10-601: Homework 1
Due: 18 September 2014 11:59pm (Autolab)
TAs: Abhinav Maurya, Jingwei Shen

Name: _____

Andrew ID: _____

Please answer to the point, and do not spend time/space giving irrelevant details. You should not require more space than is provided for each question. If you do, please think whether you can make your argument more pithy, an exercise that can often lead to more insight into the problem. Please state any additional assumptions you make while answering the questions. You need to submit a single PDF file on autolab. Please make sure you write legibly for grading.

You can work in groups. However, no written notes can be shared, or taken during group discussions. You may ask clarifying questions on Piazza. However, under no circumstances should you reveal any part of the answer publicly on Piazza or any other public website. The intention of this policy is to facilitate learning, not circumvent it. Any incidents of plagiarism will be handled in accordance with CMU's Policy on Academic Integrity.

⋆: Code of Conduct Declaration

- 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*)

- 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*).

1: The truth will set you free. (TA:- Abhinav Maurya)

State whether true (with a brief reason) or false (with a contradictory example). Credit will be granted only if your reasons/counterexamples are correct.

(a) During decision tree construction, if you reach a node where the maximum information gain for a node split using any attribute is zero, then all training examples at that node have the same label.

[2 points]

(b) Whenever a set S of labeled instances is split into two sets S_1 and S_2, the average entropy will not increase, irrespective of the split attribute or the split point.

[*2 points*]

(c) A decision tree can be represented as a decision list and vice versa. (Hint: A decision list is a sequentially applied list of decision rules of the form: *If condition$_1$ and condition$_2$ and ... condition$_n$, then output is y_i.* Each condition is a test on a single feature similar to the nodes of a decison tree.)

[*2 points*]

(d) If X_1 and X_2 are independent gaussian random variables, $X = \frac{1}{4}(X_1 - X_2)$ is a gaussian random variable.

[*2 points*]

(e) If $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are the probability density functions of independent gaussian random variables, $f(X) = \frac{1}{2}\{f_X(x) + f_Y(y)\}$ is a probability density function corresponding to a gaussian random variable.

[*2 points*]

2: Maximum Likelihood Estimation. (TA:- Jingwei Shen)

(a) $X_1, X_2, \cdots, X_n$ are random variables that are uniformly distributed between $[-\theta/2, \theta/2]$, $\theta \in \mathbb{R}$. Write down the MLE for the parameter θ and explain it. (You do not have to derive it.)

[*3 points*]

(b) We have two coins - an unbiased one with probability $p_1 = 1/2$ of showing heads on a toss, and a biased one with probability $p_2 = 1/3$ for showing heads. We choose one of the two coins. With an unknown probability p, we choose the biased coin, and with probability $1 - p$, we choose the unbiased one. We toss the chosen coin 100 times and observe 40 heads during the 100 tosses. Write down the MLE estimate of parameter p and explain it. (You do not have to derive it.)

[*3 points*]

3: Three Prisoners and a Warden (TA:- Jingwei Shen)

Three prisoners - A, B, and C - are on death row. The governor decides to pardon one of the three and chooses the prisoner to pardon at random. He informs the warden of his choice but requests the name to be kept as a secret.

Having heard of the pardon rumor through grapevine, A tries to get the warden to tell him his fate. The warden refuses. Then A asks which of B or C will be executed. The warden thinks a while and tells A that B is to be executed. (Assume that the warden picks a random legal answer for A's question).

(a) Let A, B, C denote the event that A, B, C will be pardoned respectively. Let $!B$ denote the event that the warden says B will die. Compute $P(A \,|!B)$. Does the chance of A's survival increase with the additional information about B's death? (Hint: compare $P(A|!B)$ and $P(A)$).

[*3 points*]

(b) Suppose A reveals all of the above to C. Show the probability of C surviving at this time is $2/3$. (Hint: Prove $P(C|!B) = 2/3$).

[*3 points*]

4: Probability Theory (TA:- Jingwei Shen)

(a) Let A, B, C be three discrete random variables. Show that

1. $P(A \mid B, C) = \frac{P(A, B \mid C)}{P(B \mid C)}$

2. $P(A \mid C) = \sum_B P(A, B \mid C)$

3. $P(A \mid C) = \sum_B P(A \mid B, C) \cdot P(B \mid C)$

[*3 points*]

(b) Suppose that 0.5% men and 0.25% women are color-blind. A person is chosen randomly at the university where the number of men is twice of that of women. The chosen person is color-blind. What is the probability that the person is male?

[*2 points*]

(c) Consider the probability density function $f_{X,Y}(x, y)$ over a 2-dimensional random variable $[X, Y]$.

$$f_{X,Y}(x, y) = \begin{cases} c(x + y^2) & 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Here, c is a constant appropriate for $f_{X,Y}(x, y)$ to be a density function. Find $P(X < \frac{1}{2} \mid Y = \frac{1}{2})$
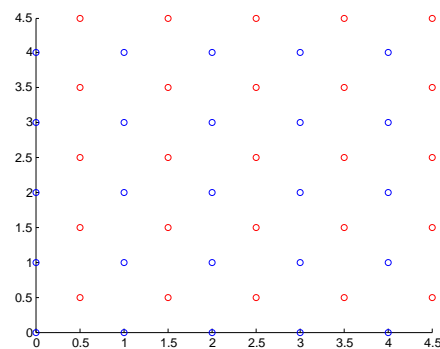
[*3 points*]

5: Nearest neighbors to the rescue. (TA:- Jingwei Shen)

(a) Consider two classes C_1, C_2 in the two-dimensional space. The data from class C_1 are uniformly distributed in a circle of radius r. The data from class C_2 are uniformly distributed in another circle of radius r. The centers of two circles are at a distance greater than $4r$. Show that the accuracy of 1-NN is greater than or equal to the accuracy of k-NN, where k is an odd integer and $k \geq 3$.

