CS 422 HW2
SHIQI LIU
1 Recitation Exercises
1.1 Chapter 3 Exercises: 2,3,5,6,7,8,12
3.2

(a) Compute the Gini

Total = 20; C0=10, C1=10,

Gini = 1 - $p(C0|Class)^2$ - $p(C1|Class)^2 = 1-2\times0.5^2 = 0.5$.

(b) Compute the Gini index for the Customer ID attribute.

The gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0.

(c) Compute the Gini index for the Gender attribute.

	M	F
Class 0	6	4
Class 1	4	6

The gini for Male is Gini (M) = 1 - $p(C0|M)^2$ - $p(C1|M)^2 = 1-(6/10)^2 - (4/10)^2= 0.48$
The gini for Female is Gini(F) = 1 - $p(C0|F)^2$ - $p(C1|F)^2 = 1-(4/10)^2 - (6/10)^2= 0.48$
Therefore, the overall gini for Gender is Gini(Gender) =[(T(M)/T(M+F)*Gini(M))] + [(T(F)/T(M+F))*Gini(F)]= (10/20)*0.48 +(10/20)*0.48 = 0.48

(d) Compute the Gini index for the Car Type attribute using multiway split.

	Fanily	Sports	Luxury	
Class0	1	8	1	10
Class1	3	0	7	10
	4	8	8	20

The gini for Family car is Gini (Family) = 1 - $p(C0|Family^2$ - $p(C1|Family)^2 = 1-(1/4)^2 - (3/4)^2= 0.375$,
Sports car is Gini (S) = 1 - $p(C0|S)^2$ - $p(C1|S)^2 = 1-(8/8)^2 - (0/8)^2= 0$,
Luxury car is Gini (L) = 1 - $p(C0|L)^2$ - $p(C1|L)^2 = 1-(1/8)^2 - (7/8)^2= 0.2188$.
The overall gini is Gini (CarType) =[(T(Family)/T(Family+S+L))*Gini(Family)] + [(T(S)/T(Family+S+L))*Gini(S)]+[(T(L)/T(Family+S+L))*Gini(L)]= (4/20)*0.375 +(8/20)*0+(8/20)*0.2188= 0.1625

(e) Compute the Gini index for the Shirt Size attribute using multiway split.

	Small	Medium	Large	extraLarge
Class0	3	3	2	2
Class1	2	4	2	2
	5	7	4	4

The gini for Small shirt size is Gini (Small) = 1 - $p(C0|Small)^2$ - $p(C1|Small)^2 = 1-(3/5)^2 - (2/5)^2= 0.48$
Medium shirt size is Gini (Medium) = 1 - $p(C0|Medium)^2$ - $p(C1|Medium)^2 = 1-(3/7)^2 - (4/7)^2= 0.4898$,

Large shirt size is Gini (Large) = 1 - $p(C0|Large)^2$ - $p(C1|Large)^2$ = 1-$(2/4)^2$ - $(2/4)^2$= 0.5,
Extra Large shirt size is Gini (extraLarge) = 1 - $p(C0|extraLarge)^2$ - $p(C1|extraLarge)^2$ = 1-$(2/4)^2$ - $(2/4)^2$= 0.5,
The overall gini for Shirt Size attribute
Gini(ShritSize)=[(T(Small)/T(ShirtSize))*Gini(Small)] +
[(T(Medium)/T(ShirtSize))*Gini(Medium)]+[(T(Large)/T(ShirtSize))*Gini(Large)])]+[(
T(extraLarge)/T(ShirtSize))*Gini(extraLarge)]= (5/20)*0.48
+(7/20)*0.4898+(4/20)*0.5+(4/20)*0.5 = 0.4914

(f) Which attribute is better, Gender, Car Type, or Shirt Size? Answer:
Car Type because it has the lowest gini among the three attributes.
Gender = .48 Car Type = .1625, and Shirt Size = .4914.
(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.
Becuase customer id is unique and thus can't be used as a predictive attribute. It would be useless becuase it divides it into all the possible nodes with out needing any sort of predictive behavior.

3.3

a. What is the entropy of this collection of training examples with respect to the class attribute?
+: 4
- :5
Total: 9
Entropy = $-\sum_{i=0}^{-1} P(i|t)\log P(i|t)$ = -(4/9)*$\log_2$(4/9) - (5/9)*$\log_2$(5/9) = .9911
b. What are the information gains of a1 and a2 relative to these training examples?

