10-601: Homework 1

Due: 18 September 2014 11:59pm (Autolab)
TAs: Abhinav Maurya, Jingwei Shen

Name:Yan Zhao
Andrew ID: yanzhao2@andrew.cmu.edu

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.
- If you answered *yes*, give full details: *Kang explained to me what is asked in Question 7.(c)(d)*
- 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*]

This statement is false.
Given the example bellow, every attribute has an entropy of $-\frac{1}{2} \times \log_2(\frac{1}{2}) - \frac{1}{2} \times \log_2(\frac{1}{2})$. So, no matter using any attribute the information, the infomation gain is zero, but not all training example have same label.

x_1	x_2	y
0	0	1
1	1	1
0	1	0
1	0	0

(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*]

This statement is true
Assuming that Y is class group we want to predict before spiltting, X is attributes.
$I(X,Y) = H(Y) - H(Y \mid X)$. Because $I(X,Y) >= 0$, then $H(Y) >= H(Y \mid X)$, So for each category s_i of S, $H(Y) >= H(Y|s_i)$. Thus, the average entropy of categories divided by s_1 and s_2 is $H(Y \mid s_1)$ and $H(Y \mid s_2)$, are both smaller than $H(Y)$. Thus, the average entropy will smaller than $H(Y)$.

(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 decision tree.)

[*2 points*]

This statement is true.
Cause in any decision tree, we can get the final prediction by traversing one path from root to bottom. And in every node on this path, there is one decision. Thus, every path of a decision tree from root to bottom can be transformed into a decision list with a sequentially applied list of decision rules of the form: *If condition$_1$ and condition$_2$ and ... condition$_n$, then output is y_i.*

(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*]

This statement is true Given X_1 and X_2 are independent gaussian random variables, according to the property of gaussian function, if If X_1 and X_2 are independent gaussian random variables, their linear combination is also follow gaussian distribution. So, we can get new gaussian variables whose mean is $\frac{(\mu_1 - \mu_2)}{4}$, variance is $\frac{(\sigma_1^2 + \sigma_2^2)}{16}$.

(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*]

This statement is false.
Suppose that $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are two gaussian functions with different Expectations $E1$ and $E2$, but the same Variance.Then for function $f(x)$, when $x = E1$ and $x = E2$, $f(x)$ gets to two peaks, which is contradict to the property of Gaussian functions.

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*]

$$\mathcal{L}(\theta; X_1, X_2....X_n) = f(X_1, X_2...X_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = (\frac{1}{\theta})^n$$

$\theta_0 = argmax\{\mathcal{L}\}$, which means to find the value of θ that maximize $\mathcal{L}$. Since $\theta/2 \geqslant \theta_1 = max\{X_1, X_2....X_n\}$ and $-\theta/2 \leqslant - \mid \theta_2 \mid = - \mid min\{X_1, X_2....X_n\} \mid$ Let $\theta_0 = max\{\theta_1, \theta_2\}$, then θ_0 is the value of θ that can maximize $\mathcal{L}$

(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*]

$$\mathcal{L} = \binom{100}{40}(p \cdot 1/3 + (1-p) \cdot 1/2)^{40} \cdot (p \cdot 2/3 + (1-p) \cdot 1/2)^{60}$$

We need to find a p that can maximize $\mathcal{L}$. To do this, we just need to find the p that can maximize $40 \cdot ln(\frac{3-p}{6}) + 60 \cdot ln(\frac{3+p}{6})$, which means $\frac{6 \cdot 40}{3-p} = \frac{6 \cdot 60}{3+p}$, and get $p = 0.6$

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 $P(A \mid !B)$. 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*]

The warden's answer is legal. So on condition of A is pardoned, the probability that the warden says $!B$ is 1/2; on condition of B is pardoned, the warden won't say $!B$, so its probability is 0; on condition of C is pardoned, the warden will definitely say $!B$, so its probability is 1.

$$P(!B) = P(!B|A) \cdot P(A) + P(!B|B) \cdot P(B) + P(!B|C) \cdot P(C) = \frac{1}{2} \times \frac{1}{3} + 0 + 1 \times \frac{1}{3} = \frac{1}{2}$$

$$P(A|!B) = \frac{P(!B|A) \cdot P(A)}{P(!B)} = \frac{P(!B|A) \cdot P(A)}{P(!B|A) \cdot P(A) + P(!B|B) \cdot P(B) + P(!B|C) \cdot P(C)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

(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*]

