Homework 3

Your Name: Srujan Shekar Shetty

Student ID: A20529733

---

1. [40] **Manually** solve the questions below by using Naïve Bayes Classifier

We conducted a survey to collect people's daily diets and try to build a model to predict whether their diets result in healthy conditions or not. The final results could be Yes, No. Note: using green rows as training, orange rows as testing.

Breakfast	**Lunch**	**Dinner**	**Healthy?**
Ham	Carnivorous	Beef	Y
Milk	Carnivorous	Beef	N
Bread	Veggie	Pork	N
Bread	Veggie	Veggie	Y
Ham	Veggie	Veggie	Y
Milk	Carnivorous	Pork	N
Bread	Carnivorous	Beef	N
Ham	Veggie	Pork	Y
Milk	Veggie	Pork	Y
Milk	Carnivorous	Veggie	N
Noddle	Carnivorous	Pork	?

1). [5 points] What is Laplace smoothing? And why we need it in the Naïve Bayesian classifier?

Ans Problem - Zero probability issue

Solution- Laplace smoothing.

How it works:

$$P(E \mid c_i) = P(e_1 \wedge e_2 \wedge \cdots \wedge e_m \mid c_i) = \prod_{j=1}^{m} P(e_j \mid c_i)$$

From the above table if we are trying to predict that breakfast is noodles using a Naïve Bayesian classifier given the health. But in our training set, we have no breakfast as noodles. In this case, one among the e feature will be zero. Meaning the whole

P(E | ci) will be zero. This is called the Zero conditional probability problem.

To replace the zero value, we use the technology called Laplace smoothing.

Laplace smoothing: P (ei = A | C = cj) = (nc + m*p) / (n + m)

– A is a value in the i-th feature; cj is a value in the label

– nc = # of training instances for which ei = A and C = cj

– n = # of training instances for which C =cj

– m = a weight factor, usually m>=1 and could be big value, for example, the size of your training data

– p = an estimate or a probability value to decrease m, usually, we can set p as 1/t and t is the number of unique values in e

2). [15 points] Using the Categorical Naive Bayesian Classification to make predictions on the test sets, present confusion matrix, and calculate accuracy, precision, recall, F1 measure, specificity, by considering Y as positive label

i) $\underbrace{\text{Breakfast}}_{e_1}$, lunch $_{e_2}$, dinner $_{e_3}$

(i) Bread, Carnivorous, Beef

$P(Y) = 3/6$

$P(N) = 3/6$

$$P(Y|E) = \frac{P(Y)\ P(E|Y)}{P(E)} \quad \text{—(1)}$$

$$P(N|E) = \frac{P(N)\ P(E|N)}{P(E)} \quad \text{—(2)}$$

We can only compare numerators of (1) & (2) To keep it simple as denominators are the same

From (1)

$P(Y|E) = P(Y)\ P(E|Y)$

$$= \frac{3}{6} \times (P(e_1|Y)\ P(e_2|Y)\ P(e_3|Y))$$

$$= \frac{3}{6} \times (\cancel{1/3} \times 1/3 \times 1/3) = \frac{1}{6} \times \frac{1}{3} \times \frac{1}{3} = \underline{0.0185}$$

$$P(N|E) = P(N)\ P(E|N) = \left(\frac{3}{6}\right) \times \left(1/3 \times 2/3 \times 1/3\right) = \frac{1}{27} = \underline{0.037}$$

~~(ii) Ham Veggie Pork~~

~~Applying the above formula~~

~~$P(Y|E) = P(Y)\ P(E|Y)$~~

~~$= \left(\frac{3}{6}\right) \times ($~~

P.T.O

~~P(N) > P(Y)~~

We have more confidence to say we can trust 'N'

$P(N|E) > P(Y|E)$

$(0.037) > (0.01851)$

(ii) Ham e_1, Veggie e_2, Pork e_3

$P(Y|E) = P(Y)\ (P(E|Y))$ (ignoring denominator as we compare it to $P(N|E)$ having same denominator)

$= \frac{3}{6} \times \left(\frac{2}{3} \times \frac{2}{3} \times \frac{0}{3} \right)$

