IDS575: Statistical Models and Methods *Updated: 01/28/2020*

Homework #1

Instructor: Moontae Lee *Total Points: 100*

Policy

1. HW1 is due by 02/09/2020 11:59PM in Central Time. One submission per each group.
2. You are allowed to work individually or as a group of up to three students.
3. Having wider discussions is not prohibited. Put all the names of students beyond your group members. However individual students/groups must write their own solutions.
4. Put all of your write-up and programming results in a single pdf report. Compress all other R codes into one zip file. Each student/group must submit only two files.
5. If you would include some graphs, be sure to include the source codes together that were used to generate those figures. Every result must be reproducible!
6. Maximally leverage Piazza to benefit other students by your questions and answers. Try to be updated by checking notifications in both Piazza and the class webpage.
7. Late submissions will be penalized 20% per each late day. No assignment will be accepted more than 3 days after its due date. For HW1, it will be 02/12/2020 11:59pm

Problem 1: Instance-based Learning [25 points]

Alice just started her program at UIC this semester. As Alice is new in town, she has tried several restaurants based on Opentable reviews, but none of them is truly satisfactory. She concludes that the review system does not match well with her taste because the star-ratings are too detailed. Opentable reviews consist of four different topics: food, ambiance, service, and noise level. Now she wants to **simplify** evaluation criteria as follows:

- *Food:* Only three categories: *poor*, *average*, or *good* food.
- *Ambiance:* Only three categories such as *inferior*, *normal*, or *superior* ambiance.
- *Service:*: Only two categories: *bad* or *fair* service.
- *Noise-level:* Ignore as this is too subjective and sensitive to visiting hours.

Based on these simplified criteria, Alice tries to predict whether or not she would *like* a certain restaurant. The next table shows her initial research for five different restaurants in West Loop.

Restaurant	Food?	Ambiance?	Service?	Like?
1	good	normal	fair	yes
2	average	superior	bad	yes
3	poor	superior	bad	no
4	good	inferior	fair	no
5	average	normal	bad	no

Our purpose is to learn a function: $F \times A \times S \to L$ using the table with the training data D of five examples.

(a) Define each attribute $F(ood), A(mbiance), S(ervice)$, and the output space $L(ike)$, respectively as a set of possible values. What is the size of the instance space X?

(b) Suppose that the hypothesis space consists of all possible function $h : F \times A \times S \to L$. What is the size of the hypothesis space H?

(c) Initially Alice thought that *she will like any restaurant that serves good food.* Is this hypothesis h consistent with the training set D? Why or why not?

Alice later realizes that the hypothesis space H under (b) is too large. She now tries to evaluate each criterion by assigning numeric scores such as ($poor = -1$, $average = 0$, $good = 1$) for Food, ($inferior = -1$, $normal = 0$, $superior = 1$) for Ambiance, ($bad = -1$, $fair = 1$) for Service. Her goal is to restrict hypotheses to certain functions that only uses the **sum of any two attribute values** for predicting her potential likeness. In particular, $Like = yes$ if $sum > 0$, $Like = no$ if $sum \leq 0$. For instance, Alice now *likes* the Restaurant 4 because it achieves 2 points from its *good food* and *fair service* even with *inferior ambiance.*

(d) Let H' be the new hypothesis space satisfying the above numeric formulation. Measure the size of the new hypothesis space H'. (Hint: Try to see each of the three cases where only $\{F, A\}$, $\{F, S\}$, and $\{F, S\}$ matters, respectively)

Problem 2: k-Nearest Neighbors Algorithm [25 points]

Consider a binary classification problem with two real valued features x_1 and x_2. Figures 1 & 2 illustrate two different training sets D_1 and D_2. White circles denote positively labeled examples, whereas black squares denote the negatively labeled examples. In order to classify a new instance point, we will use (unweighted) k-Nearest Neighbors with Euclidean distance and various k. Thus the label for a new point will be predicted by the majority class (*i.e.*, positive or negative) among the k closest examples around the query point to be classified.