Because A and C all know B is going to die, if A is about to die, C must be pardoned, vice versa. Then A and C are complementary events. Thus $P(A|!B) + P(C|!B) = P(!B) = 1$. Now $P(A|!B) = \frac{1}{3}, P(C|!B) = 1 - P(A|!B) = \frac{2}{3}$

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. $P(A \mid (B, C)) = \frac{P(A, B, C)}{P(B, C)} = \frac{P((A, B) \mid C) \cdot P(C)}{P(B|C) \cdot P(C)} = \frac{P(A, B \mid C)}{P(B \mid C)}$
2. $\sum_B P(B \mid A, C) = 1$, $\sum_B P(B \mid A, C) = \frac{\sum_B P(A, B, C)}{P(A, C)} = \frac{\sum_B P(A, B \mid C) \cdot P(C)}{P(A, C)} = 1$, $\sum_B P(A, B \mid C) = \frac{P(A, C)}{P(C)} = P(A \mid C)$
3. $\sum_B P(B \mid A, C) = \frac{\sum_B P(A, B, C)}{P(A, C)} = 1$, so $\sum_B P(A, B, C) = \sum_B P(A \mid B, C) \cdot P(BC) = P(AC)$, with $P(BC) = \frac{P(B \mid C)}{P(C)}$ and $P(AC) = \frac{P(A \mid C)}{P(C)}$, 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*]

Assuming that M is a random variable denote men. F is a random variable denote women. B is a random variable denote color-blind.
Thus,$P(M) = \frac{2}{3}, P(F) = \frac{1}{3}, P(B \mid M) = \frac{1}{200}, P(B \mid F) = \frac{1}{400}$

$$P(M \mid B) = \frac{P(M,B)}{P(B)} = \frac{P(B \mid M) \cdot P(M)}{P(B \mid M) \cdot P(M) + P(B \mid F) \cdot P(F)} = \frac{\frac{1}{200} \cdot \frac{2}{3}}{\frac{1}{200} \cdot \frac{2}{3} + \frac{1}{400} \cdot \frac{1}{3}} = \frac{4}{5}$$

(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*]

$$\int_0^1 \int_0^1 (c \cdot (x+y^2)) dx dy = c \int_0^1 (\frac{1}{2} + y^2) dy = c(\frac{1}{2} + \frac{1}{3}) = 1, c = \frac{6}{5}$$

,

$$P(X < \frac{1}{2} \mid Y = \frac{1}{2}) = \int_0^{1/2} (c \cdot x + \frac{1}{4}) dx = \frac{6}{5} \cdot (\frac{1}{2} \cdot (\frac{1}{2})^2 + \frac{1}{4} \cdot \frac{1}{2}) = \frac{3}{10}$$

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*]

Assuming that C_1 has n_1 training data, C_2 has n_2 training data.

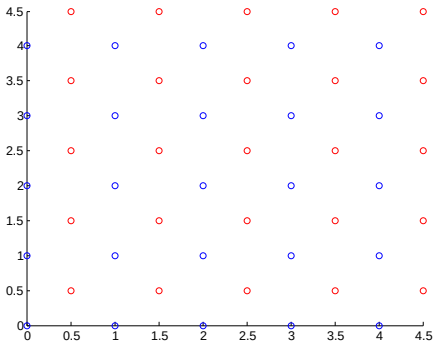
1. If any of the class has no training data($n_1 = 0$ or $n_2 = 0$), K-NN classifier would be wrong no matter what the value of k is.

2. Else, For 1-NN, the nearest neighbor of a test data must be in the same class, cause the smallest distance between two cycles is $2r$. So the accuracy of 1-NN is 100%.

3. As for the k-NN, cause k is an odd integer, assume the test data is in C_1, if $n_1 > k/2$ and $n_2 < k/2$, the accuracy of k-NN is 100%, the same accuracy of 1-NN; if $n_1 < k/2$ and $n_2 > k/2$, the accuracy of k-NN is 0%, wrong. It is the same when test data is in C_2

4. In conclusion, the the accuracy of 1-NN is greater than or equal to the accuracy of k-NN

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

[*2 points*]

The leave-one-out accuracy of the k-NN method when $k = 2$ is 0. From the picture shown above, as for a RED test data point, its 4 nearest neighbors are all opposite(BLUE) data points. It is the same with a BLUE test data point. Because of the requirements for leave-one-out, we cannot choose the test data point itself. Thus, for $k = 2$, the 2 nearest data points are all with an opposite Color.

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

[*2 points*]

As is said in above question, for any test data point, 2 nearest neighbors are all opposite data points, making the accuracy of k-NN method 0. However, if we randomly guess, since the number of RED test data point and BLUE test data are equal, we can get a near 50% accuracy. Thus, k-NN performs worse than randomly guessing

---

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

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

they have.

