Homework 3

Your Name: Srujan Shekar Shetty

Student ID: <u>A20529733</u>

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 <u>Yes</u>, <u>No</u>. 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:

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

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

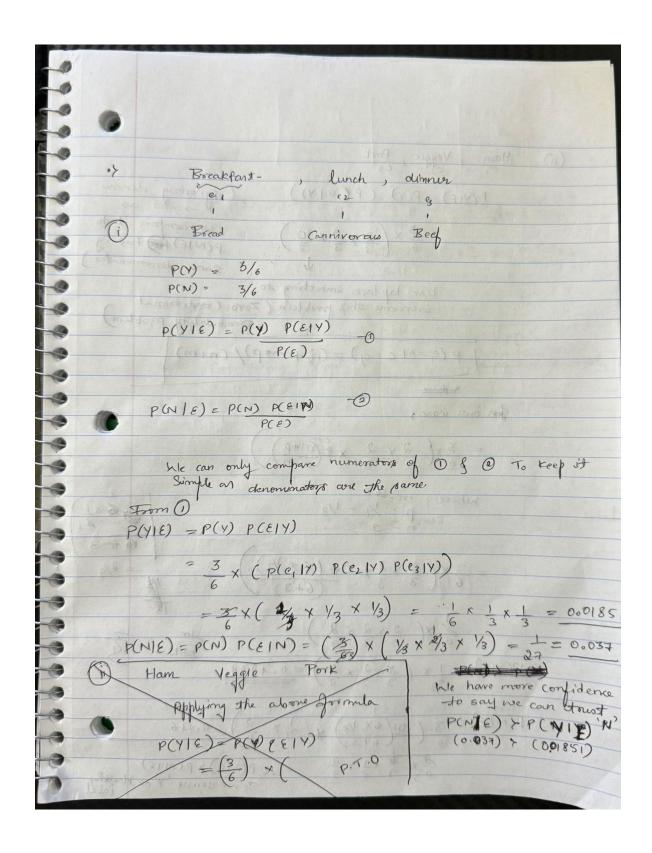
P(E | ci) will be zero. This is called the Zero conditional probability problem.

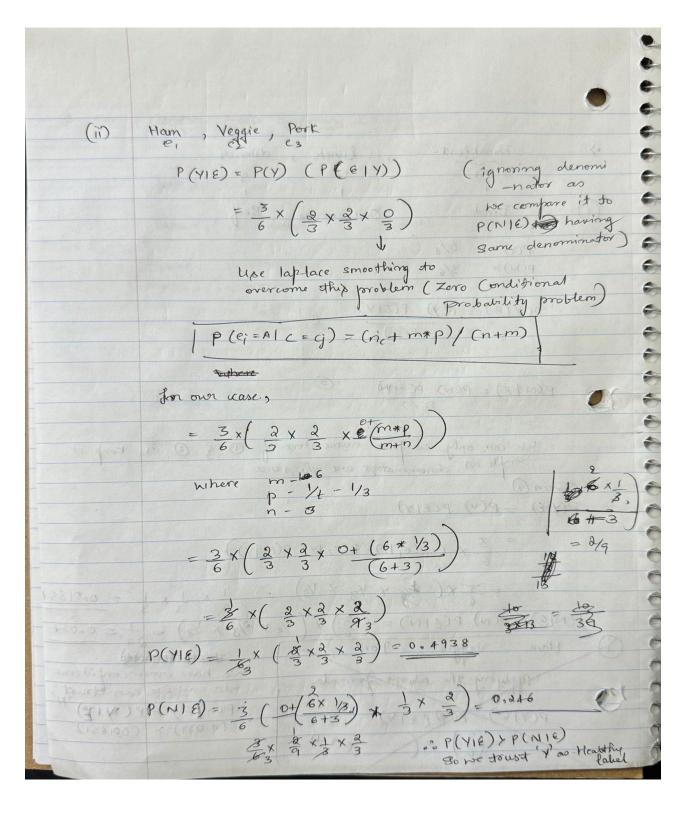
To replace the zero value, we use the technology called Laplace smoothing.