$\downarrow$

Use laplace smoothing to overcome this problem (Zero Conditional Probability problem)

$$\boxed{P(e_i = A \mid c = c_j) = (n_c + m * p) / (n + m)}$$

~~where~~

For our case,

$= \frac{3}{6} \times \left(\frac{2}{3} \times \frac{2}{3} \times \frac{0 + (m * p)}{m + n} \right)$

where $m - 6$
$p - 1/t - 1/3$
$n - 3$

$\frac{\cancel{6}^{2} \times \frac{1}{3}}{6 + 3} = 2/9$

$= \frac{3}{6} \times \left(\frac{2}{3} \times \frac{2}{3} \times \frac{0 + (6 * 1/3)}{(6+3)} \right)$

$= \frac{\cancel{3}^{1}}{\cancel{6}} \times \left(\frac{2}{3} \times \frac{2}{3} \times \frac{2}{\cancel{9}_{3}} \right)$

$P(Y|E) = \frac{1}{\cancel{6}_{3}} \times \left(\frac{\cancel{2}^{1}}{3} \times \frac{2}{3} \times \frac{2}{3} \right) = 0.4938$

$P(N|E) = \frac{3}{6} \left(\frac{0 + \left(\cancel{6}^{2} \times 1/3\right)}{6+3} * \frac{1}{3} \times \frac{2}{3} \right) = 0.246$

$\frac{\cancel{3}}{\cancel{6}_{3}} \times \frac{\cancel{2}^{1}}{9} \times \frac{1}{\cancel{3}} \times \frac{2}{3}$

$\therefore P(Y|E) > P(N|E)$
So we trust 'Y' as Healthy label

(iii) Milk Veggie Pork

$$P(Y|E) = P(Y)\,P(E|Y) = \frac{3}{6} \times \left(\frac{6 \times 1/3}{6+3}\right) \times \frac{2}{3} \times \left(\frac{6 \times 1/3}{6+3}\right)$$

$$= \frac{3}{6} \times \frac{2}{9} \times \frac{2}{3} \times \frac{2}{9} = \underline{0.01646}$$

$$P(N|E) = \frac{3}{6} \times \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} = \underline{0.7407}$$

$P(N|E) > P(Y|E)$
∴ So we can trust 'N' as health label.

(iv) Milk Carnivorous Veggie

$$P(Y|E) = \frac{3}{6} \times \left(\frac{2}{9} \times \frac{1}{3} \times \frac{2}{3}\right) = \underline{0.0246}$$

$$P(N/E) = \frac{3}{6} \times \left(\frac{2}{3} \times \frac{2}{3} \times \frac{2}{9}\right) = \underline{0.04938}$$

$P(N|E) > P(Y|E)$
∴ We can trust 'N' as healthy label

Confusion Matrix

Predicted labels

	Y	N
Actual labels Y	True positives (TP) 1	False Negatives (FN) 1
Actual label N	False positives (FP) 0	True Negatives (TN) 2

∴ Accuracy = (TP + TN) / All
= (1 + 2) / 4 = 3/4 = 75%

∴ Precision = (TP) / (TP + FP)
= (1 / 1 + 0) = 100%

∴ Recall = (TP) / (TP + FN) = 1 / 1+1 = 50%

∴ F1 Measure = $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{precision} + \text{Recall}}$

$$= \frac{2 \times 100 \times 50}{100 + 150} = \frac{2 \times 1 \times 1/2}{1 + \frac{1}{2}}$$

$$= \frac{1}{1 + \frac{1}{2}} = \frac{1}{\frac{3}{2}} = 0.666$$

$$= \underline{66\%}$$

∴ Specificity considering Y as positive label

Specificity = TN / (FP + TN)
= 2 / (0 + 2) = 100%

NOTE :- The last observation with '?' label is not used for the laplace calculation. ONLY training & testing dataset is used.