A1	+	-
T	3	1
F	1	4

Delta = I(parent) -I(children)
I(a1|t) = $-\sum_{i=0}^{-1} P(i|t)\log P(i|t)$ = -(3/4)*$\log_2$(3/4) - (1/4)*$\log_2$(1/4) = .8113
I(a1|f) = $-\sum_{i=0}^{-1} P(i|t)\log P(i|t)$ = -(1/5)*$\log_2$(1/5) - (4/5)*$\log_2$(4/5) = .7219
Delta= I(parent) - I(a1)=.9911-$\sum_j \frac{N(Vj)}{N} * I(Vj)$=.9911-(4/9)*I(a1|t)+(5/9)I(a1|f)=0.2345

A2	+	-
T	2	3
F	2	2

Delta = I(parent) -I(children)
I(a2|t) = $-\sum_{i=0}^{-1} P(i|t)\log P(i|t)$ = -(2/5)*$\log_2$(2/5) - (3/5)*$\log_2$(3/5) = .971
I(a2|f) = $-\sum_{i=0}^{-1} P(i|t)\log P(i|t)$ = -(2/4)*$\log_2$(2/4) - (2/4)*$\log_2$(2/4) = 1
Delta= I(parent) - I(a2)=.9911-$\sum_j \frac{N(Vj)}{N} * I(Vj)$=.9911-
(5/9)*I(a2|t)+(4/9)I(a2|f)=0.007211

c. For a3, which is a continuous attribute, compute the information gain for every possible split.

A3	Class Label	Split Point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

The best split for a3 occurs at split point equals to 2.

d. What is the best split (among a1, a2, and a3) according to the information gain?

The best split is using attribute a1 becuase it has the largest delta difference in entropy with .11427.

e. What is the best split (between a1 and a2) according to the misclassification error rate?

For attribute a1: error rate = 2/9.

For attribute a2: error rate = 4/9.

The best split is A1 since it has the lower classification error.

f. What is the best split (betweena1and a2)according to the Gini index?

A1:

Gini(T) = 1 - p(+|T)2 - p(-|T)2 = 1 - (3/4)2 - (1/4)2 = 0.375

Gini(F) = 1 - p(+|F)2 - p(-|F)2 = 1 - (1/5)2 - (4/5)2 = 0.32

TGini(a1) = [(Total(T)/Total(a1)]*Gini(T) + [(Total(F)/Total(a1)]*Gini(F) =(4/9)*0.375 + (5/9)*0.32 = 0.3444

A2:

Gini(T) = 1 - p(+|T)2 - p(-|T)2 = 1 - (2/5)2 - (3/5)2 = 0.48

Gini(F) = 1 - p(+|F)2 - p(-|F)2 = 1 - (2/4)2 - (2/4)2 = 0.5

TGini(a2) = [(Total(T)/Total(a2)]*Gini(T) + [(Total(F)/Total(a2)]*Gini(F) =(5/9)*0.48 + (4/9)*0.5 = 0.4889

The best split is A1 since the subsets for attribute a1 have a smaller Gini index.

3.5

a. Calculate the information gain when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

A	T	F
+ -	4 0	0 3

B	T	F

+	3	1
-	1	5

Overall entropy before splitting is:

$E_{orig} = -0.4\log(0.4)-0.6\log(0.6) = 0.9710$

The information gain after splitting on A is:

$E_{A=T} = -4/7\log4/7 – 3/7\log3/7 = 0.9852$

$E_{A=F} = -3/3\log3/3 – 0/3\log0/3 = 0$

$\Delta = E_{orig} - 7/10E_{A=T} - 3/10E_{A=F} = 0.2813$

Therefore, attribute A will be chosen to split the node

b. Calculate the gain in the Gini index when splitting on *A* and *B*. Which attribute would the decision tree induction algorithm choose?

