Assignment 3

10/13/2020

Name: Yusuf Elnady

Problem 1. Probability

P1.1

We have the formula for Expected value as $E(X) = \sum_{x,y}(x * P(x, y))$.
$E(X) = 0 * \alpha + 0 * \beta + 1 * \gamma + 1 * \delta = \gamma + \delta$
$E(Y) = 0 * \alpha + 1 * \beta + 0 * \gamma + 1 * \delta = \beta + \delta$

P1.2

If we have independent variables X and Y, then E[X · Y] = E[X] · E[Y]. (i.e. $P_{X,Y}(x, y) = P_X(x) * P_Y(y)$).
First, we have the following equations: $E[X.Y] = (0 * 0 * \alpha) + (0 * 1 * \beta) + (1 * 0 * \gamma) + (1 * 1 * \delta) = \delta$
So, the necessary and sufficient condition is $\delta = (\gamma + \delta) * (\beta + \delta)$
We can also use the marginal proabilities to find the conditions for having $P(X, Y) = P_X(x) * P_Y(y)$

$$P(X = 0, Y = 0) = P_X(x = 0) * P_Y(y = 0) = \alpha = (\alpha + \beta)(\alpha + \gamma)$$

$$P(X = 0, Y = 1) = P_X(x = 0) * P_Y(y = 1) = \beta = (\alpha + \beta)(\beta + \delta)$$

$$P(X = 1, Y = 0) = P_X(x = 1) * P_Y(y = 0) = \gamma = (\gamma + \delta)(\alpha + \gamma)$$

$$P(X = 1, Y = 1) = P_X(x = 1) * P_Y(y = 1) = \delta = (\gamma + \delta)(\beta + \delta)$$

Moreover, we can divide the last two formulas to get a more compact formula:

$$\frac{\gamma}{\delta} = \frac{(\gamma + \delta)(\alpha + \gamma)}{(\gamma + \delta)(\beta + \delta)} = \frac{(\alpha + \gamma)}{(\beta + \delta)}$$

$$\delta * \alpha + \delta * \gamma = \gamma * \beta + \gamma * \delta$$

$$\delta\alpha = \gamma\beta$$

Problem 2. Bayes Theorem Naïve Bayes Classifier

P2.1

$$P(recovered) = 0.65$$

$$P(\neg recovered) = 1 - 0.65 = 0.35$$

$$P(drug, recovered) = P(drug|recovered) * P(recovered) = 0.5 * 0.65 = 0.325$$

$$P(\neg drug, recovered) = 0.5 * 0.65 = 0.325$$

$$P(drug, \neg recovered) = P(drug|\neg recovered) * P(\neg recovered) = 0.55 * 0.35 = 0.1925$$

In order to see if taking the drug helps a patient to recover from the disease, we will need to compare between the probability of the class "recovered" given that you have taken this drug, and the probability of "not recovered" given that you have taken this drug.

$$P(recovered|drug) = \frac{P(drug, recovered)}{P(drug)}$$

.

So, we need to calculate first the proportion of test subjects who were given the real drug as

$$P(drug) = P(drug|recovered) * P(recovered) + P(drug|\neg recovered) * P(\neg recovered)$$

$$P(drug) = 0.325 + 0.1925 = 0.5175$$

$$P(\neg drug)1 - 0.5175 = 0.4825$$

$$P(recovered|drug) = \frac{P(drug, recovered)}{P(drug)} = \frac{0.325}{0.5175} = 0.628$$

$$P(recovered|\neg drug) = \frac{P(\neg drug, recovered)}{P(\neg drug)} = \frac{0.325}{0.4825} = 0.674$$

So, we can infer that taking drug will not help in recovering from the disease.

P2.2

a.As we did in P1.b, to know if two variables are independent, we can check if this condition $P(X, Y) = P(x) * P(y)$ holds or not. So, we can start by $P(X_1 = 1, X_2 = 1) = 7/25$ and $P(x_1 = 1) * P(x_2 = 1) = 13/20 * 41/100 = 533/2000$. So, it's clear that X_1 and X_2 are **not independent**, and we don't need further calculations.

b. To check if two variables are conditionally independent given the class we have the formula $P(X_1, X_2|C) = P(X_1|C) * P(X_2|C)$. The class can be + or -, and X_1 and X_2 can either be 0 or 1. For the +ve class:

$$P(X_1 = 0, X_2 = 0|+) = 1/10 = P(X_1 = 0|+) * P(X_2 = 0|+) = 10/50 * 25/50 = 1/10$$

$$P(X_1 = 0, X_2 = 1|+) = 1/10 = P(X_1 = 0|+) * P(X_2 = 1|+) = 10/50 * 25/50 = 1/10$$
$$P(X_1 = 1, X_2 = 0|+) = 2/5 = P(X_1 = 1|+) * P(X_2 = 0|+) = 40/50 * 25/50 = 2/5$$
$$P(X_1 = 1, X_2 = 1|+) = 2/5 = P(X_1 = 1|+) * P(X_2 = 1|+) = 40/50 * 25/50 = 2/5$$

