CS 5525

Biplav Choudhury (906089826)

7 October 2020

1 Solutions to Assignment 3

1.1 Problem 1 - Probability

1.1.1 Part 1

X takes on values 1 and 0, hence the average is calculated as

$$\begin{aligned} E(X) = 0 \times P(X=0) + 1 \times P(X=1) \\ \implies E(X) = P(X=1) \\ P(X=1) = P(X=1,Y=0) + P(X=1,Y=1) = \gamma + \delta \end{aligned}$$

Therefore expected value of X is $\gamma + \delta$.

Y takes on values 1 and 0, hence the average is calculated as

$$\begin{aligned} E(Y) = 0 \times P(Y=0) + 1 \times P(Y=1) \\ \implies E(Y) = P(Y=1) \\ P(Y=1) = P(X=1,Y=1) + P(X=0,Y=1) = \delta + \beta \end{aligned}$$

Therefore expected value of X is $\delta + \beta$.

1.1.2 Part 2

The necessary and sufficient condition for X and Y to be independent is $P(X,Y) = P(X).(Y)$. With four different conditions for x=0/1 and y=0/1, the conditions are :

$$(\alpha+\beta)(\alpha+\gamma) = \alpha, (\alpha+\beta)(\beta+\delta) = \beta, (\gamma+\delta)(\alpha+\gamma) = \gamma, (\gamma+\delta)(\beta+\delta) = \delta$$

(1)

1.2 Problem 2 - Bayes Theorem & Naïve Bayes Classifier

1.2.1 Part 1

We can construct a scenario shown below where the people are the samples, the two mutually exclusive features are real and placebo, and the final label is recovered/not-recovered.

Table 1: Dataset

Samples	**Placebo**	**Real**	**Recovered**
Person 1	Yes	No	No
Person 2	No	Yes	Yes

Total recovered $= 6500$. Real $=$ Placebo $= 0.5 \times 6500 = 3250$

Total not recovered $= 3500$. Real $= 0.55 \times 3500 = 1925$. Placebo $= 1575$

Total users for real drug $= 3250+1925=5175$, Total users for placebo drug $= 3250+1575=4825$.

Proportion of users given real drug $= 5175/10000 = 0.5175$

To check efficiency of the drug, we use BT to see the probability of recovered given that the sample was real drug.

$$\text{P(Recovered}|Real) = \frac{P(Real|Recover) \times P(Recover)}{P(Real)} = \frac{0.5 \times 0.65}{0.5175} = 0.6280 which is more than 50\% chance.$$

(2)

To confirm, lets check the efficiency of the placebo.

$$\text{P(Recovered}|Placebo) = \frac{P(Placebo|Recover) \times P(Recover)}{P(Placebo)} = \frac{0.5 \times 0.65}{0.4825} = 0.6735$$

(3)

It shows that placebo is better in treating the disease. So probably the drug should **not** be used.

1.2.2 Part 2

a) Independence check between X and Y is by $P(X,Y) = P(X).P(Y)$. P(X1=0)=$\frac{35}{100}$, P(X1=1)=$\frac{65}{100}$, P(X2=0)=$\frac{59}{100}$, P(X2=1)=$\frac{41}{100}$.
P(X1=0,X2=0) = $\frac{25}{100}$. This is not equal to P(X1=0).P(X2=0)=$\frac{35 \times 59}{100 \times 100}$.

b) Conditional Independence check between X and Y is by $P(X,Y|Z) = P(X|Z).P(Y|Z)$.