2). [20 points] Using the Categorical Naive Bayesian Classification to make prediction on the unseen data (note: building the model by using both the green and orange rows, and predicting the label for unseen data/last row)

2) (b) All the rows considering as training data

$E =$ Noodle, Carnivorous, Pork

$C_1 = Y \quad C_2 = N$

• $P(\text{Noodle} \mid C_1) = \frac{m_c + m \times p}{m + n} = \frac{0 + \cancel{10}\,5 \times 1/\cancel{4}\,2}{10 + 5} = \frac{5/2}{15} = \frac{5}{30} = \frac{1}{6} = \underline{0.1667}$

m - Total Rows

p - 1/t

n - no of 'Y' in lables

• $P(\text{Carnivorous} \mid C_1) = \frac{1}{5}$

$P(\text{Pork} \mid C_1) = \frac{2}{5}$

$P(E \mid C_1) = 0.1667 \times 1/5 \times 2/5 = \underline{0.013336}$

$P(\text{Noodle} \mid C_2) = \frac{0 + 10 \times \frac{1}{4}}{10 + 5} = \underline{0.1667}$

$P(\text{Pork} \mid C_2) = 2/5$

$P(\text{Carnivorous} \mid C_2) = 4/5$

$P(E \mid C_2) = 0.1667 \times 4/5 \times 2/5 = \underline{0.05334}$

$P(C_1) = 5/10 = 1/2$; $P(C_2) = 5/10 = 1/2$

e

$P(E) = \frac{1}{2} \times 0.01336 + \frac{1}{2} \times 0.053344$

$= \frac{1}{2}(0.01336 + 0.053344) = \cancel{0.03334}$

0.03334

$P(C_1|E) = \frac{P(C_1)(P(E|C_1))}{P(E)} = \frac{\frac{1}{2} \times 0.13336}{0.03334} = 0.200$

$P(C_2|E) = \frac{P(C_2)P(E|C_2)}{P(E)} = \frac{1/2 \times 0.053344}{0.03334}$

$= 0.800$

$P(C_2|E) > P(C_1|E)$

∴ We can ~~simply~~ say we have more confidence that the label is 'N'.

Healthy – 'N'

2. (60 points) Python practice for Naïve Bayes classification
Use the Malware_MultiClass.csv data (predicting the column "classification"), and run 5 Naïve Bayes techniques by using hold-out evaluations (80% as training)

Note:

- You need to change different/multiple parameters to find the best NB model.
- You should evaluate the models by using accuracy, micro-precision and micro-recall, micro-F1 and micro-AUC
- Give conclusions about the best model by comparing the evaluation metrics above

Submission

- The ipynb and saved html files
- The comparison and conclusions of different models

Ans- Comparison

Model

Results of Categorical Naive Bayes using df_binary.

Hold-out Evaluation: Accuracy = 0.7458

Micro Precision = 0.7648960114630021

Micro Recall = 0.7460409171725716

Micro F1 = 0.7412574511263368

Micro AUC = 0.8057403055985901

Results of Categorical Naive Bayes using df_binary.

Hold-out Evaluation: Accuracy = 0.7458

Micro Precision = 0.7648960114630021

Micro Recall = 0.7460409171725716

Micro F1 = 0.7412574511263368

Micro AUC = 0.8057403055985901

Results of Multinomial Naive Bayes using df_num.

Hold-out Evaluation: Accuracy = 0.62225

Micro Precision = 0.6265632168368146

Micro Recall = 0.6222222310000007

Micro F1 = 0.6189721844020803

Micro AUC = 0.6038926643503398

Results of Gaussian Naive Bayes using df_num.

Hold-out Evaluation: Accuracy = 0.6174

Micro Precision = 0.6203667501296704

Micro Recall = 0.6170361858205575

Micro F1 = 0.6145891369097447

Micro AUC = 0.7118165207678421

Results of Gaussian Naive Bayes using df_num_std.

Hold-out Evaluation: Accuracy = 0.49495

Micro Precision = 0.706251630151278

Micro Recall = 0.7024574526847485

Micro F1 = 0.7005634219234647

Micro AUC = 0.5

Results of Bernoulli Naive Bayes using df_binary.

Hold-out Evaluation: Accuracy = 0.74935

Micro Precision = 0.7689996022120577

Micro Recall = 0.7500736021462386

Micro F1 = 0.7450445669649486

Micro AUC = 0.8097013358909546

Comparing the above metrics
Results of Bernoulli Naive Bayes using df_binary.

Hold-out Evaluation: Accuracy = 0.74935

Gives more accuracy Precision recall and we can conclude this model is better comparatively.