IT255 mumps and spam classifier

Report: Analysis of Mumps Probability

This report presents the analysis of the probability of having the mumps based on the given data. The analysis is performed using the PyMC3 library in Python. The model considers the variables I (Infection), R (Roommate's infection), and M (Mumps).

Model Specification
The model is specified as follows:

1. Infection (I): Categorical variable with two levels. The prior probabilities are P(I=+i) = 0.8 and P(I=-i) = 0.2.

2. Roommate's infection (R): Categorical variable with two levels. The prior probabilities are P(R=+r) = 0.4 and P(R=-r) = 0.6.

3. Mumps (M): Categorical variable with four levels. The conditional probabilities are as follows:

- P(M=+m | I=+i, R=+r) = 0
- P(M=+m | I=+i, R=-r) = 1
- P(M=+m | I=-i, R=+r) = 0
- P(M=+m | I=-i, R=-r) = 0.7
- P(M=-m | I=+i, R=+r) = 0
- P(M=-m | I=+i, R=-r) = 0
- P(M=-m | I=-i, R=+r) = 1
- P(M=-m | I=-i, R=-r) = 0.3

Analysis Results

1. Table with I R M values and P(I,R,M):

The table provides a summary of the joint probabilities P(I,R,M) for different combinations of I, R, and M. It includes the mean, standard deviation, highest density interval (HDI), and other statistics for each probability value.

Example table entry:
- I[0]: Probability of I=+i: 0.8323
- R[0]: Probability of R=+r: 0.4079

- M[0,0]: Probability of M=+m given I=+i, R=+r: 0.3844
- M[1,0]: Probability of M=+m given I=+i, R=-r: 0.6156
- M[2,0]: Probability of M=+m given I=-i, R=+r: 0.6155
- M[3,0]: Probability of M=+m given I=-i, R=-r: 0.2322
- p_IRM[0]: Probability of I=+i, R=+r, M=+m: 0.2562
- p_IRM[1]: Probability of I=+i, R=+r, M=-m: 0.0470
- p_IRM[2]: Probability of I=+i, R=-r, M=+m: 0.3462
- p_IRM[3]: Probability of I=+i, R=-r, M=-m: 0.3506

2. Marginal probability P(+m) that you have the mumps:

The marginal probability P(+m) represents the overall probability of having the mumps regardless of the infection and roommate's infection. The calculated value is presented.

3. Probability P(+r | +m) that your roommate has the mumps given that you have

the mumps:

This probability represents the likelihood of your roommate having the mumps given that you have the mumps. The calculated value is presented.

Conclusion

Based on the analysis, the following conclusions can be drawn:

- The marginal probability of having the mumps (P(+m)) is approximately 0.5. This indicates that there is an equal chance of having the mumps or not.

- The probability of your roommate having the mumps (P(+r | +m)) is approximately 0.616. This suggests that there is a relatively high likelihood that your roommate also has the mumps if you have the mumps.

These results provide valuable insights into the probabilities associated with mumps infection and can be used to make informed decisions or further investigate the transmission dynamics of the disease.

Screenshot

```
Table with I R M values and P(I,R,M):
            mean     sd      hdi_3%   hdi_97%  mcse_mean  mcse_sd  ess_mean  ess_sd  ess_bulk  ess_tail  r_hat
I[0]      0.8323  0.3748    0.0000    1.0000     0.0054   0.0039    7847.0  7847.0    7877.0    5435.0    1.0
R[0]      0.4079  0.4912    0.0000    1.0000     0.0063   0.0044    8060.0  8060.0    8104.0    5389.0    1.0
M[0,0]    0.3844  0.4864    0.0000    1.0000     0.0043   0.0030    8022.0  8006.0    8103.0    5527.0    1.0
M[1,0]    0.6156  0.4864    0.0000    1.0000     0.0043   0.0030    8022.0  8006.0     798.0    5527.0    1.0
M[2,0]    0.6155  0.4863    0.0000    1.0000     0.0043   0.0030    8019.0  7997.0    8025.0    5435.0    1.0
M[3,0]    0.2322  0.4227    0.0000    1.0000     0.0057   0.0040    7931.0  7931.0    7918.0    5343.0    1.0
p_IRM[0]  0.2562  0.4364    0.0000    1.0000     0.0054   0.0038    7852.0  7852.0    7827.0    5389.0    1.0
p_IRM[1]  0.0470  0.2112    0.0000    0.6475     0.0025   0.0018    7981.0  7981.0    7943.0    5527.0    1.0
p_IRM[2]  0.3462  0.4755    0.0000    1.0000     0.0063   0.0044    8080.0  8080.0    8123.0    5435.0    1.0
p_IRM[3]   0.3506  0.4779    0.0000    1.0000     0.0063   0.0044    8062.0  8062.0    8099.0    5359.0    1.0

Marginal probability P(+m) that you have the mumps: 0.5

Probability P(+r | +m) that your roommate has the mumps given that you have the mumps: 0.6156
```

Report: Naive Bayes Classifier Analysis

This report presents the analysis of a Naive Bayes classifier implemented using the `MultinomialNB` class from the scikit-learn library. The analysis involves three parts: (a), (b), and (c).

Part (a)
In this part, a Naive Bayes classifier is trained to classify text messages as either 'spam' or 'ham' based on a given training dataset. The classifier is then used to predict the label of a test instance. The threshold value for the classifier is calculated using the predicted probabilities.

- Threshold value: The threshold value is calculated as the ratio of the predicted probability of 'ham' to the predicted probability of 'spam'. It determines the decision boundary for classifying instances as 'ham' or 'spam'.

Part (b)
In this part, another Naive Bayes classifier is trained to classify text messages as either 'spam' or 'ham' based on a different training dataset. The classifier is then used to calculate the following conditional word probabilities:

- P(W=sir | Y=spam): The probability of the word 'sir' occurring in a 'spam' message.
- P(W=watch | Y=ham): The probability of the word 'watch' occurring in a 'ham' message.
- P(W=gauntlet | Y=ham): The probability of the word 'gauntlet' occurring in a 'ham' message.
- P(Y=ham): The prior probability of a message being classified as 'ham'.

These probabilities are calculated based on the feature log probabilities and class log priors obtained from the trained classifier.

Part (c)

In this part, the analysis focuses on the vocabulary size of the text data. The minimal number of conditional word probabilities needed to represent the text data is calculated. The vocabulary size (V) is assumed to be 4, and the minimal number of probabilities is calculated based on the formula V + V^2 + V^3 + V^4.

The calculated minimal number of conditional word probabilities represents the number of unique combinations of words and their occurrences required to fully capture the information in the text data.

Conclusion

The Naive Bayes classifier analysis provides insights into the classification of text messages as 'spam' or 'ham'. The threshold value determines the decision boundary for classifying instances, while the conditional word probabilities and vocabulary size give a deeper understanding of the underlying probabilities and complexity of the text data.

These results can be used to further refine the classifier, optimize the feature set, or improve the overall classification accuracy for text message classification tasks.

```
PS C:\Users\vivek\Desktop\sem4\IT255-Ai> pytho
4\IT255-Ai\211AI041_IT255_Spam.py"

a.
Threshold value: 0.7346938775510207

b.
0.08333333333333333
0.0625
0.12500000000000003
0.6666666666666666

c.
340
PS C:\Users\vivek\Desktop\sem4\IT255-Ai>
```