So they are conditionally independent given the +ve class. Next, we check for the -ve class:

$$P(X_1 = 0, X_2 = 0|-) = 17/50 = P(X_1 = 0|-) * P(X_2 = 0|-) = 25/50 * 34/50 = 17/50$$
$$P(X_1 = 0, X_2 = 1|-) = 8/50 = P(X_1 = 0|-) * P(X_2 = 1|-) = 25/50 * 16/50 = 8/50$$
$$P(X_1 = 1, X_2 = 0|-) = 17/50 = P(X_1 = 1|-) * P(X_2 = 0|-) = 25/50 * 34/50 = 27/50$$
$$P(X_1 = 1, X_2 = 1|-) = 8/50 = P(X_1 = 1|-) * P(X_2 = 1|-) = 25/50 * 16/50 = 8/50$$

From all of these equations, it's clear that X_1 and X_2 are **conditionally independent given the class**.

c.

$$P(X_1 = 1|+) = 40/50 = 4/5$$
$$P(X_1 = 1|-) = 25/50 = 1/2$$
$$P(X_2 = 1|+) = 25/50 = 1/2$$
$$P(X_2 = 1|-) = 16/50 = 8/25$$
$$P(X_3 = 1|+) = 20/50 = 2/5$$
$$P(X_3 = 1|-) = 8/50 = 4/25$$

d. The way to solve this problem is to calculate the conditional probability of having the +ve or the -ve class, given each of the 4 combinations of the attributes we have in the data set. So, I will start one row by row:

$$P(+|X_1 = 1, X_2 = 1, X_3 = 1) = \frac{P(X_1 = 1, X_2 = 1, X_3 = 1|+)P(+)}{P(X_1 = 1, X_2 = 1, X_3 = 1)} = \frac{20/50 * 1/2}{7/25} = 5/7$$
$$P(-|X_1 = 1, X_2 = 1, X_3 = 1) = \frac{P(X_1 = 1, X_2 = 1, X_3 = 1|-)P(-)}{P(X_1 = 1, X_2 = 1, X_3 = 1)} = \frac{8/50 * 1/2}{7/25} = 2/7$$

$$P(+|X_1 = 1, X_2 = 0, X_3 = 0) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 0|+)P(+)}{P(X_1 = 1, X_2 = 0, X_3 = 0)} = \frac{2/5 * 1/2}{37/100} = 20/37$$
$$P(-|X_1 = 1, X_2 = 0, X_3 = 0) = \frac{P(X_1 = 1, X_2 = 0, X_3 = 0|-)P(-)}{P(X_1 = 1, X_2 = 0, X_3 = 0)} = \frac{17/50 * 1/2}{37/100} = 17/37$$

$$P(+|X_1 = 0, X_2 = 1, X_3 = 0) = \frac{P(X_1 = 0, X_2 = 1, X_3 = 0|+)P(+)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} = \frac{1/10 * 1/2}{13/100} = 5/13$$

$$P(-|X_1 = 0, X_2 = 1, X_3 = 0) = \frac{P(X_1 = 0, X_2 = 1, X_3 = 0|-)P(-)}{P(X_1 = 0, X_2 = 1, X_3 = 0)} = \frac{4/25 * 1/2}{13/100} = 8/13$$

$$P(+|X_1 = 0, X_2 = 0, X_3 = 0) = \frac{P(X_1 = 0, X_2 = 0, X_3 = 0|+)P(+)}{P(X_1 = 0, X_2 = 0, X_3 = 0)} = \frac{1/10 * 1/2}{11/50} = 5/22$$

$$P(-|X_1 = 0, X_2 = 0, X_3 = 0) = \frac{P(X_1 = 0, X_2 = 0, X_3 = 0|-)P(-)}{P(X_1 = 0, X_2 = 0, X_3 = 0)} = \frac{17/50 * 1/2}{11/50} = 17/22$$

So, given $X_1 = 1, X_2 = 1, X_3 = 1$ or $X_1 = 1, X_2 = 0, X_3 = 0$ we will choose +ve, and if $X_1 = 0, X_2 = 1, X_3 = 0$ or $X_1 = 0, X_2 = 0, X_3 = 0$ we will choose -ve, therefore the training error of Bayes naive classifier $= \frac{8+17+5+5}{100} = 0.350$

P2.3

a: We see that C depends on A, and that B is already a parent node that doesn't depend on anything else. Therefore C and B are independent. Moreover, B is the parent node of E, but there's no relation between E and C. So, **C and B are independent**

