****

**NAIVE BAYES**

**ARVIND PAWAR**

**NORTHEASTERN UNIVERSITY**

**WEEK 4**

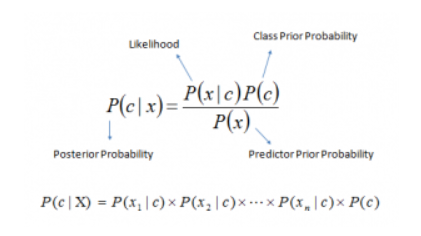
**ALY6020 PREDICTIVE ANALYTICS**

**INSTRUCTOR NAME- DR. MARCO MONTE DE OCA**



**Introduction**

* Naive Bayes is a simple and effective probabilistic classifier. It uses a maximum ‘Posteriori Decision’ rule in a Bayesian setting.
* As this technique is based on Bayes’ Theorem, it assumes that the predictors are independent on each other.
* For example, a fruit considered to be guava if it is green, round and about 3 inches in diameter. Even if these features depend upon the existence of the other features, the properties of these features independently contribute to the probability that the fruit is guava, and therefore, it is known as ‘Naive’.



Where,

* P(c|x) refers to posterior probability of class (c, target) given predictor (x, attributes).
* P(c) denotes the prior probability of class.
* P(x|c) is the likelihood which is the probability of predictor given class.
* P(x) is the prior probability of predictor.

