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*]

False. In the dataset of question 7, split at the root on any of the three attributes gives an information gain of zero. However, the root node clearly does not have datapoints with the same label.

(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*]

True. Splitting a node on an attribute always provides additional information and never leads to loss of discriminatory information. Hence, entropy must never increase on any split. (Any algebraic proof works as well.)

(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*]

True. Each path from the root of a decision tree to a leaf provides a decision rule. Also, decision rules can be merged to form a decision tree.

(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*]

True. Any linear combination of a finite number of Gaussian random variables is also a Gaussian random variable. (Any algebraic proof works as well.)

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

[*2 points*]

False. The resulting probability density function is bimodal in general and therefore does not correspond to a gaussian random variable whose density is quasiconcave and has a single maximum occurring at the mean.

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*]

The likelihood funtion:

$$l(\theta) = \frac{1}{\theta^n} I_{|X^{(1)}| \leq \theta/2}$$

where $X^{(1)}$ denotes the observation that has largest absolute value. $I_{|X^{(1)}| \leq \theta/2} = \begin{cases} 1 \text{ if } |X^{(1)}| \leq \theta/2 \\ 0 \text{ otherwise} \end{cases}$

The likelihood function value increases as θ decreases, so the mle is the smallest possible value of θ, which is $2X^{(1)}$.

(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 do 100 tosses. Each time 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. And we 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]

The probability to observe a Head in one experiment is $q = p/3 + (1-p)/2$. Since the 100 observations are iid, the total number of Heads follows a binomial distribution $B(100, q)$. The log likelihood function :

$$\begin{aligned}\log(l(p)) &= 40\log(q) + 60\log(1-q) + C \\ &= 40\log(1/2 - 1/6p) + 60\log(1/2 + 1/6p) + C\end{aligned}$$

where C is a constant unrelated to p.

Take the derivative w.r.t p, and make it zero.

$$\frac{40}{1/2 - 1/6p} \times (-1/6) + \frac{60}{1/2 + 1/6p} \times 1/6 = 0$$

We solve $p = 0.6$.

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]

To be pardoned	warden's answer
A (1/3)	B (1/2) or C (1/2)
B (1/3)	C
C (1/3)	B

So

$$P(A|!B) = \frac{P(A,!B)}{P(!B)} = \frac{P(A,!B)}{P(A,!B) + P(B,!B) + P(C,!B)} = \frac{1/6}{1/6 + 0 + 1/3} = 1/3$$

(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*]

Similar to (a),

$$P(C|!B) = \frac{P(C,!B)}{P(!B)} = \frac{1/3}{1/2} = 2/3$$

4: Probability Theory (TA:- Jingwei Shen)

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

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

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

2.
$$P(A|C) = \frac{P(A,C)}{P(C)} = \frac{\sum_B P(A,B,C)}{P(C)} = \sum_B \frac{P(A,B,C)}{P(C)} = \sum_B P(A,B|C)$$

3.
$$P(A,B|C) = \frac{P(A,B,C)}{P(C)} = \frac{P(A|B,C)P(B,C)}{P(C)} = P(A|B,C)P(B|C)$$

From the above question we know that $P(A \mid C) = \sum_B P(A, B \mid C)$. So simply plug in, and we get $P(A \mid C) = \sum_B P(A \mid B, C) \cdot P(B \mid C)$

(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*]

Let M denotes that a person is a male, and F denotes the person is a female. B denotes the event that a person is color blind.

$$P(M|B) = \frac{P(M,B)}{P(B)} = \frac{P(B|M)P(M)}{P(B|M)P(M) + P(B|F)P(F)} = \frac{P(B|M)}{P(B|M) + P(B|F)P(F)/P(M)}$$
$$= \frac{0.5\%}{0.5\% + 0.25\% \times 0.5} = 0.8$$

(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 $P(X < \frac{1}{2}|Y = \frac{1}{2})$

[3 points]

Marginal probability for Y: $p_Y(y) = \int_X f(x,y)dx = c(1/2 + y^2)$.