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

[*1 points*]

$$P(Emotion = S \mid Wig = Y) = \frac{2}{3}, P(Emotion = H \mid Wig = Y) = \frac{1}{3}$$

$$Entropy(Emotion \mid Wig = Y) = -\sum_{Emotion} P(Emotion \mid Wig = Y) \cdot \log_2(P(Emotion \mid Wig = Y))$$

$$= -\frac{2}{3} \cdot \log_2(\frac{2}{3}) - \frac{1}{3} \cdot \log_2(\frac{1}{3}) = 0.918$$

.

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

[*2 points*]

We choose "Color" as the root of the decision tree. Cause "Ears" must have a much smaller info gain than any of the other attributes, and compared with "Colors" and "Wig", "Colors" has the largest info gain.

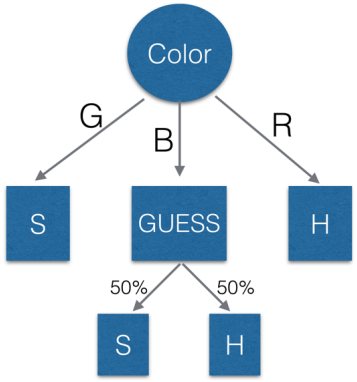
$$Entropy(Emotion) = -\frac{4}{9} \times \log_2(\frac{4}{9}) - \frac{5}{9} \times \log_2(\frac{5}{9}) = 0.918$$

$$InfoGain(Color) = Entropy(Emotion) - \frac{2}{9} \times (-\frac{1}{2} \times \log_2(\frac{1}{2}) - \frac{1}{2} \times \log_2(\frac{1}{2})) = 0.696$$

$$InfoGain(Wig) = Entropy(Emotion) - \frac{1}{3} \times (-\frac{1}{3} \times \log_2(\frac{1}{3}) - \frac{2}{3} \times \log_2(\frac{2}{3})) - \frac{2}{3} \times (-\frac{1}{3} \times \log_2(\frac{1}{3}) - \frac{2}{3} \times \log_2(\frac{2}{3})) = 0.029$$

(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]

For this training set, when Color is B, we have 50% to misclassify. So the total training set error is $\frac{1}{9}$

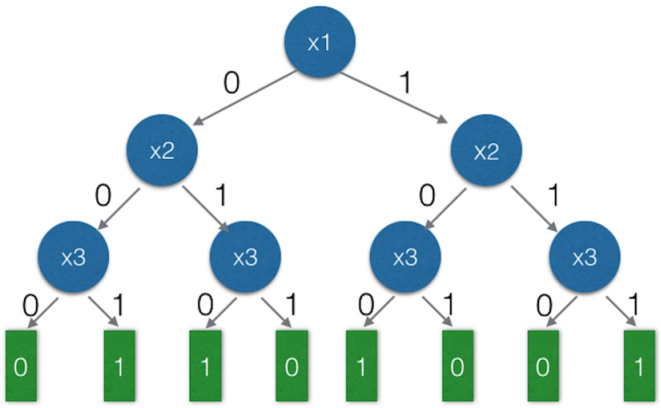
7: Digging up the dense binary tree. (TA:- Abhinav Maurya)

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*]

We could not prune any of the lower levels of above decision tree. Cause for XOR operation, every attribute of x_1, x_2, x_3 is equal weighted, all decide the result of prediction. If we prune any of these levels on the decision tree, we will get the wrong prediction. And only with all the 7 nodes in decision tree, can we predict correctly.

(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*]

The decision tree and KNN classifier with $k = 1$ have an equal accuracy. Cause according to the questions above, a decision tree without any pruning has a 100% accuracy on prediction. For KNN classifier, based on l_1 distance, $k = 1$, that means we choose the closest binary string, with a distance of 0, which is the string itself. We already get the whole instances of training set, so the prediction based on string itself is 100% accuracy. Thus, decision tree and KNN classifier have an equal accuracy.