Overall gini before splitting is:

$G_{orig} = 1-0.4^2 -0.6^2= 0.48$

The gain in gini after splitting on A is:

$G_{A=T} = 1-(4/7)^2 –(3/7)^2= 0.4898$

$G_{A=F} = 1-(3/3)^2 –(0/3)^2= 0$

$\Delta = G_{orig} - 4/10G_{A=T} - 6/10G_{A=F} = 0.1633$

Therefore, attribute B will be chosen to split the node

c. **Figure 3.11** shows that entropy and the Gini index are both monotonically increasing on the range [0, 0.5] and they are both monotonically decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

3.6

	P	C1	C2
Class0	7	3	4
Class1	3	0	3

a. Calculate the Gini index and misclassification error rate of the parent node P .
Gini = $1-(3/10)^2-(7/10)^2=1-0.09-0.49=0.42$

b. Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?
Class 0 = $1-(3/7)^2-(4/7)^2=0.5$
Class 1 = $1-(3/3)^2-(0/3)^2=1-1=0$

c. Calculate the weighted misclassification rate of the child nodes. Would you consider this attribute test condition if misclassification rate is used as the impurity measure?
Gini(children)=(7/10)*0.5+(3/10)*0=0.35

3.7

a. Compute a two-level decision tree using the greedy approach described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

Splitting at level 1

X	C1	C2
0	60	60
1	40	40

Error rate using attribute x is (60+40)/200=0.5

Y	C1	C2
0	40	60
1	60	40

Error rate using attribute Y is (40+40)/200=0.4

Z	C1	C2
0	30	70
1	70	30

Error rate using attribute x is (30+30)/200=0.3

Since z gives the lowest error rate, it is chosen as the splitting attribute at level 1

Splitting at level 2
After splitting on attribute Z, the subsequent test condition may involve either attribute X or Y . This depends on the training examples distributed to the Z = 0 and Z = 1 child nodes.
For Z = 0, the corresponding counts for attributes X and Y are the same, as shown in the table below.

X	C1	C2
0	15	45
1	15	25

Error rate (x,z=0)=(15+15)/100 =0.3

Y	C1	C2
0	15	45
1	15	25

Error rate (y,z=0)=(15+15)/100 =0.3

For Z = 1, the corresponding counts for attributes X and Y are shown in the tables below.

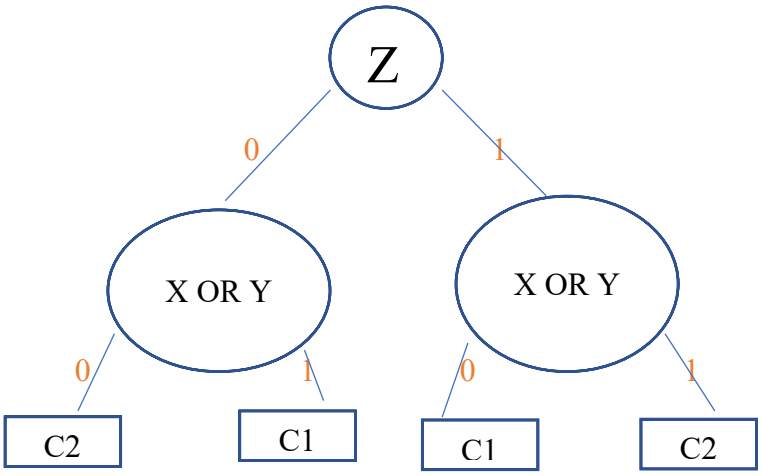
X	C1	C2
0	45	15
1	25	15

Error rate (x,z=1)=(15+15)/100 =0.3

Y	C1	C2
0	25	15
1	45	15

Error rate (y,z=1)=(15+15)/100 =0.3

Therefore, correspoding two-level decision tree is:



b. Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?
If the X attribute is already choosen we starrt with the next choice.

X=T

CE(ZT) = 1 - max[25/40 , 15/40] = .375
CE(ZF) = 1 - max[15/40 , 25/40] = .375
WE = 40/80 * .375 + 40/80 * .375 = .375

CE(YT) = 1 - max[5/40 , 35/40] = .125
CE(YF) = 1 - max[35/40 , 5/40] = .125
WE = 60/100 * .25 + 40/100 * .375 = .125

X=F

CE(ZT) = 1 - max[45/60 , 15/60] = .25
CE(ZF) = 1 - max[15/60 , 45/60] = .25
WE = 40/100 * .375 + 60/100 * .25 = .25

CE(YT) = 1 - max[55/60 , 5/60] = .083
CE(YF) = 1 - max[15/60 , 45/60] = .083
WE = 60/120 * .083 + 60/120 * .083 = .083

Since YT and YF on the X=F branch becuase they are lowst value. Then the ZT AND ZF are off the X=T branch.

c. Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.