Table 2: Positive Class

$P(X_1, X_2\|+)$	$P(X_1\|+)$	$P(X_2\|+)$	Independence
$P(X_1 = 0, X_2 = 0\|+) = \frac{5}{50}$	$P(X_1 = 0\|+) = \frac{10}{50}$	$P(X_2 = 0\|+) = \frac{25}{50}$	Yes
$P(X_1 = 0, X_2 = 1\|+) = \frac{5}{50}$	$P(X_1 = 0\|+) = \frac{10}{50}$	$P(X_2 = 1\|+) = \frac{25}{50}$	Yes
$P(X_1 = 1, X_2 = 0\|+) = \frac{20}{50}$	$P(X_1 = 1\|+) = \frac{40}{50}$	$P(X_2 = 0\|+) = \frac{25}{50}$	Yes
$P(X_1 = 1, X_2 = 1\|+) = \frac{20}{50}$	$P(X_1 = 1\|+) = \frac{40}{50}$	$P(X_2 = 1\|+) = \frac{25}{50}$	Yes

For positive class, they are independent.

Table 3: Negative Class

$P(X_1, X_2\|-)$	$P(X_1\|-)$	$P(X_2\|-)$	Independence
$P(X_1 = 0, X_2 = 0\|-) = \frac{17}{50}$	$P(X_1 = 0\|-) = \frac{25}{50}$	$P(X_2 = 0\|-) = \frac{34}{50}$	Yes
$P(X_1 = 0, X_2 = 1\|-) = \frac{8}{50}$	$P(X_1 = 0\|-) = \frac{25}{50}$	$P(X_2 = 1\|-) = \frac{16}{50}$	Yes
$P(X_1 = 1, X_2 = 0\|-) = \frac{17}{50}$	$P(X_1 = 1\|-) = \frac{25}{50}$	$P(X_2 = 0\|-) = \frac{34}{50}$	Yes
$P(X_1 = 1, X_2 = 1\|-) = \frac{8}{50}$	$P(X_1 = 1\|-) = \frac{25}{50}$	$P(X_2 = 1\|-) = \frac{16}{50}$	Yes

For Negative class, they are conditionally independent of each other given the class.

c)

$P(X1 = 1|+) = \frac{40}{50}$, $P(X1 = 1|-) = \frac{25}{50}$, $P(X2 = 1|+) = \frac{25}{50}$, $P(X2 = 1|-) = \frac{16}{50}$, $P(X3 = 1|+) = \frac{20}{50}$, $P(X3 = 1|-) = \frac{8}{50}$.

These values will be used to calculate the class of the samples shown in the table.

d)

P(X1=1,X2=1,X3=1)

For positive, P(X1 =1|+) * P(X2 =1|+) * P(X3 = 1|+) = 0.8.*0.5*0.4=0.16.
For negative, P(X1 =1|-) * P(X2 =1|-) * P(X3 = 1|-) = 0.5*0.32*0.16 = 0.0256
As the probability of it being in the positive class is more, hence it belongs to the positive class.

P(X1=1,X2=0,X3=0)

For positive, P(X1 =1|+) * P(X2 =0|+) * P(X3 = 0|+) = 0.8*0.5*0.6=0.24.
For negative, P(X1 =1|-) * P(X2 =0|-) * P(X3 = 0|-) = 0.5*0.68*0.84. = 0.2856
As the probability of it being in the negative class is more, hence it belongs to the negative class.

P(X1=0,X2=1,X3=0)

For positive, P(X1 =0|+) * P(X2 =1|+) * P(X3 = 0|+) = 0.2*0.5*0.6 = 0.06
For negative, P(X1 =0|-) * P(X2 =1|-) * P(X3 = 0|-) = 0.5*0.32*0.84 = 0.5*0.32*0.84 = 0.1344
As the probability of it being in the negative class is more, hence it belongs to the negative class.

P(X1=0,X2=0,X3=0)

For positive, P(X1 =0|+) * P(X2 =0|+) * P(X3 = 0|+) = 0.2*0.5*0.6 = 0.06
For negative, P(X1 =0|-) * P(X2 =0|-) * P(X3 = 0|-) = 0.5*0.68*0.84 = 0.2856
As the probability of it being in the negative class is more, hence it belongs to the negative class.

Therefore the error rate is $\frac{38}{100}$.