(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*]

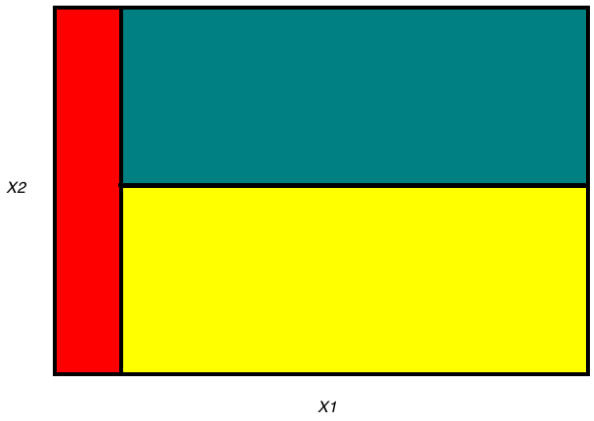
The decision tree will cost less time to predict. Cause for a decision tree, the height of the tree is the number of binary input features, which is n. As for the KNN classifier, we need $O(2^n)$

steps(may be only one step in optimal situation, but $O(2^n)$ on average) to detect if one string s has a 0 distance with the test datapoint.

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*]

Two nodes. One is based on x_1, for $x_1 < 1$, to classify C_1 from the rectangular region; and for $x_1 > 1$, use x_2 to classify C_2 and C_3 from the left region.

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

[*2 points*]

We need three nodes in the decision tree. One is to cut horizontal, dividing the rectangular region into two parts based on x_2. Then for $x_2 < 3$, use x_1 to identify class $C_1$and C_3, for $x_1 < 1$ is $C_1$and for $X_1 > 1$ is C_3; Then for $x_2 > 3$, use x_1 to identify class $C_1$and C_2, for $x_1 < 1$ is $C_1$and for $X_1 > 1$ is C_2

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

[*3 points*]

Cause using entropy criterion, if we classify C_1 first, the region of C_2 and C_3 are of the same size, meaning the entropy very large. Thus, to minimize entropy to get a maximum information gain, we need to cut horizontal, which creates one more nodes than the optimal situation.

$$Entropy(Cut-Vertical) = \frac{7}{8} \times (-\frac{1}{2} \times \log_2(\frac{1}{2}) - \frac{1}{2} \times \log_2(\frac{1}{2})) = 0.875$$

$$Entropy(Cut-Horizontal) = \frac{1}{2}\times(-\frac{1}{8}\times\log_2(\frac{1}{8})-\frac{7}{8}\times\log_2(\frac{7}{8}))+\frac{1}{2}\times(-\frac{1}{8}\times\log_2(\frac{1}{8})-\frac{7}{8}\times\log_2(\frac{7}{8})) = 0.54$$

(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*]

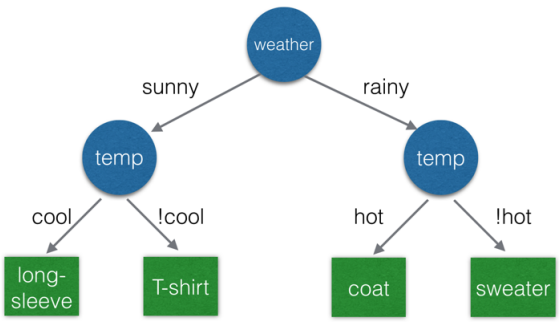
The toy dataset is to predict which clothes to wear based on the weather and the temperature.

Weather	Temperature	Clothes
sunny	hot	T-shirt
rainy	mild	sweater
sunny	cool	long-sleeve
rainy	cool	sweater
sunny	mild	T-shirt
rainy	hot	coat
sunny	cool	long-sleeve
rainy	hot	coat
sunny	mild	T-shirt
rainy	hot	coat
sunny	cool	long-sleeve

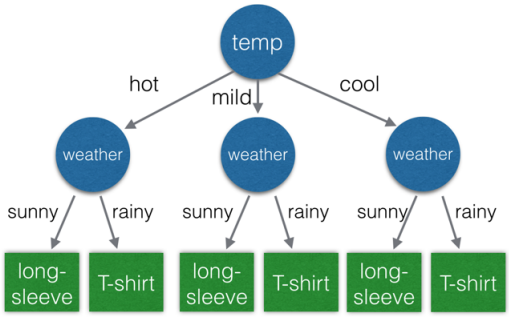
(e) For your suggested dataset, draw the optimal decision tree as well as the decision tree obtained using the entropy minimization criterion.

[*3 points*]

Using entropy minimization criterion, the attributes "Temperature" would be first split, cause it has a more information gain.
optimal decision tree shown as below



decision tree obtained using the entropy minimization criterion shown as below



(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*]

There are mainly two conditions for a decision tree to provide 100% test accuracy

1. There is no noise in the training data. For example, for two instances, if all the attributes have the same value, then the result should be the same. If not, that means the training data is contaminated by noises, will influence the accuracy of decision tree.
2. For different attributes, they must be independence, meaning no correlation between different attributes. If not, for the example above, the Class will present as an arbitrary shape rather than an rectangular. Thus, it cannot be identified as decision tree.

Total: 70