When comparing the results from part (a) and (b), the suitability of the greedy heuristic does not produce optimum outcomes.

3.8

a. According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

The error rate for the data without partitioning on any attribute is
$E_{orig} = 1 - \max(50/100,50/100)=0.5$
After splitting on attribute A, the gain in error rate is:

A	T	F
+	25	25
-	0	50

$E_{A=T} = 1 - \max(25/25, 0/25) = 0$
$E_{A=F} = 1 - \max(25/75, 50/25) = 25/75$
$\Delta A = E_{orig} - 25/100 E_{A=T} - 75/100 E_{A=F} = 25/100$

After splitting on attribute B, the gain in error rate is:

B	T	F
+	30	20
-	20	30

$E_{B=T} = 1 - \max(30/50, 20/50) = 20/50$
$E_{B=F} = 1 - \max(20/50, 30/50) = 20/50$
$\Delta B = E_{orig} - 50/100 E_{B=T} - 50/100 E_{B=F} = 10/100$

After splitting on attribute C, the gain in error rate is:

C	T	F
+	25	25
-	25	25

$E_{C=T} = 1 - \max(25/50, 25/50) = 25/50$
$E_{C=F} = 1 - \max(25/50, 25/50) = 25/50$
$\Delta C = E_{orig} - 50/100 E_{C=T} - 50/100 E_{C=F} = 0/100 = 0$

The algorithm chooses attribute A because it has the highest gain

b. Repeat for the two children of the root node.

Because the A = T child node is pure, no further splitting is needed. For the A = F child node, the distribution of training instances is:

B	C	+	-
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

$E_{orig} = 25/75$

After splitting on attribute B, the gain in error rate is:

B	T	F
+	25	0
-	20	30

$E_{B=T} = 20/45$
$E_{B=F} = 0$
$\Delta B = E_{orig} - 45/75 E_{B=T} - 20/75 E_{B=F} = 5/75$

After splitting on attribute C, the gain in error rate is:

C	T	F

+	0	25
-	25	25

$E_{C=T} = 0/25$
$E_{C=F} = 25/50$
$\Delta C = E_{orig} - 25/75E_{c=T} - 50/75E_{c=F} = 0$

The split will be made on attribute B

d. How many instances are misclassified by the resulting decision tree?

20 instances are misclassified. (The error rate is 20/100)

e. Repeat parts (a), (b), and (c) using C as the splitting attribute.

For the C = T child node, the error rate before splitting is:

$E_{orig} = 25/50$

After splitting on attribute A, the gain in error rate is:

A	T	F
+	25	0
-	0	25

$E_{A=T} = 0$
$E_{A=F} = 0$
$\Delta A=25/50$

After splitting on attribute B, the gain in error rate is:

B	T	F
+	5	20
-	20	5

$E_{B=T} = 5/25$
$E_{B=F} = 5/25$
$\Delta B= 15/50$

Therefore, A is chosen as the splitting attribute

For the C = F child node, the error rate before splitting is:

$E_{orig} = 25/50$

After splitting on attribute A, the gain in error rate is:

A	T	F
+	0	25
-	0	25

$E_{A=T} = 0$
$E_{A=F} = 25/50$
Δ A=0

After splitting on attribute B, the gain in error rate is:

B	T	F
+	25	0
-	0	25

$E_{B=T} = 0$
$E_{B=F} = 0$
Δ B= 25/50

Therefore, B is chosen as the splitting attribute
The overall error rate of the induced tree is 0

f. Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.

The greedy heuristic does not necessarily lead to the best tree.