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

b: To check if D and B are conditionally independent given A is to use d-separation: First, we draw the ancestral graph of only the variables mentioned in the probability expression, and we can see directly that B and D are independent given A without going further into the moralize and disorient steps since node B is not connected to the others. Even after we moralize, no edges will be added between any two parents, and after disorienting, there's no path between D and B **Therefore, D and B are conditionally independent given A.**

c: To check if C and E are conditionally independent given B is to use d-separation: First, we draw the ancestral graph of only the variables mentioned in the probability expression. Second, we will moralize by connecting undirected edge between Parents of (E) (i.e. (A) and (E)). Then, we disorient by replacing all directed edges with undirected ones, those are edges (A,C) and (A,E). Finally, we delete the given node (B). It's clear that there's a path from (C) to (E). **Therefore, E and C are not conditionally independent given B**

P2.4

First, we need to calculate $P(C = 0|A = 1, B = 1, D = 1)$ which is equal to

$$P(C = 0|A = 1, B = 1, D = 1) = \frac{P(A = 1, B = 1, D = 1|C = 0)P(C = 0)}{P(A = 1, B = 1, D = 1)}$$

$$= \frac{P(A = 1)P(B = 1)P(D = 1|C = 0)P(C = 0)}{P(A = 1, B = 1, D = 1)}$$

Using the law of Total Probability, we can compute $P(C = 0)$.

$$P(C = 1) = P(C = 1|A = 0)P(A = 0) + P(C = 1|A = 1)P(A = 1) = 0.4 * 0.5 + 0.7 * 0.5 = 0.55$$

Therefore $P(C = 0) = 1 - 0.55 = 0.45$. Now we can compute

$$P(C = 0|A = 1, B = 1, D = 1) = \frac{P(A = 1)P(B = 1)P(D = 1|C = 0)P(C = 0)}{P(A = 1, B = 1, D = 1)}$$

$$= \frac{0.5 * 0.6 * 0.4 * 0.45}{P(A = 1, B = 1, D = 1)} = \frac{0.054}{P(A = 1, B = 1, D = 1)}$$

We also need to calculate $P(C = 1|A = 1, B = 1, D = 1)$

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

$$= \frac{P(A = 1)P(B = 1)P(D = 1|C = 1)P(C = 1)}{P(A = 1, B = 1, D = 1)} = \frac{0.5 * 0.6 * 0.3 * 0.55}{P(A = 1, B = 1, D = 1)}$$

$$\frac{0.0495}{P(A = 1, B = 1, D = 1)}$$

I am just ignoring the denominator while comparing the results from the two formulas, because it just the same normalizing.

As 0.054 > 0.0495 , so C is more likely to be (0).

P3.1

Using the euclidean distance to calculate the distance between the object A and object B, we get the following : $\sqrt{(3 - 9.1)^2 + (2.1 - 0.7)^2 + (4.8 - 2.2)^2 + (5.1 - 5.1)^2 + (6.2 - 1.8)^2} = 8.080$. Using the euclidean distance to calculate the distance between the object A and object B, we get the following : $|3 - 9.1| + |2.1 - 0.7| + |4.8 - 2.2| + |5.1 - 5.1| + |6.2 - 1.8| = 14.500$

So, they are closer in 5-D space using the **euclidean distance**.

P3.2

If we chose the k to be 5, therefore we will have 4 points that are misclassified out of a total of 14 points, which means the classification error on the training set is $4/14 = 0.286$.

P3.3

If (K = 1) that means for every point it will be classified as itself (as its ground truth value), and that will result in a training error = 0.

P3.4

- If we decided to assign x the label of its nearest neighbor,then we didn't solve the problem, because we still don't know which one should be assigned (the green or the red), because the four surrounding points are all at the same distance from x.
- If we decided to Flip a coin to randomly assign a label to x (from the labels of its 4 closest points), then that will definitely work and will be a way to break ties.
- Using k=3 doesn't solve the problem because we will not be sure how to decide which three out of the four to be chosen (given our specific case), therefore, we won't be able to decide how to classify the point (x).
- Using k=5, can be used as another solution to avoid ties, because the 5th point (will be a green one), because it's much closer to the (x) from the other third red point. Therefore, x will be classified as a green point.

The answer is: Flip a coin to randomly assign a label to x (from the labels of its 4 closest points) or Use k=5 instead.

P3.5

In order to choose the 3 nearest neighbors, I will need to calculate the euclidean distance between each student with the student (9).

So, the distances are as the following: Student 1,9: 1.315. Student 2,9: 0.806. Student 3,9: 0.2. Student 4,9: 0.6403. Student 5,9: 0.7071. Student 6,9: 2.6. Student 7,9: 0.5. Student 8,9: 1.1704.

