15-780: Graduate AI
Homework Assignment #4A Solutions

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

The following solution was provided by Revanth Bhattaram (with slight modifications). In the decision trees - the color blue denotes an impure node, the color red denotes a node with $OK = 0$ and the color green denotes a node with $OK = 1$.

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.

Let's look at the first level. We will be trying all attributes and checking which ones yields the maximum information gain upon splitting.

Information gain of the whole dataset is :

$$-\frac{6}{11}\log(\frac{6}{11}) - \frac{5}{11}\log(\frac{5}{11}) = 0.994$$

- **Type :** When type is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Fast Food	2	1	0.918
Casual Dining	3	2	0.970
Ethnic	1	2	0.918

Conditional Entropy turns out to be : 0.9416.

- **Price :** When price is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
$	2	2	1
$$	2	3	0.970
$$$	2	0	0

Conditional Entropy turns out to be : 0.8045.

- **Neighborhood :** When neighborhood is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Oakland	2	1	0.918
Squirrel Hill	2	2	1
Shadyside	2	2	1

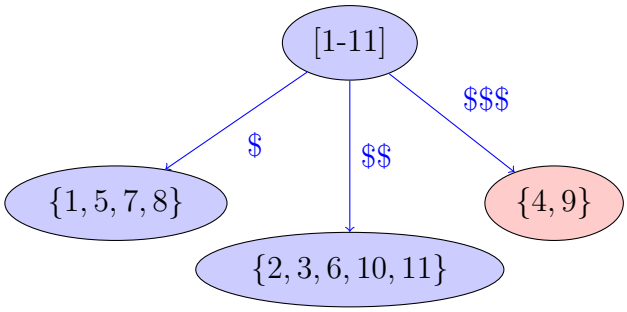
Conditional Entropy turns out to be : 0.9776.

- **Restriction :** When restriction is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Vegetarian	2	2	1
Gluten Free	2	1	0.918
None	2	2	1

Conditional Entropy turns out to be : 0.9776.

Splitting on **price** gives the maximum information gain. The decision tree so far is as follows :



We now have two nodes to further split on.

Child - 1 {1,5,7,8} : Let's look at splitting the first child first.

- **Type :** When type is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Fast Food	1	0	0
Casual Dining	1	1	1
Ethnic	0	1	0

Conditional Entropy turns out to be : 0.50.

- **Neighborhood :** When neighborhood is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Oakland	1	1	1
Squirrel Hill	0	1	0
Shadyside	1	0	0

Conditional Entropy turns out to be : 0.5.

- **Restriction :** When restriction is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Vegetarian	1	1	1
Gluten Free	1	0	0
None	0	1	0

Conditional Entropy turns out to be : 0.5.

All three attributes result in the same information gain. So, let's split on neighborhood (random).

Child - 2 {2,3,6,10,11} : Let's look at splitting the second child now.

- **Type :** When type is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Fast Food	0	1	0
Casual Dining	1	1	1
Ethnic	1	1	1

Conditional Entropy turns out to be : 0.80.

- **Neighborhood :** When neighborhood is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Oakland	0	0	0
Squirrel Hill	2	1	0.918
Shadyside	0	2	0

Conditional Entropy turns out to be : 0.55.