(a) Draw the decision boundaries of D_1 and D_2 when $k = 1$.

(b) Which label would you suggest for (3, 2) and (4, 2) in D_1? Ties are broken by predicting the **positive** class.

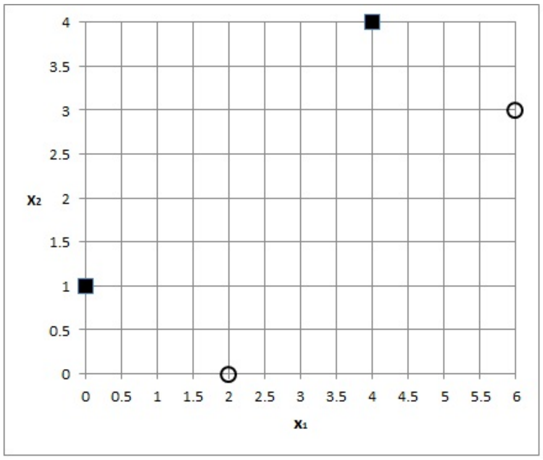

Figure 1: The training set S_1

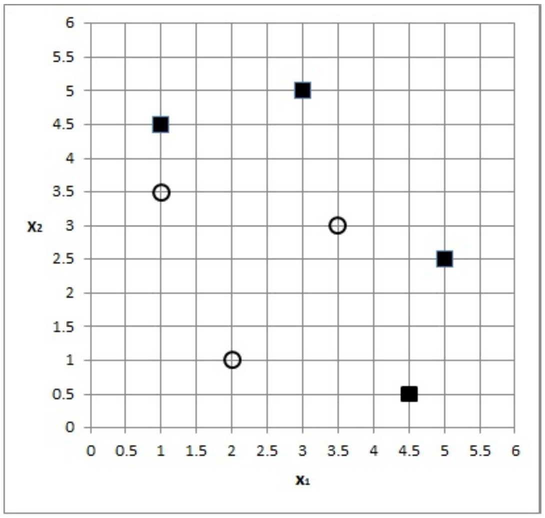

Figure 2: The training set S_2

(c) Which label would you suggest for (4.5, 4) and (4, 2) be in D_2? Ties are broken by predicting the **positive** class.

(d) When $k > 1$, a partition of spaces like the above is called the k^{th}-order Voronoi Diagram or Voronoi Tesselation. Try to draw the decision boundaries of D_1 when $k = 3$. (Hint: Try to draw every bisector between all pairs of positive and negative examples)

(e) (Optional +5pts): If the x_2-coordinate of four example points in D_1 were multiplied by 5, what would happen to its decision boundary when $k = 1$? Draw another picture. Could this effect cause problems when working with real data?

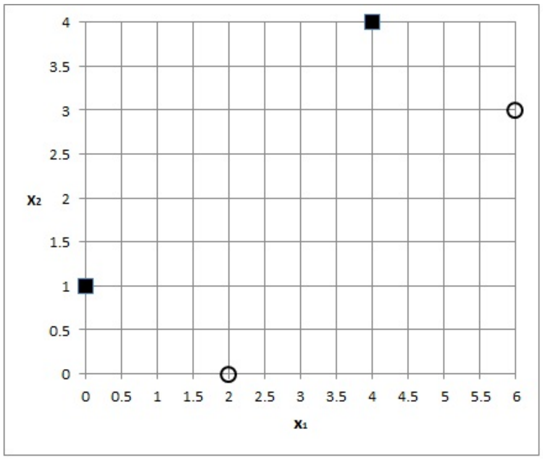

Figure 3: The training set S_1

Problem 3: Decision Trees [25 points]

Bob has to write his letters politely to various people with attaching the right title such as Mr and Ms. However, they use their cryptologic names whose genders are not evident just based on the naming conventions. Thus he tries to build a decision tree model for predicting the right gender based on the information that he has. He is willing to consider the following three attributes:

- Length: Length of hair rounded to the nearest integer value in inches. This attributes can take categorical values in $\{0, 1, 2, ..., 10\}$, but it can be binarized by asking whether or not it is less than or equal to 5".
- Weight: Positive integer value, but it can be binarized by asking whether or not it is less than or equal to 160lbs.
- Age: Integer value, but it can be binarized by asking whether or not it is less than or equal to 40 years old.

