Problem Set 2: Naive Bayes Algorithm

Denver Cooper
Artificial Intelligence @ 5:25
University of Arkansas – Fort Smith

February 24, 2023

Abstract

A Naive Bayes classifier written in Java with the goal of calculating probabilities on a set of training data and using those probabilities to successfully classify a given set of features into either class 0 or class 1.

1 Introduction

This is an implementation of the Naive Bayes classification algorithm to predict the probabilities of a set of features belonging to one class or another. The Naive Bayes classifier operates under the "naive" assumption that all features are conditionally independent of each other.

2 Background

2.1 Liver Disease Prediction using SVM and NB

Dr. S. Vijayarani and Mr.S.Dhayanand conducted an experiment with the Naive Bayes algorithm to predict liver diseases such as Cirrhosis, Bile Duct, and Chronic Hepatitis from a Liver Function Test dataset [Moh15]. Their results were Naive Bayes correctly classifying 61.28% of the data

2.2 Improved Naive Bayes Classification Algorithm for Traffic Risk Management

Hua Rui, Songhua Hu, and Hong Chen conducted an experiment to attempt to improve the accuracy of the standard Naive Bayes algorithm by combining feature weighting and Laplace calibration to overcome the shortcomings of the original Naive Bayes such as assuming attribute independence. [Hua+21] Their experiment analyzed a sample of random traffic violation cases in a city form January 2019 to December 2019 and recognized two kinds of traffic violations, speeding and running red lights. Their experiment showed an increase in accuracy from 49.5% to 92% when comparing the original naive Bayes results and their imporved naive Bayes algorithm.

3 Specification

For this implementation the following formula was used to classify features

$$\hat{c} = \underset{j \in i \ldots k}{argmax} \frac{P(C_j) \prod_{i=1}^{n} P(f_i|C_j)}{\sum_{k}^{|C|} Pr(C_k) \prod_{i=1}^{n} P(f_i|C_k)}$$

4 Implementation

Several hash tables were used in this implementation to facilitate O(1) search time and the following algorithms were used:

1. void train(String filename)– This method performed the necessary pre-processing to calculate probabilities. For discrete data, probabilities were calculated using conditional probability. For continuous data, probabilities were calculated with the Cumulative Distribution Function

2. void test(String filename)- This method splits a given line in the file and calls the classify(f1,f2,f3,f4,f5) method to predict a class of either 0 or 1.

3. String classify(String f1, String f2, String f3, double f4, double f5)- This method is called by the test method with the 5 features on a given line as parameters and attempts to predict which class is more likely using the probabilities calculated in the train method.

5 Evaluation

Due to the training dataset containing heavy bias towards class 0, class 1 is only predicted once and the prediction is wrong. The Naive Bayes classification algorithm is only as good as its training data.