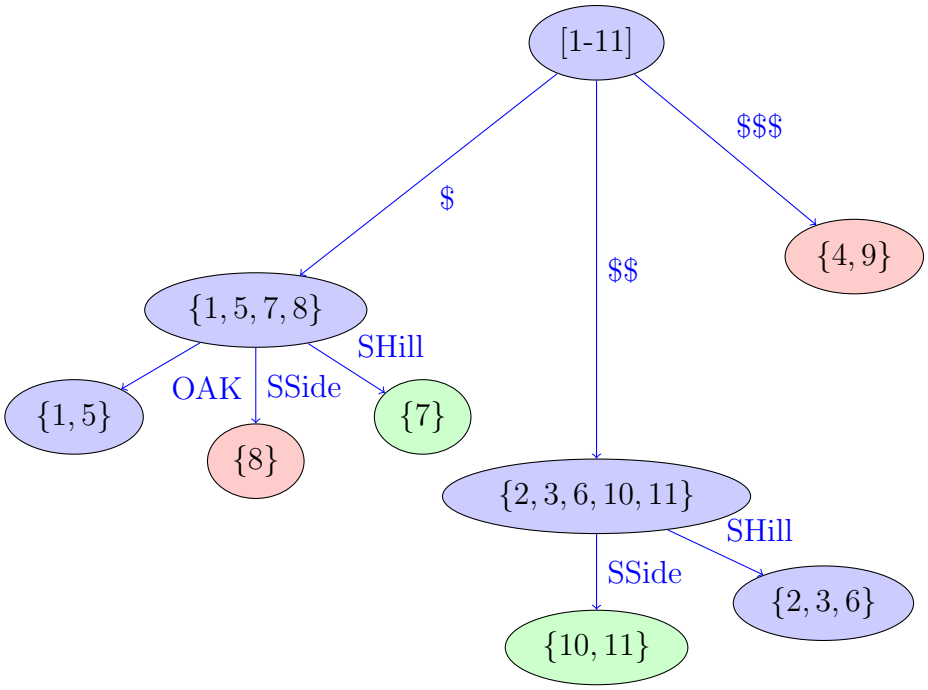
- **Restriction :** When restriction is used as an attribute, we get the following split:

Value	$OK = 0$	$OK = 1$	Entropy
Vegetarian	0	1	0
Gluten Free	1	1	1
None	1	1	1

Conditional Entropy turns out to be : 0.8.

The best attribute to split on here is neighborhood.

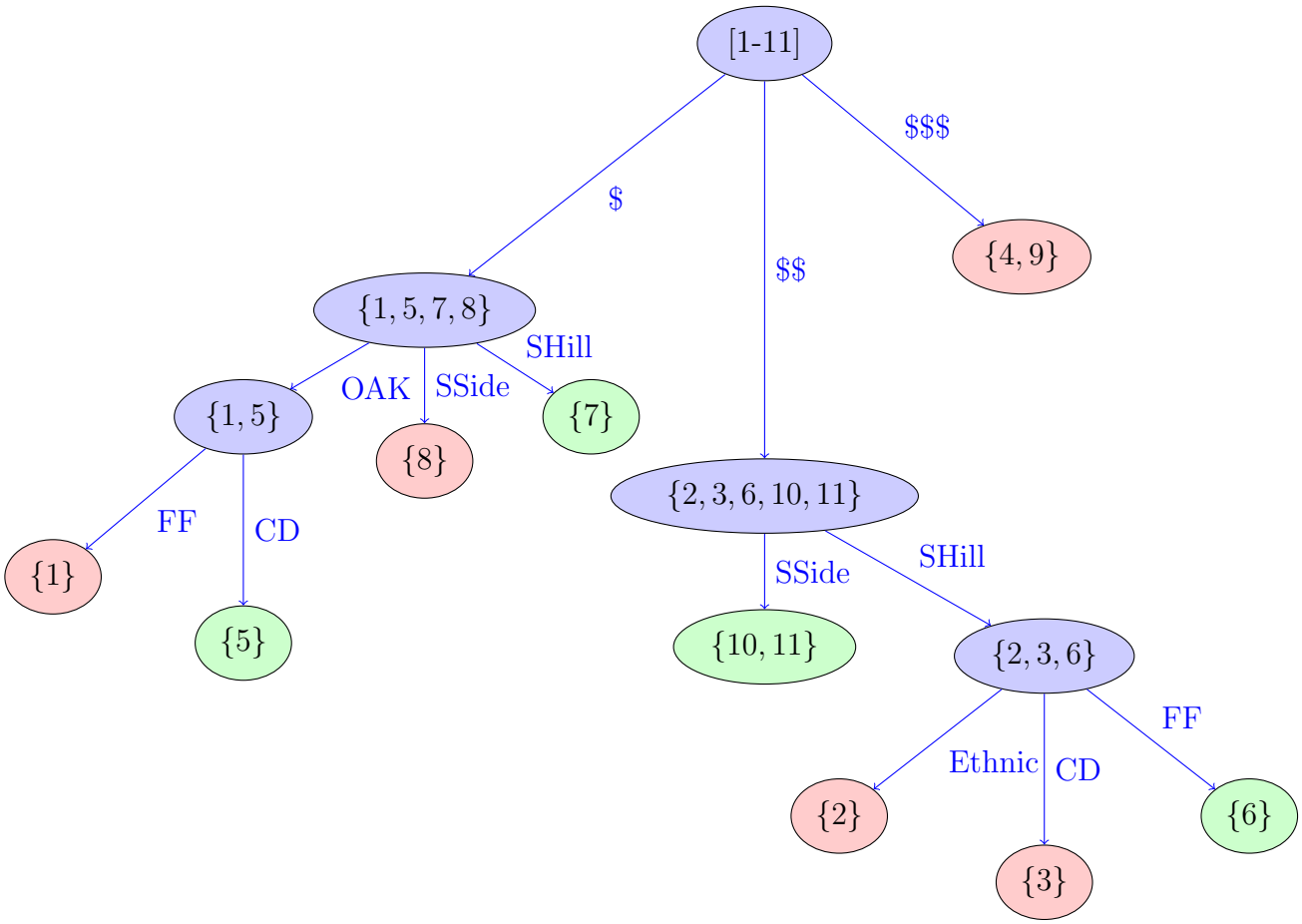
The decision tree now looks like :