[*3 points*]

Figure 1: Q4 Dataset

(b) In the dataset shown in figure 1, what is the leave-one-out accuracy of the k-NN method when $k = 2$? Remember that a data point cannot be considered its own neighbor since it is left out.

[*2 points*]

(c) In this problem, explain briefly why you think k-NN performs worse than randomly guessing, which has an accuracy near 50%?

[2 points]

6: A tree about the important things in life. (TA:- Abhinav Maurya)

The following dataset will be used to learn a decision tree for predicting whether a person is Happy (H) or Sad (S) based on the color of their shoes, whether they wear a wig and the number of ears they have.

Color	Wig	Num. Ears	Emotion
G	Y	2	S
G	N	2	S
G	Y	2	S
B	N	2	S
B	N	2	H
R	N	2	H
R	N	2	H
R	N	2	H
R	Y	3	H

(a) What is Entropy(Emotion | Wig=Y)?

[1 points]

(b) Which attribute would the decision-tree building algorithm choose to use for the root of the tree (assume no pruning)?

[2 points]

(c) Draw the full decision tree that would be learned for this data (assume no pruning).

[3 points]

(d) What would be the training set error for this dataset? Express your answer as the percentage of records that would be misclassified.

[2 points]

7: Digging up the dense binary tree. (TA:- Abhinav Maurya)

Consider the following data with three binary attributes, where x^i denotes the ith datapoint, x_j denotes the jth feature of the datapoint, and y denotes the class label:-

	x_1	x_2	x_3	y
x^0	0	0	0	0
x^1	0	0	1	1
x^2	0	1	0	1
x^3	0	1	1	0
x^4	1	0	0	1
x^5	1	0	1	0
x^6	1	1	0	0
x^7	1	1	1	1

(a) Draw the decision tree for the above dataset using the entropy criterion to decide node splits (assume no pruning).

[3 points]

(b) Decision trees are often pruned so that they can better generalize for prediction on the test set. Do you think you could prune any of the lower levels of the above decision tree used to predict the XOR of 3 binary digits? Give reasons for your decision.

[2 points]

(c) Considering a generalization of the above problem, let's say that we train a decision tree without any pruning to output the XOR function using *all* possible binary strings of length n. Out of the decision tree and KNN classifier (using l_1 distance and $k = 1$), which one would be more accurate when the test samples are also binary strings of length n?
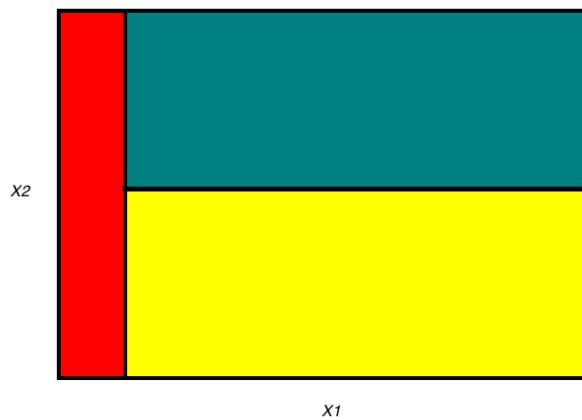
[2 points]

(d) Out of the decision tree and KNN classifiers considered in the previous question, which one will take lesser time to predict the output label of a new test datapoint? Why? (Hint: Note that there are 2^n possible datapoints due to n binary input features. Consider the number of nodes traversed by the decision tree and the number of distance computations performed by the KNN classifier to predict the label of a test datapoint with n binary input features.)

[2 points]

8: On the hardness of learning optimal binary decision trees (TA:- Abhinav Maurya)

In figure 2, assume that the rectangular region consisting of two features x_1 and x_2 is densely packed with points. The red, green, and yellow subrectangles represent the three classes C_1, C_2, and C_3 of datapoints. The x_1 X x_2 dimensions of the red, green, and yellow rectangles are 1 X 6, 7 X 3, and 7 X 3 respectively. The red rectangle is uniformly populated with 6,000 datapoints of class C_1. The green rectangle is uniformly populated with 42,000 datapoints of class C_2. The yellow rectangle is uniformly populated with 42,000 datapoints of class C_3.

Figure 2: A 2D dataset with three classes

(a) What is the minimum number of nodes that a decision tree needs to have in order to classify the above dataset correctly?

[*2 points*]

(b) What is the number of nodes in the decision tree trained on the above dataset using the entropy criterion?

[*2 points*]

(c) Are the number of nodes in the two cases identical or different? Why do you think that is?

[*3 points*]

(d) Construct another toy dataset where the entropy gain criterion leads to a suboptimal decision tree i.e. one with more nodes than another tree of comparable accuracy. Your dataset should have at least four labels and be sufficiently different from the given toy dataset.

[3 points]

(e) For your suggested dataset, draw the optimal decision tree as well as the decision tree obtained using the entropy minimization criterion.

[3 points]

(f) A decision tree can classify the dataset in figure 2 with 100% test accuracy (assuming that there is no label noise). What are the general conditions on a dataset under which a decision tree can provide 100% test accuracy? (Hint: Each internal node of a decision tree performs a split based on a single feature. Think about the class of separation functions such a decision tree entails.)

[3 points]

$$\boxed{\textbf{Total: 70}}$$