The following table consists of the training data D (9 first rows) and the test data (the last row) with the corresponding feature values.

Person	**Length?**	**Weight?**	**Age?**	**Gender?**
Alice	0 inches	250 lbs	36 years old	male
Bob	10 inches	150 lbs	34 years old	female
Charlie	2 inches	90 lbs	10 years old	male
Dave	6 inches	78 lbs	8 years old	female
Eve	4 inches	20 lbs	1 year old	female
Frank	1 inches	170 lbs	70 years old	male
Grace	8 inches	160 lbs	41 years old	female
Heidi	10 inches	180 lbs	38 years old	male
John	6 inches	200 lbs	45 years old	male
Kyle	8 inches	290 lbs	38 years old	??

(a) Create the best decision tree splitting based on the training accuracy. Justify your split by showing the training accuracy. Does the tree achieve zero training error?

(b) What is the most plausible gender for Kyle based on what you have built in (a)?

In Information Theory, **Entropy** measures the homogeneity of samples. Completely homogeneous samples have the minimum possible entropy of 0, whereas equally divided samples between two labels have the entropy of 1. For any subset of data S, it is evaluated by

$$\mathit{Entropy}(S) = -\sum_{l \in \mathit{Labels}(S)} p(l) \log_2 p(l)$$

, where $p(l)$ is the probability of that labels in S. For example in our dataset D, we have two possible labels where $p(\mathit{male}) = 5/9$ and $p(\mathit{female}) = 4/9$. Then the $\mathit{Entropy}(D) = -5/9 \log_2(5/9) - 4/9 \log_2(4/9) = 0.9911$. Information Gain (IG) is another popular metric to decide the best split. Assuming a split divides S into $\{S_1, ..., S_K\}$,

$$IG(\mathit{split}) = \mathit{Entropy}(S) - \sum_{k=1}^{K} w_k \cdot \mathit{Entropy}(S_k) \text{ where } (w_k = \frac{|S_k|}{|S|})$$

Then you can choose the attribute with the highest information gain as the next split.

(c) Create the best decision tree splitting based on the information gain. Justify your split by comparing gains for each of possible splits. Does the tree achieve zero training error?

(d) What is the most plausible gender for Kyle based on what you built in (c)?

(e) (Optional +5pts) What if we use categorical values from 0 to 10 as they are rather than binarizing Length attribute? Which would be better criterion between training accuracy and information gain? Why?

Problem 4: Linear Regression [25 points]

In RStudio, try to download ISLR packages. You will work with 'Auto' dataset, which becomes immediately available after an installation of ISLR. Verify the first 10 examples that have various attributes. Each of these could be used as an input feature (often called *predictors* in Statistics) or as an output value (often called as response).

(a) Take a look at our data. 1) What is the number of training examples m and the number of features n except the *name* attribute? 2) Thinking this data as a matrix $X \in \mathbb{R}^{m \times n}$, is X a skinny/tall matrix or fat/wide matrix?

(b) Perform the basic exploratory analysis by computing and visualizing correlations. 1) Draw the plot of correlations between every pair of features. 2) Which features are highly correlated one another? (Hints: You can use **cor** and **corrplot** if necessary.)

(c) Perform linear regression by putting *mpg* as an output variable based on all other features except the *name* attribute. 1) Is there any relationship between the input features and the output response? 2) Which features appear to have a statistically significant relationship to the output response? 3) What does the coefficient for the *year* variable suggest? (Hint: You can use **lm** if necessary)

(d) Produce diagnostic plots of the linear regression fit by using **plot**. 1) Any problems in the fit? 2) Do the residual plots suggest any unusually large outliers?

(e) (Optional +5pts) Try a few different transformations of the variables, such as $\log X$, $\sqrt{X}$, X^2. Can you make a better fit? Comment on your findings.