Xiaodong Wang xwang322@wisc.edu

CS 760 Machine Learning Homework #3

Name: Xiaodong Wang
Email: xwang322@wisc.edu
Student Id: 9066383432

1. (a) Mutual information Gain about the class by knowing whether or nor the value of C is less than 475:

$$I(X \mid Y)$$
$$= H(Y) - H(Y \mid X)$$
$$H(Y) = -0.4 * \log_2 0.4 - 0.6 * \log_2 0.6$$
$$= 0.97095$$
$$H(Y \mid X) = P(C < 475) * H(Y \mid C < 475) + P(C >= 475) * H(Y \mid C >= 475)$$
$$= 0.4 * (-0.5 * \log_2 0.5 - 0.5 * \log_2 0.5) + 0.6 * (-\frac{2}{3} * \log_2 \frac{2}{3} - \frac{1}{3} * \log_2 \frac{1}{3})$$
$$= 0.4 + 0.55098$$
$$= 0.95098$$
$$I(X \mid Y) = 0.0200$$

In class, "Benign" is 40% while "Malignant" is 60%, when calculating H(Y|X), we need to have probability of attribute C less than 475 and greater than 475. Also, for each entropy calculation, we need to get counts when C larther 475, how many instances goes to "benign", how many goes to "Malignant" and their corresponding percentage. Similar things to C greater than 475.

(b) Similar procedures here as above, but when AB is not different, all of them class is "Malignant" and when AB is different, they are all "Benign", so there is no entropy.

$$I(X \mid Y)$$
$$= H(Y) - H(Y \mid X)$$
$$H(Y) = -0.4 * \log_2 0.4 - 0.6 * \log_2 0.6$$
$$= 0.97095$$
$$H(Y \mid X) = P(ABdifferent) * H(Y \mid ABdifferent) + P(ABnotdifferent) * H(Y \mid ABnotdifferent)$$
$$= 0.4 * (-1 * \log_2 1) + 0.6 * (-1 * \log_2 1)$$
$$= 0$$
$$I(X \mid Y) = 0.9710$$

2. By choosing LOOCV, every time we are choosing 6 instances as 5 of them is training set and one left is test set, we are following the procedures mentioned in the lecture notes by applying Manhattan distance:
For k = 1:3

Xiaodong Wang xwang322@wisc.edu

For i = 1:6

Learn the model using all folds but Si, evaluate accuracy on Si

The result is as follow:

We set up a table of distance between each node:

instance	1	2	3	4	5	6
1	0	3	4	4	6	7
2	3	0	1	3	5	4
3	4	1	0	4	6	5
4	4	3	4	0	2	3
5	6	5	6	2	0	1
6	7	4	5	3	1	0

When k = 1

1. When S1 is test set, we choose S2 from training set, the prediction is “positive” and the real is “positive”, predict is correct;
2. When S2 is test set, we choose S3 from training set, the prediction is “negative” and the real is “positive”, predict is incorrect;
3. When S3 is test set, we choose S2 from training set, the prediction is “positive” and the real is “negative”, predict is incorrect;
4. When S4 is test set, we choose S5 from training set, the prediction is “negative” and the real is “positive”, predict is incorrect;
5. When S5 is test set, we choose S6 from training set, the prediction is “negative” and the real is “negative”, predict is correct;
6. When S6 is test set, we choose S5 from training set, the prediction is “negative” and the real is “negative”, predict is correct;

When k = 2

1. When S1 is test set, we choose S2, S3 from training set, the prediction is “positive” and the real is “positive”, predict is correct;
2. When S2 is test set, we choose S1, S3 from training set, the prediction is “correct” and the real is “positive”, predict is correct;
3. When S3 is test set, we choose S1, S2 from training set, the prediction is “positive” and the real is “negative”, predict is incorrect;
4. When S4 is test set, we choose S2, S5 from training set, the prediction is “positive” and the real is “positive”, predict is correct;
5. When S5 is test set, we choose S4, S6 from training set, the prediction is “positive” and the real is “negative”, predict is incorrect;
6. When S6 is test set, we choose S4, S5 from training set, the prediction is “positive” and the real is “negative”, predict is incorrect;

When k = 3

1. When S1 is test set, we choose S2, S3, S4 from training set, the prediction is “positive” and the real is “positive”, predict is correct;
2. When S2 is test set, we choose S1, S3, S4 from training set, the prediction is “positive” and the real is “positive”, predict is correct;
3. When S3 is test set, we choose S1, S2, S4 from training set, the prediction is “positive” and the real is “negative”, predict is incorrect;

4. When S4 is test set, we choose S2, S5, S6 from training set, the prediction is "negative" and the real is "positive", predict is incorrect;
5. When S5 is test set, we choose S2, S4, S6 from training set, the prediction is "positive" and the real is "negative", predict is incorrect;
6. When S6 is test set, we choose S2, S4, S5 from training set, the prediction is "positive" and the real is "negative", predict is incorrect;

Accuracy for k =1 and k = 2 is the same, we choose k =1 because of Occam's razor.