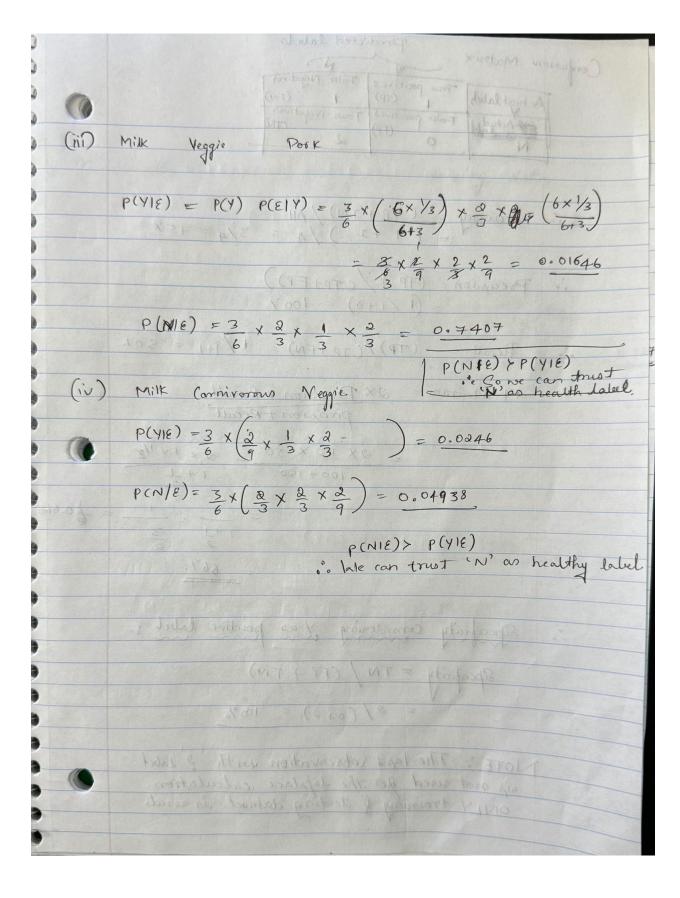
Laplace smoothing: $P(ei = A \mid C = cj) = (nc + m*p) / (n + m)$

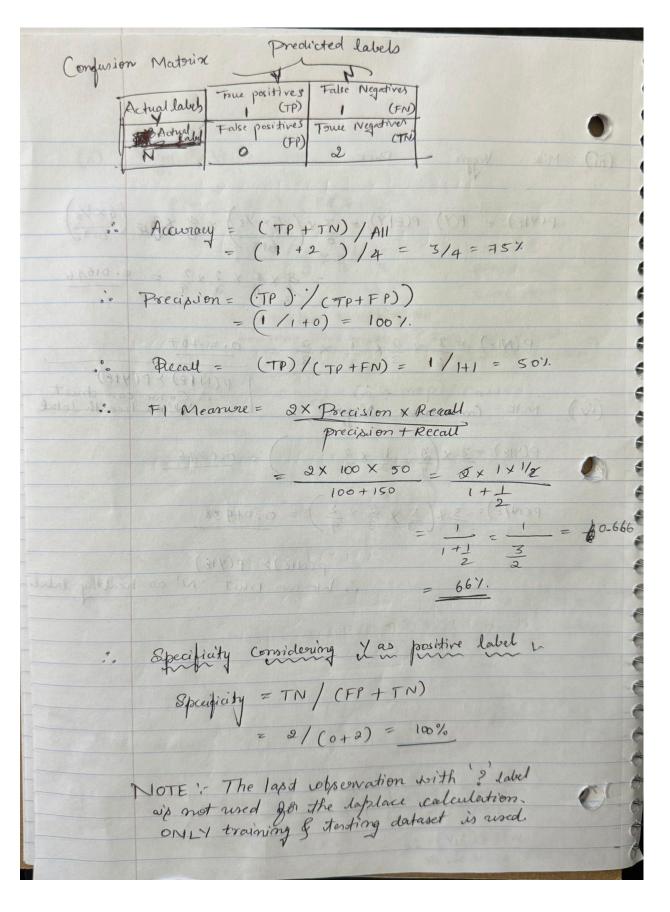
- − A is a value in the i-th feature; cj is a value in the label
- -nc = # of training instances for which ei = A and C = cj
- -n = # of training instances for which C = cj
- -m = a weight factor, usually m>=1 and could be big value, for example, the size of your training data
- -p = an estimate or a 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









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 orones considering as torcining data E= Noodle, Carrivorous, Posk 3. C,- Y (2=N • P(Noodle | Ci) = $m_c + m \times p = 0 + t \times 1 \times p$ m-Total Roms p - 1/t n - no of (y') im lables n - no of (y') im lables. P(Carnivorous / C1) = 10 = 1.9(0)9= (3/0)99 p (Pork 14) = '& P(E1C1) = 0.1667 x 1/5 x 0/5 = 0.013336 P (Noodle | C2) = 10+ 10x 4 10- 001667 P(POSK 1 (2) = 2/5 P (lownivorous | C2) = 4/5 P(E(c2)= 0.1667 x 4/5 x 2/5 = 0.05334 P(C1) = 5/10=1/2 3 P(C2) = 5/10 = 1/2

P(E)=1 x 0.01336+1 x 0.053344 = 1 (0.01336+0.053344) = 0 0.03334 6 M Karto = dxua + ou - (1) I apparent)d $P(C||E) = P(C|)(P(E|C|)) = 1 \times 0.73336$ P(E) = 0.2000.03374 establing by look P(C2/E) = P(C2) P(E/C2) = 1/2 x 0.053344 P(E) 0.03334 =0.800 P(c218) > P(c,18) ... We can simply say her 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 = 0.622222310000007

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.