So, the nearest 3 students to the student (9) are (3), (7), (4) with Salary of 91, 163, 142 respectively, then we will average them.

The prediction for Student 9's salary is 132 thousands of dollars per year

---

P4.a

First, we have only (BloodPressure $>$ 150) → HeartDisease=Severe. If we are adding one more conjunct (CholesterolLevel $>$ 245), that means that the conditions are more restrict, therefore we may have **the same or less coverage**. The accuracy cannot be determined because we don't have a dataset that we can refer back to, so we may have **the same, more, or less accuracy** based on the all new data will be covered, we don't know how many of them is classified correctly.

P4.b

If we have now the (BloodPressure $>$ 200) instead of 150. That means we still cover all the instances that are greater than 200, but if there are instances between 151 and 200, they will not be covered. That means we may have **the same or less coverage**. Similarily, as in P4.a, we don't know anything about the actual dataset, so we may have **the same, more, or less accuracy**.

P4.c

Now, there're only two classes, and our rule is (BloodPressure $>$ 150) $\rightarrow$ HeartDisease=Yes. As, the left hand side (the conjunct) didn't change so we expect to have **the same coverage**. Then, previously in P4.a, our rule may have been covering some mild cases that were classified as Severe, but now they will be classified correctly as (Yes). So, we expect **the accuracy to remain the same or increase**.

P4.d

First, the data instances in our dataset will be categorized by the patients, and for every patient it will have rows that includes the blood pressure and chief compliant for this visit. If the patient has three visits that has chief compliant = Heart-Related, then it's classified as severe. In terms of the coverage, we are assuming now that our dataset instead of having for example 4 patients, we may have 5*4 rows, and each may contain any result of the blood pressure and any compliant of Heart-related. If we assumed that still have the same blood pressures as previously in the dataset, and the new visits has blood pressures less than 150, we will also have a smaller coverage due to the increase in the number of records in the dataset.

Overall, we cannot determine the coverage, so we may have **the same, more, or less coverage**.

To understand the accuracy of these modified instances, we still cannot determine what is the type of the new visits that will occur and it may result in an accuracy that is lower than having just the patients without calculating how many visits they did and so on. Also, if all have more than three Heart-Related, and they were classified all as HeartDisease = Severe, then we will have very high accuracy, that is higher than the original case we have. So, overall we cannot determine the accuracy and it may be **the same, more, or less accuracy**.

P5

a: Accuracy:

$$\frac{TP+TN}{TP+FN+FP+TN} = \frac{98+143}{20+37+98+143} = \frac{241}{298} = 0.809$$

b: Precision:

$$\frac{TP}{TP+FP} = \frac{98}{37+98} = 0.726$$

c: Recall:

$$\frac{TP}{TP+FN} = \frac{98}{20+98} = 0.831$$

d: F-Measure:

$$\frac{2rp}{r+p} = \frac{2 * \frac{49}{59} * \frac{98}{136}}{\frac{49}{59} + \frac{98}{136}} = 0.775$$

e: Cost:

$$-98 + 2000 + 37 = 1939$$

f: Sensitivity (recall) = TPR:

$$\frac{TP}{TP + FN} = \frac{98}{20 + 98} = 0.831$$

g: Specifity =TRN:

$$\frac{TN}{TN + FP} = \frac{143}{143 + 37} = 0.795$$

h: FPR =

$$\frac{FP}{FP + TN} = \frac{37}{37 + 143} = 0.206$$

P6

First, we know that the actual error rate is 50% for any classifier, so having the mean CV error rate to be 2.7% means that **we did not perform the Cross Validation correctly**.

Repeating the configuration 100 times is so much, and that means that we only have 50 samples but we are training 100 times which doesn't make sense and will result in having the model estimated on the validation set two times and the error we will get is not a representing number for the generalization error that we may counter when trying this classifier on a test data set (totally new points). The experiments have proven that 10-fold cross-validation is the best number to get an accurate generalization error on the validation data to estimates the performance of the model, because in 10-folds you are covering 90% of the samples in your training data, and 10% for validation data, which will never be used as validation data two times. Moreover, in the given problem it's said that D = 1000 and that these features sampled from a standard normal distribution such that these features are independent of the labels. But we have used only 100 features out of the available 1000, so in order to make use out of all available features, we can randomly make 10 groups of the features, each group of 100 features. The process will be to use this set of features (that are randomly chosen) to train the model, then we repeat this step using the next group till we cover all the 1000 features, which means we will need to do that 10 times. Each iteration of those 10 times consists of 10-fold cross-validation which means there are two inner loops. Finally, we will have 10 average errors for each group of features we selected and we can calculate the overall mean CV error rate and it will be easy to have it 50% (a really representing generalization error) not 2.7% as we had previously.