3. (a) (i). Mutual information between Z and X:

$$I(X, Z) = \sum_{x \in values(X)} \sum_{z \in values(Z)} P(x,z) \log_2 \frac{P(x,z)}{P(x)P(z)}$$

$$= 0.38 * \log_2 \frac{0.38}{0.5 * 0.55} + 0.12 * \log_2 \frac{0.12}{0.5 * 0.45}$$

$$+0.17 * \log_2 \frac{0.17}{0.5 * 0.55} + 0.33 * \log_2 \frac{0.33}{0.5 * 0.45}$$

$$=0.17730-0.10883-0.11796+0.18234$$

$$=0.1328$$

0.38 is the percentage where both X and Z are positive, 0.12 is the percentage where X is true while Z is false. 0.17 is percentage where X is F and Z is T and 0.33 is both of them are F. The percentage of X is true is 0.5 and Z is true is 0.55.

(ii) Similar procedures here as above, for mutual information between Z and Y.

$$I(Y, Z) = \sum_{y \in values(Y)} \sum_{z \in values(Z)} P(y,z) \log_2 \frac{P(y,z)}{P(y)P(z)}$$

$$= 0.45 * \log_2 \frac{0.45}{0.5 * 0.55} + 0.05 * \log_2 \frac{0.05}{0.5 * 0.45}$$

$$+0.10 * \log_2 \frac{0.10}{0.5 * 0.55} + 0.40 * \log_2 \frac{0.40}{0.5 * 0.45}$$

$$=0.31972-0.10850-0.14594+0.33203$$

$$=0.3973$$

0.45 is the percentage where both Y and Z are positive, 0.04 is the percentage where Y is true while Z is false. 0.10 is percentage where Y is F and Z is T and 0.40 is both of them are F. The percentage of Y is true is 0.5 and Z is true is 0.55.

(b) Based on the result from (a), Y would become candidate for parent of Z because the higher value of mutaul information.

(c) By definition of netwoek, we assume the probability of X is the root, Y is X's child and Z is Y's child.

The CPT will be like this:

Xiaodong Wang xwang322@wisc.edu

P(X) = 0.5

X	Probability(Y)
True	0.8
False	0.2

This result if obtained by when X is true (50 samples), Y is true, there are 40 cases, the P = 0.8, and for when X is false, Y is true, there are 50, 10 cases respectively, so probability is 0.2

	Probability (Z)
Y=true	0.9
Y=false	0.2

The result is obtained as follow: when Y is true (50 cases), Z has 45 which is true, so the probability is 90%, and when Y is false, Z is true there are 10 cases, so the probability is 20%.

(d) The calculation procedure based on requirement is as follow:

$$D_{k,L} = (\hat{P}(X,Z) \,||\, P_{net}(X,Z))$$

$$= \sum_{x,z} P(x,z) \log \frac{P(x,z)}{P_{net}(x,z)}$$

$$= P(x=true, z=true) \log \frac{P(x=true, z=true)}{P_{net}(x=true, z=true)}$$

$$+P(x=true, z=\ false) \log \frac{P(x=true, z=\ false)}{P_{net}(x=true, z=\ false)}$$

$$+P(x=\ false, z=\ false) \log \frac{P(x=\ false, z=\ false)}{P_{net}(x=\ false, z=\ false)}$$

$$+P(x=\ false, z=true) \log \frac{P(x=\ false, z=true)}{P_{net}(x=\ false, z=true)}$$

$$= 0.38 \log \frac{0.38}{0.5*(0.8*0.9+0.2*0.2)} + 0.12 \log \frac{0.12}{0.5*(0.8*0.1+0.2*0.8)}$$

$$+0.33 \log \frac{0.33}{0.5*(0.8*0.8+0.1*0.2)} + 0.17 \log \frac{0.17}{0.5*(0.2*0.9+0.8*0.2)}$$

$$= 0$$

(e) As (d) shows that Kullback-Leibler divergence is 0 which means that those two distributions are exactly the same. So it does not matter which one to choose. This means we do not need to choose X as candidate for Z because it has covered all the information of distributions from X to Z.