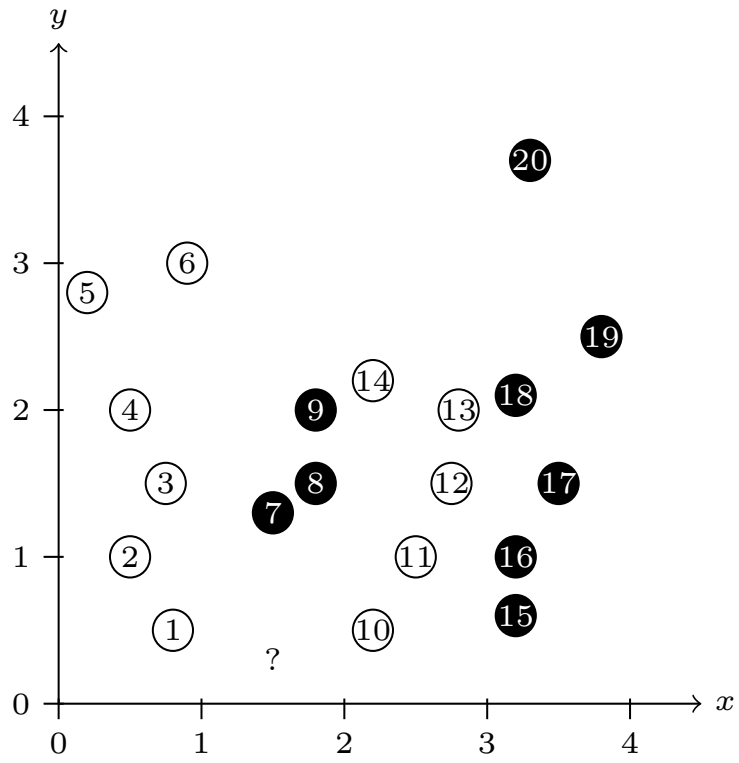

Figure 1: Training Set for Problem 3