1.2.3 Part 3

C and B are independent

P(C,B) = $\sum_A$P(C,A,B) = $\sum_A$P(C|A)P(A)P(B)

= $\sum_A$(P(A)P(C|A)P(B) = $\sum A$P(C,A)P(B) = P(C)P(B)

D and B are independent given A

P(D,B|A) = $\frac{P(A,B,D)}{P(A)}$ = $\sum_C \frac{P(A,B,C,D)}{P(A)}$) = $\sum_C \frac{P(A)P(B)P(C|A)P(D|C)}{P(A)}$

Again $P(D|A) = \frac{P(D,A)}{P(A)} = \sum_C \frac{P(A,C,D}{P(A)} = \sum_C \frac{P(A)P(C|A)P(D|C)}{P(A)}$

and $P(B|A) = \frac{P(B,A)}{P(A)} = \frac{P(B)P(A)}{P(A)} = P(B)$

Multiplying $P(D|A)$ and $P(B|A)$, we get $\sum_C \frac{P(B)P(C|A)P(D|C)}{1}$

which is equal to $P(D,B|A)$.

C and E are not independent given B

P(C,E|B) = $\frac{C,E,B}{P(B)} = \sum_A \frac{P(A,B,C,E}{P(B)}$

= $\sum_A P(A)P(B)P(C|A)P(E|A,B)$

Then P(C,B) = $\frac{P(C,B)}{P(B)} = \sum_A \frac{A,B,C}{P(B)} = \sum_A \frac{P(A)P(B)P(C,A)}{P(B)}$

And P(E|B) = $\frac{P(E,B)}{P(B)} = \sum_A \frac{A,B,E}{P(B)} = \sum_A \frac{P(A)P(B)P(E|A,B)}{P(B)}$

There P(C,E|B) is not equal to P(C,B)*P(E|B) and they are not independent.

1.2.4 Part 4

Given A=1, B=1, D=1
P(C=0|A=1,B=1,D=1) is what we need to find

= P(A=1,B=1,C=0,D=1) / P(A=1, B=1, D=1) converting conditional to joint
= P(A=1) P(B=1) P(C=0|A=1) P(D=1|C=0)/P(A=1) P(B=1) P(D=1|C)
= P(C=0|A=1) using naiver bayes relation from graph
= 1 – P(C=1|A=1) = 1 – 0.7 = 0.3

Similarly like the above approach, P(C=1|A=1,B=1D=1)
= P(A=1,B=1,C=0,D=1)/P(A=1, B=1, D=1)
= P(A=1) P(B=1) P(C=1|A=1) P(D=1)/P(A=1)P(B=1)P(D=1|C)
= P(C=1|A=1) = 0.7

Since P(C=0|A=1,B=1,D=1) is less than P(C=1|A=1,B=1,D=1), C = 1 is less likely.

1.3 Problem 3 - Analyzing the kNN

1.4 Part 1

Euclidean distance = 8.08 and Manhattan Distance = 14.5, therefore the points are closer in Euclidean Distance.

1.5 Part 2

for [1, 5], closest neighbors are [[1, 5], [2, 6], [2, 7], [3, 7], [3, 8]]. Negative correct.

for [2, 6], closest neighbors are [[2, 6], [2, 7], [1, 5], [3, 7], [3, 8]]. Negative correct.

for [2, 7], closest neighbors are [[2, 7], [2, 6], [3, 7], [3, 8], [1, 5]]. Negative wrong.

for [3, 7], closest neighbors are [[3, 7], [2, 7], [3, 8], [2, 6], [4, 8]]. Negative correct.

for [3, 8], closest neighbors are [[3, 8], [3, 7], [4, 8], [2, 7], [2, 6]]. Negative correct.

for [4, 8], closest neighbors are [[4, 8], [3, 8], [3, 7], [5, 9], [2, 7]]. Negative correct.

for [5, 9], closest neighbors are [[5, 9], [4, 8], [3, 8], [3, 7], [2, 7]]. Negative correct.

for [5, 1], closest neighbors are [[5, 1], [6, 2], [7, 2], [7, 3], [8, 3]]. Positive correct.

for [6, 2], closest neighbors are [[6, 2], [7, 2], [5, 1], [7, 3], [8, 3]]. Positive correct.