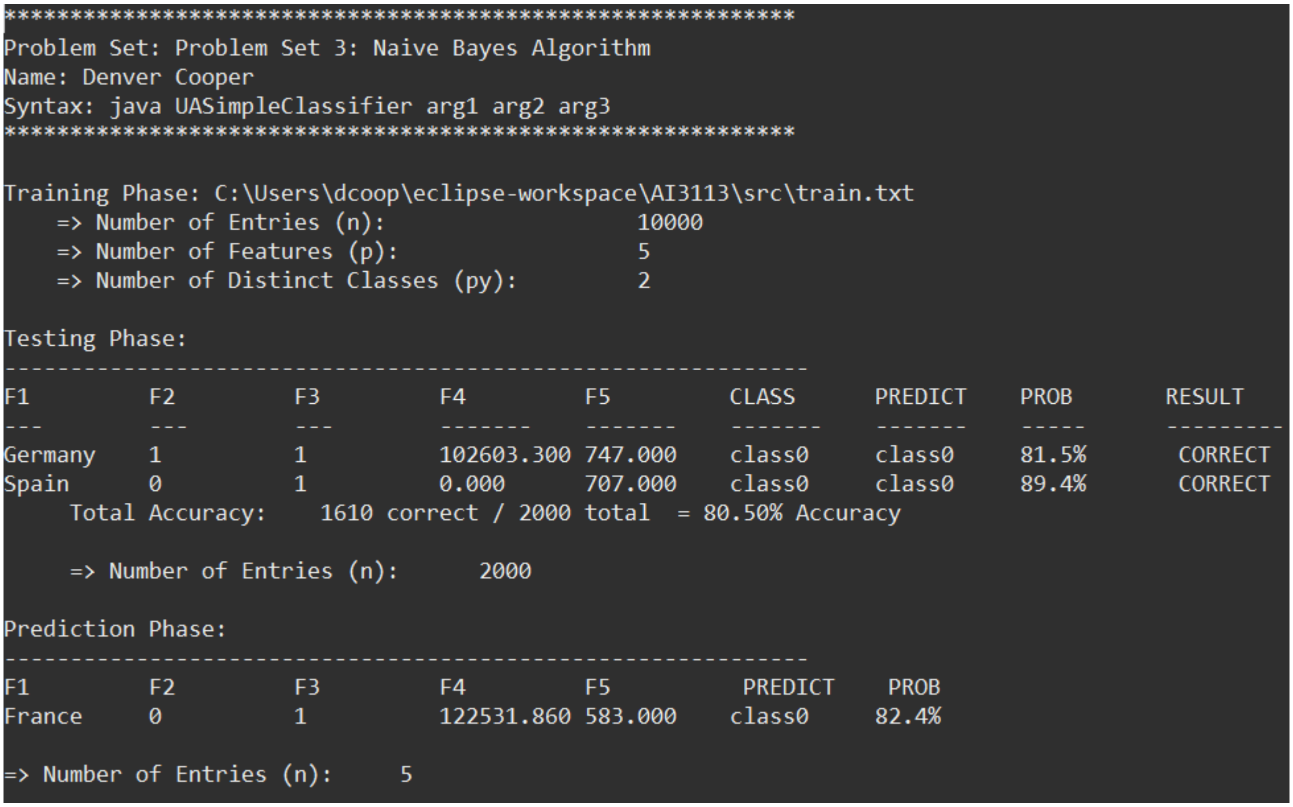
6 Conclusions

At first glance it would seem that this implementation was unsuccessful due to only class 0 being predicted. However this was simply due to the nature of the training data and the probabilities were in fact correct

Bibliography

[Moh15] Vijayarani Mohan. "Liver Disease Prediction using SVM and Naïve Bayes Algorithms". In: (Apr. 2015).

[Hua+21] Rui Hua et al. *Improved Naive Bayes Classification Algorithm for Traffic Risk Management.* Mar. 2021. DOI: 10.21203/rs.3.rs-355037/v1.



```
************************************************************
Problem Set: Problem Set 3: Naive Bayes Algorithm
Name: Denver Cooper
Syntax: java UASimpleClassifier arg1 arg2 arg3
************************************************************

Training Phase: C:\Users\dcoop\eclipse-workspace\AI3113\src\train.txt
    => Number of Entries (n):                   10000
    => Number of Features (p):                  5
    => Number of Distinct Classes (py):         2

Testing Phase:
-------------------------------------------------------------
F1         F2         F3         F4         F5         CLASS      PREDICT    PROB       RESULT
---        ---        ---        -------    -------    -------    -------    -----      ---------
Germany    1          1          102603.300 747.000    class0     class0     81.5%       CORRECT
Spain      0          1          0.000      707.000    class0     class0     89.4%       CORRECT
     Total Accuracy:    1610 correct / 2000 total  = 80.50% Accuracy

     => Number of Entries (n):      2000

Prediction Phase:
-------------------------------------------------------------
F1         F2         F3         F4         F5          PREDICT    PROB
France     0          1          122531.860 583.000    class0     82.4%

=> Number of Entries (n):     5
```