So the conditional density function :

$$p(x|y) = \frac{p(x,y)}{p(y)}$$
$$= \frac{c(x+y^2)}{c(1/2+y^2)}$$
$$= \frac{x+y^2}{1/2+y^2}$$

$P(X < 1/2|Y = 1/2) = \int_0^{1/2} p(x|y = 1/2)dx = \int_0^{1/2} \frac{x+1/4}{1/2+1/4} = 1/3$

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

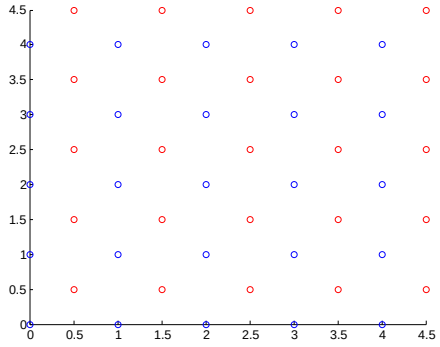
(a) Consider two classes C_1, C_2 in the two-dimensional space. The data from class C_1 are uniformly distributed in a circle of radius r. The data from class C_2 are uniformly distributed in another circle of radius r. The centers of two circles are at a distance greater than $4r$. Show that the accuracy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

[3 points]

The distance of two points in one circle will be shorter than that of two points from two different circles. So as long as we have training data points from both circles, 1-NN will give 100% accuracy while k-NN may have error predictions. For example, if we have 10 pints from circle 1 and 20 points from circle 2, and if we choose $k = 21$, then the 21 nearest neighbors of any point in circle 1 will have 10 points from circle 1 and 11 from circle 2, which results in a wrong prediction.

(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. (Ignore the datapoints that have an output tie for $k = 2$ nearest neighbors.)

Figure 1: Q4 Dataset



[*2 points*]

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

[*2 points*]

k-NN methods classify data points according to the distance. In the example of 4(b), the distance of two points doesn't contain any information about the classification. So k-NN fails.

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*]

$$\text{Entropy(Emotion | Wig=Y)} = -\frac{1}{3} * \log(\frac{1}{3}) + -\frac{2}{3} * \log(\frac{2}{3})$$

$$= \frac{1}{3} * \log(3) + \frac{2}{3} * \log(\frac{3}{2}) = 0.918$$

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

[*2 points*]

Information gains on various attributes are as follows:

$$IG(Color) = 0.76885$$
$$IG(Wig) = 0.07278$$
$$IG(Ears) = 0.10219$$

Hence, assuming entropy based information gain to be the goodness of an attribute, we would use color for a split at the root of the tree.

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

[*3 points*]

The first split is on color. For branch corresponding to color B, neither wig nor number of ears offers any further information gain. Hence, we can assign a label of S or H to the leaf at the end of the branch for color B, and stop building the tree.

(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*]

One example in the training set would be misclassified (either 4^{th} or 5^{th} example in the dataset depending on the tree), leading to a train error of 11.11 percent.

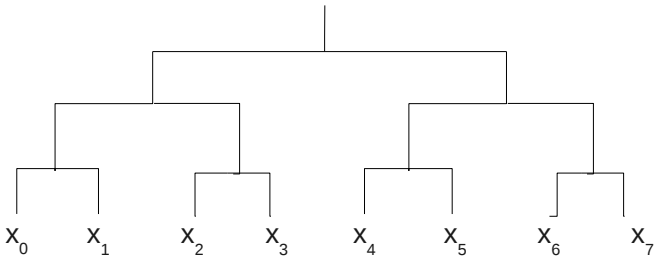
7: Digging up the dense binary tree. (TA:- Abhinav Maurya)

Consider the following data with three binary attributes, where x^i denotes the i^{th} datapoint, x_j denotes the j^{th} 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*]

Removing any of the decision nodes would decrease the accuracy of the decision tree in predicting XOR function. Hence, we cannot prune the tree at all.

(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*]

Given a dataset of all n-digit binary strings and XOR as the binary output, both the decision tree (without pruning) and KNN algorithm can be trained to provide 100% accuracy on the classification problem. Hence, the two algorithms have comparable accuracy in this case.

(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.)