3.12

a. Based on the accuracies shown in Table 3.7 , which classification model would you expect to have better performance on unseen instances?
We will choose T10 on unseen data. This is because it has a better accuracy on unseen dataset. It also does not over fit to noise of training dataset unlike T100 which captures even noise on the training dataset to yield an accuracy of 0.97
b. Now, you tested T10 and T100 on the entire data set (A+B) and found that the classification accuracy of T10 on data set (A+B) is 0.85, whereas the classification accuracy of T100 on the data set (A+B) is 0.87. Based on this new information and your observations from Table 3.7 , which classification model would you finally choose for classification?
The new information is nothing but just the avergae accuracies on train and test data. The umbers obtained are avergae of earlier numbers. This essentially again means that on an unseen dataset we prefer amodel which performs better on the unseen portion (B) . Hence our decision will be the same. We would choose T10 on unseen dataset

Problem 2:
2.1

Based on the above, the tree with the max-depth of 2 indicated the highest recall for all three classses (0-1,1-1,2-0.89,macro avg:0.96 and weighted avg:0.97).

After max-depth of 2, the recall becomes constant among all three classes (0-1, 1-1, 2-0.89/0.78, macro and weighted avg:0.93) meaning that the tree is a pure as it can possibly

be based on the training set. Also, these data show after 2th depth, three classes might be splited mostly.
That indicates that at max-depth of 2, the recall is the highest because it has not done enough splits within the tree to calculate the proper recall.

Precision is the lowest at max-depth of 1. which is (0-1.0,1-0.50,2-0.0, macro avg:0.50, weighted avg:0.55) Because it has not predicted one of the classes (class 2).

F1-score is based on both precision and recall, and since recall and precision were highest in the tre of max-depth of 2, it is also the highest in the tree of max-depth of 2. Also, after depth 2,f1 becomes comstant.

Micro-average will aggregate the contributions of all classes to compute the average metric.
Macro-average will straight forward. Just take the average of the precision and recall of the system on different sets. compute the metric independently for each class and then take the average (hence treating all classes equally)
In a multi-class classification setup, micro-average is preferable if suspect there might be class imbalance (i.e have many more examples of one class than of other classes).
Weighted-average is similar to macro-average, but each metric is given an additional weight to further balance it out.

2.2

From the above calculations, we can determine that:

The entropy of the first split is: 0.884

The gini of the first split is: 0.422

The misclassification error of the first split is: 0.302

The information gain of the first split is: 0.544

The feature that was selected for the first split is X[1],which is ('mean texture')'cell size', which was determine throughout the training phase as the most valuable.(x[1]=2.5)

2.3

Summary:

Original data produced the following in (B, M) format:

precision (0.91, 0.98), recall (0.98, 0.88), and F1-Score (0.95, 0.92).

It's confusion matrix shows the following (tn, fp, fn, tp) format:

(64,1,6,43)

FPR (Fallout) = FP/(FP + TN) = 1/(1+64) = 1/65 = 0.001538

TPR (Recall) = TP/(TP + FN) = 43/(43+6) = 43/49 = 0.87755102

FPR/TPR = 0.00175371

PCA with component of 1:

precision (0.95, 0.97), recall (0.99, 0.89), and F1-Score (0.97, 0.93)

It's confusion matrix shows:

(77, 1, 4, 32)

FPR (Fallout) = FP/(FP + TN) = 1/(1+77) = 1/78 = 0.01282

TPR (Recall) = TP/(TP + FN) = 32/(32+4) = 32/36 = 0.88889

FPR/TPR = 0.14422

PCA with component of 2:

precision (0.94, 0.94), recall (0.97, 0.87), and F1-Score (0.95, 0.90)

It's confusion matrix shows:

(74, 2, 5, 33)

FPR (Fallout) = FP/(FP + TN) = 2/(2+74) = 2/76 = 0.02631

TPR (Recall) = TP/(TP + FN) = 33/(33+5) = 33/38 = 0.868421

FPR/TPR = 0.03029

According to summary, the F1, precision and recall continues to slowly decreases when we utilized pca with one component, and then evens out when we utilize pca with two components. As for the confusion matrix of each, we can see that the original continuous data shows a higher TP compared to pca with one or two components. FN is also the lowest with the original data. Because of this, I'm unsure on how the continuous data affects this model.
However, seeing as the split occurs around 142/143 samples for both original continous data and with pca components, I calculated the missclassification error rate to comapre the continous vs discrete. The error rate is higher for the continous data, which means it is not beneficial to the model compared to the discrete.

2.4

The empirical distribution for the above decision tree is 1.0%.

The threshold above split in a manner that 793/800 of the samples split to the right and 7 to the left. it is less than 10% split to the left.

This follows similar to the empirical distribution.