Let's split the three children now :

- **Child - 1 {1,5} :** These can simply be split on the Type attribute which gives an entropy of zero.
- **Child - 2 {2,3,6} :** Similar to the previous child, we see that splitting on the Type attribute gives an entropy of zero.

The final decision tree is as follows :



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)?

The training error here is 0. All training points are correctly classified.

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?

Following the paths of the decision trees, we get the following values for each node:

Restaurant	Predicted Value	Actual Value
R_{12}	1	0
R_{13}	1	1
R_{14}	0	0
R_{15}	0	1
R_{16}	1	0

Note that for this R_{14}, the path ends at an impure node and so you are free to decide how to label it.We decided to assign the label that is the majority by walking up the path to the parent.

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?

Only two out of five are correct. So, test error is 0.4.

Number of true positives = 1.
Number of false positives = 2.
Number of false negatives = 1.
Precision = $\frac{1}{1+2} = \frac{1}{3}$
Recall = $\frac{1}{1+1} = \frac{1}{2}$
F1 Measure = $2 * \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{3}+\frac{1}{2}} = \frac{2}{5} = 0.4$.

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

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

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

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

   The VC-dimension is 3. First, we will show that we can shatter a set of three points. (In fact, we can shatter *any* set of three points that do not lie on a line.) We will use two facts: (1) there is only circle that goes through any set of three points that do not lie on a line, and (2) more than one circle goes through any two points (in fact, there are infinitely many such circles, but we don't need that). These two facts imply that (1) we can label all the points with +1 by drawing a circle through them, and (2) we can label any set of two points as +1 by drawing a circle through them that is not the same as the circle that goes through all the points. Furthermore, we can trivially draw a circle that includes none of the points, and we can trivially draw a circle that goes through only one point. Thus we can shatter three points.

   We now show that we cannot shatter any set of four points. To shatter four points, these points need to all lie on a circle. We can choose four points such that there is a unique circle that all of these points can lie on, but according to fact (1) above, each subset of three points also only share this unique circle. Thus we cannot label any set of three points with +1 without labeling the other point as +1.

   Thus the VC-dimension is 3.

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)?

   The VC-dimension is 6, because a seven node tree with only one attribute can have one root node and 6 leaf nodes. Each of these leaves can be labeled with +1 or −1. Thus we can shatter any set of 6 points where each point takes on a different value. If we have an additional point, it must also be assigned to one of these leaves. (Actually, normally for decision trees we branch out for any possible value the attribute can take, so if this attribute takes on 7 or more values, we can't really construct a proper decision tree for it.)

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

The VC-dimension is ∞: since we are allowed to have an arbitrary number of nodes, we can have an arbitrary number of leaves, and thus we can shatter any set of points with an arbitrary number of leaves (as in the analysis for 7 nodes).

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?