for [7, 2], closest neighbors are [[7, 2], [6, 2], [7, 3], [8, 3], [5, 1]]. Positive wrong.

for [7, 3], closest neighbors are [[7, 3], [7, 2], [8, 3], [6, 2], [8, 4]]. Positive correct.

for [8, 3], closest neighbors are [[8, 3], [7, 3], [8, 4], [7, 2], [6, 2]]. Positive wrong.

for [8, 4], closest neighbors are [[8, 4], [8, 3], [7, 3], [9, 5], [7, 2]]. Positive correct.

for [9, 5], closest neighbors are [[9, 5], [8, 4], [8, 3], [7, 3], [7, 2]]. Positive correct.

Classification error is 4/14.

1.6 Part 3

With each point itself added as a neighbor, a nearest neighbor of 1 will be the best.

1.7 Part 4

1 No as there is no **1** nearest neighbor, 4 of them have the same distance.
2 Yes
3 No as there is no **3** nearest neighbor, 4 of them have the same distance.
4 Yes

1.8 Part 5

Based on the GPAs, the 3 nearest neighbors are the students 4, 3 and 7. Therefore the predicted salary based on the average is 132,000$.

1.9 Problem 4 - Rule based classifier

1.10 Part 1

Coverage will decrease, but accuracy may increase or decrease. Example in a scenario of 1000 samples, 50 samples are covered and 30 of them satisfy the rule. Now when we bring in additional restriction of coverage in addition to the present condition, coverage will decrease.

Considering extreme cases, let's assume coverage becomes 45 from 50. Now the 5 samples that get excluded, they could have had severe heart disease or not. If they had, new accuracy is $\frac{30-5}{45} = \frac{25}{45}$ else it becomes $\frac{30}{45}$. So the accuracy can go in either direction.

1.11 Part 2

Coverage will decrease, but accuracy may increase or decrease based on similar argument of the previous question.

1.12 Part 4

Coverage will remain same as the LHS of the rule remains the same. But accuracy may increase or will remain the same. This is because the new class of severe can bring in positive samples satisfying the rule which can increase the accuracy, and in the worst case, accuracy will remain the same.

1.13 Part 5

Coverage will decrease or remain the same as now we are no longer taking the maximum of BP. For each visit, there might be a few instances when the BP was less than 150 and that will reduce the coverage.

Accuracy may increase or decrease as it depends on the relationship of BP with heart disease.

1.14 Problem 5 - Evaluation Measures

Accuracy = 0.8087,
Precision = 0.7259,
Recall = 0.8305,
F-measure = 0.7746,
Cost = 1939,
Sensitivity = Recall = 0.8305,
Specificity = 0.7944,
False Positive Rate = 0.2055.

1.15 Problem 6 - Cross-validation

- One obvious reason could be that the data while doing CV was not randomized properly to be a representative of the main data, due to which the samples of the same label end up in the same validation set thereby reducing the error. But the fact that this was repeated 100 times makes this fact less probable.

- Another possible cause might be that due to the very less number of samples compared to the features, which leads to overfitting. But cross-validation is supposed to be a preventive measure against overfitting and this approach still performs better with CV, so this too seems not very probable.
- One possible scenario is that when the algorithm picks useful features through correlation, it is actually picking up irrelevant features like the serial number. This will create a problem as features like serial numbers are not indicative of the actual label and because of the way the data is organized, serial numbers might have a high correlation with the label and therefore lead to high accuracy while training and validation. Therefore the features selected should be investigated.

  The cross-validation performed may not be right as by doing feature selection before performing cross-validation, even the data in all the folds get changed which could bias the performance. source. Another article discussing this phenomenon is this. So the current approach of doing CV after feature selection is considered wrong as it tampers with the data with which we will perform cross-validation.Based on the answer at this link, in a scenario with fewer samples and large features, we can do leave-one-out cross-validation which is an extreme case of k-fold validation. Additionally, sample duplication to generate additional samples similar to the original samples. The links have been given in the quiz.