**ANALYSIS**

1. First, we calculated prior probabilities for each digit.



Fig. 1: Calculating prior probabilities

1. In the second step we made the subsets for each label from 0 to 9.

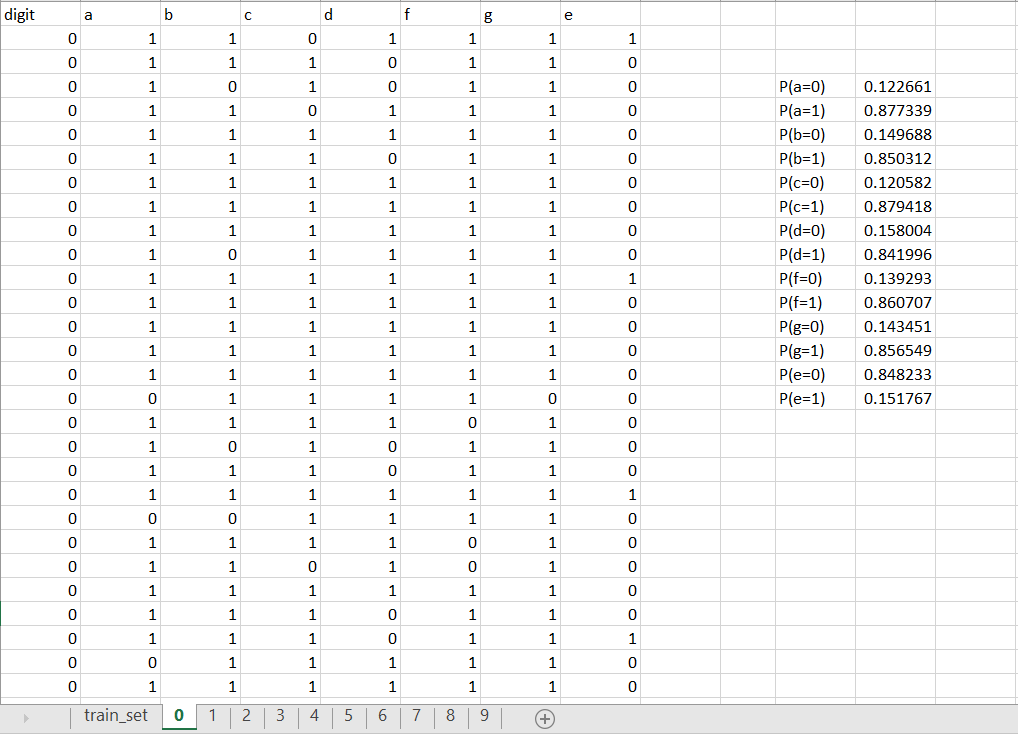


Fig. 2: subset for each digit

1. Then we calculated probabilities of each LED with respect to given digit.

|  |  |
| --- | --- |
| P(a=0) | =COUNTIF(B2:B482,"=0")/COUNT($A$2:$A$482) |
| P(a=1) | =COUNTIF(B2:B482,"=1")/COUNT($A$2:$A$482) |
| P(b=0) | =COUNTIF(C2:C482,"=0")/COUNT($A$2:$A$482) |
| P(b=1) | =COUNTIF(C2:C482,"=1")/COUNT($A$2:$A$482) |
| P(c=0) | =COUNTIF(D2:D482,"=0")/COUNT($A$2:$A$482) |
| P(c=1) | =COUNTIF(D2:D482,"=1")/COUNT($A$2:$A$482) |
| P(d=0) | =COUNTIF(E2:E482,"=0")/COUNT($A$2:$A$482) |
| P(d=1) | =COUNTIF(E2:E482,"=1")/COUNT($A$2:$A$482) |
| P(f=0) | =COUNTIF(F2:F482,"=0")/COUNT($A$2:$A$482) |
| P(f=1) | =COUNTIF(F2:F482,"=1")/COUNT($A$2:$A$482) |
| P(g=0) | =COUNTIF(G2:G482,"=0")/COUNT($A$2:$A$482) |
| P(g=1) | =COUNTIF(G2:G482,"=1")/COUNT($A$2:$A$482) |
| P(e=0) | =COUNTIF(H2:H482,"=0")/COUNT($A$2:$A$482) |
| P(e=1) | =COUNTIF(H2:H482,"=1")/COUNT($A$2:$A$482) |

|  |  |
| --- | --- |
| P(a=0) | 0.122661 |
| P(a=1) | 0.877339 |
| P(b=0) | 0.149688 |
| P(b=1) | 0.850312 |
| P(c=0) | 0.120582 |
| P(c=1) | 0.879418 |
| P(d=0) | 0.158004 |
| P(d=1) | 0.841996 |
| P(f=0) | 0.139293 |
| P(f=1) | 0.860707 |
| P(g=0) | 0.143451 |
| P(g=1) | 0.856549 |
| P(e=0) | 0.848233 |
| P(e=1) | 0.151767 |

Fig. 3: Calculating likelihood probabilities

1. Calculating the probability for each digit based on the given state of LED’s, that is, 0 or 1 for ‘a’ to ‘g’ alphabets. Used natural logarithm to normalize the calculated probabilities for each digit. This method tends to improve the efficiency of the entire model. The predicted values refer to the digit corresponding to the maximum probability digit. Here we used index() and match() function to map, the maximum posterior probability to the corresponding digit.



Fig. 4: Predicting the digit in test dataset

1. After making predictions, we obtained the confusion matrix, which determines the number of observations which predict the digit from 0 to 1 when the observed value was fixed (say 0). 61 observations were correctly predicted as 0 when there observed value is 0. There were 2, 4, 0, 2, 3, 15, 1, 5, 4 observations were incorrectly predicted when the observed label was 0. Similarly, the table records the count value for every predicted value corresponding to each observed label.



Fig. 5: Confusion Matrix

1. The accuracy is calculated by dividing the sum of all the correctly predicted labels to the total number of labels. Using Naïve Bayes model, we got 62.70% accuracy.



Fig. 6: Accuracy

**Conclusion:**

* The accuracy of the model is relatively average. Therefore, the model must be optimized to improve the predictions made, which would enhance the accuracy of the entire model.
* Use of normalization and smoothing also improves the accuracy of the model. However, since smoothing is not required, none of the likelihood probabilities become entirely zero.