   The VC-dimension is 81. Consider a dataset of all unique points with 4 attributes that each take on 3 values: there are exactly $3^4 = 81$ such points. Notice that since each point is unique, our algorithm will always construct a decision tree that assign each point to a leaf that gives it the correct label, regardless of the labeling. Thus we can shatter 81 points. We can't shatter any more because then we would have two points with the same set of features that possibly have different labels.

   Some considered the fact that we only have 11 data points in the dataset, so we can only construct a decision tree that can classify at most 11 points, and hence under that consideration the VC-dimension is 11. In some sense it is true that the VC-dimension will never be greater than the number of data points we have. As we begin to see in the next part the VC-dimension is generally only useful when it is much less than the number of points in our dataset. This is also an acceptable answer.

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?

   Using $E_{train}(h) = 0$ and $VC(\mathcal{H}) = 81$, we get that $E_{true}(h) < 1.45i$ when $\delta = 0.9$ and $E_{true}(h) < 1.38i$ when $\delta = 0.1$. Of course these bounds are meaningless because the number of data points is too small for the bound to hold.

   If we instead use $VC(\mathcal{H}) = 11$, we get that $E_{true}(h) < 1.35$ when $\delta = 0.9$ and $E_{true}(h) < 1.42$ when $\delta = 0.1$. These bounds are now meaningful but they are still useless, because we already know that $E_{true}(h) \leq 1$ with probability 1 (i.e. with $\delta = 0$)!

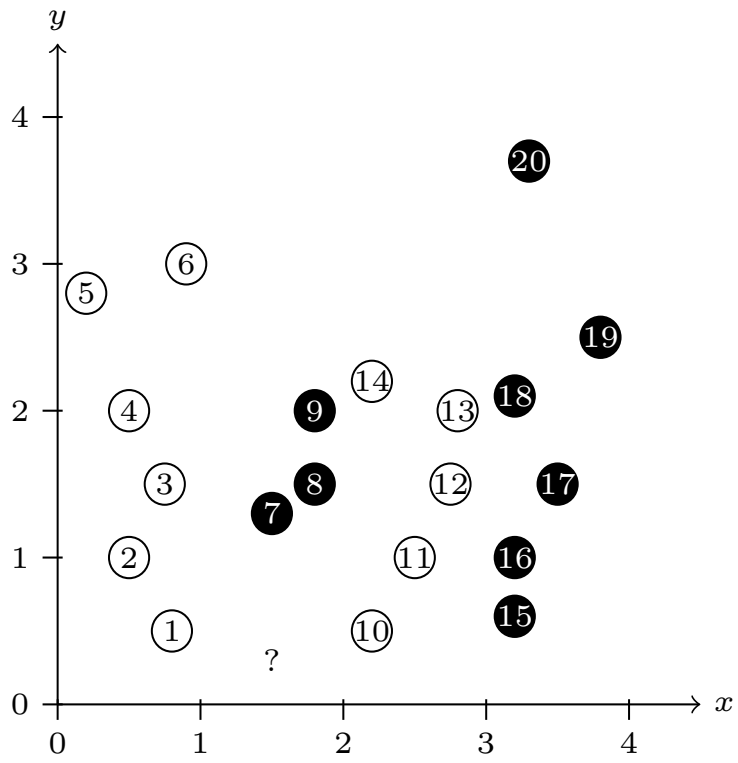


Figure 1: Training Set for Problem 3

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?

   It's classified as white 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.

   Yes, points 9, 14, 13, and 18 would be misclassified using $k = 1$. (Of course we mean if that point wasn't in the training dataset it would be misclassified by $k = 1$.)

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

   A simple distance metric would be to use the distance on the x-axis.

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

   Using $k = 5$ some points in the training set would be misclassified: points 7,8, and 9 would all be classified as white, and points 12 and 13 would be classified as black.

(Interestingly, we would still classify unknown points near 7,8, and 9 correctly, because the nearest three neighbors are all black. If you pointed this out, you would also get full credit, because the question was ambiguous. The point is though that if k gets large enough an entire region that should be classified as black will be misclassified as white.)

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

   It's classified as black for $k = 1$, 2, and 3.