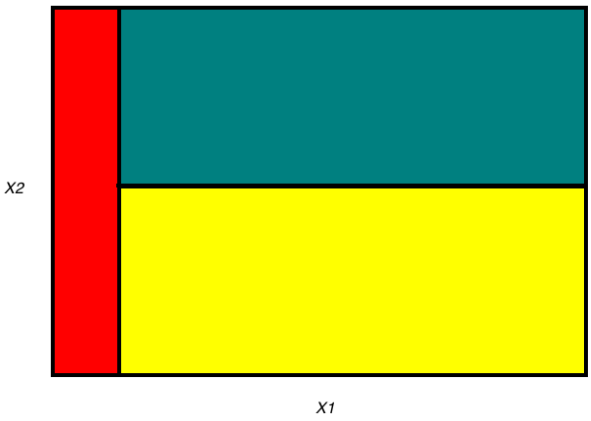
[*2 points*]

For any binary string of length n, a decision tree does n tests and therefore takes $O(n)$ time. KNN computes the distance to all the datapoints in the training set of size 2^n and therefore takes time $O(2^n)$. Hence, a decision tree is asymptotically much faster.

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*]

The optimal decision tree needs two internal nodes i.e. attribute tests to classify the given dataset correctly. It first splits on x_1 and then on x_2.

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

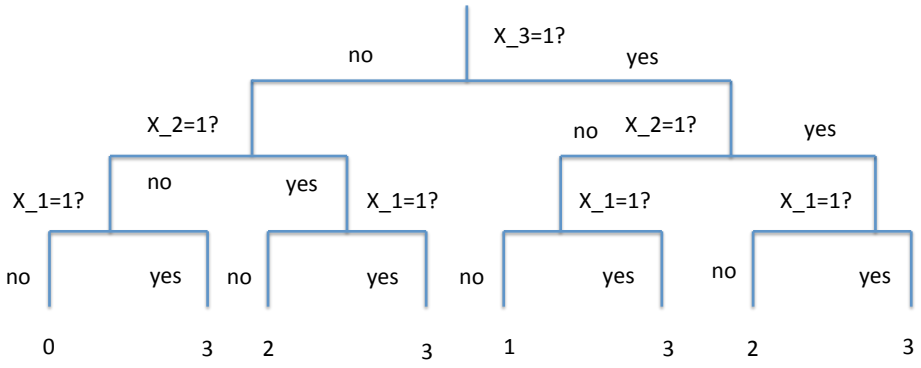
[*2 points*]

The decision tree obtained using entropy minimization criterion needs three internal nodes i.e. attribute tests to classify the given dataset correctly. It first splits on x_2, and then makes two splits on x_1 to separate red-green and red-blue datapoints.

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

[*3 points*]

Figure 3: Decision Tree using Entropy Minimization



The number of nodes are different in the two cases because the entropy minimization criterion is a heuristic used by a greedy decision tree building algorithm, which cannot yield the optimal decision tree.

(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*]

Consider the below dataset with three binary attributes, where x^i denotes the i^{th} datapoint, x_j denotes the j^{th} feature of the datapoint, y denotes the class label, and n denotes the number of copies of the datapoint.

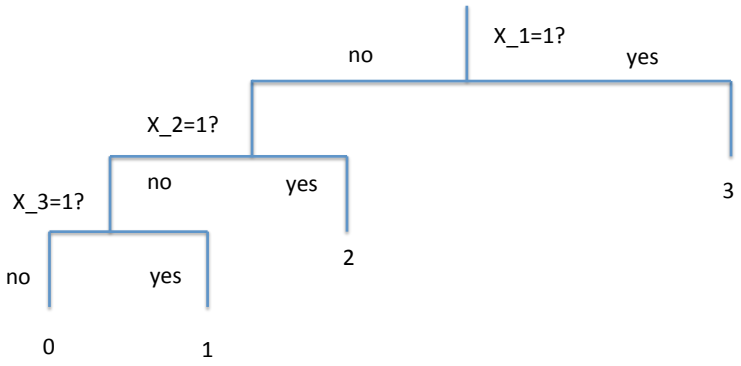
	n	x_1	x_2	x_3	y
x^0	5000	0	0	0	0
x^1	6000	0	0	1	1
x^2	20	0	1	0	2
x^3	10	0	1	1	2
x^4	1	1	0	0	3
x^5	1	1	0	1	3
x^6	1	1	1	0	3
x^7	1	1	1	1	3

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

[*3 points*]

See figures 3 and 4.

Figure 4: Optimal Decision Tree



(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]

The separation boundaries between the various classes need to be axis-aligned.

OR

The datapoints of each class can be contained in a finite number of axis-aligned rectangles.

(There are many ways of saying this. You will get full credit as long as your condition is effectively same as this one.)

Total: 70