Assignment 4: REPORT
NAÏVE BAYES (NB) CLASSIFIER

Siddharth Saklecha

201505570

Breast Cancer Wisconsin (Diagnostic) Data Set

About Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

URL: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

RESULTS

Best Accuracy: 67.74193548387096
Mean Accuracy: 60.48212754685352
Deviation Accuracy: 4.450932267824146
Misclassified Data:
[555977,5,6,6,8,6,10,4,10,4,4,
1017023,4,1,1,3,2,1,3,1,1,2,
1182404,5,1,4,1,2,1,3,2,1,2]
Confusion Matrix
1867 1029
319 195

Bank Marketing Data Set

Data Set Information:
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The dataset: bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Url: http://archive.ics.uci.edu/ml/datasets/Adult

RESULTS

Best Accuracy: 88.98916168989162
Mean Accuracy: 88.80424684804245
Deviation Accuracy: 0.11366325575935866
Misclassified Data:
["management","married","tertiary","no","no","no","cellular","apr","failure","yes",
"blue-collar","married","secondary","no","yes","no","cellular","jun","other","yes",
"self-employed","married","unknown","no","yes","no","cellular","may","unknown","yes"]
Confusion Matrix
7704 6618
18690 193038

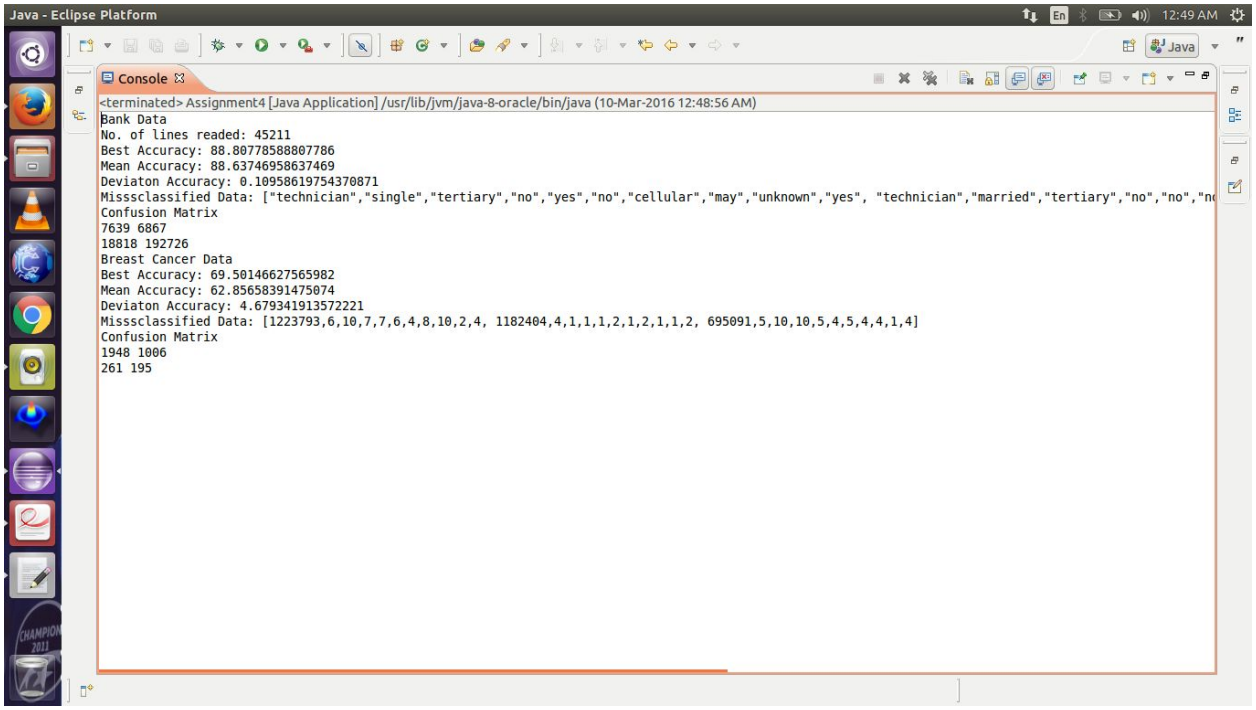

Figure1:Shows Result of the Bayes Classifier

1. Take three example records that are misclassified in each dataset and explain why these were misclassified.

Sol:

Dataset1

- ["management","married","tertiary","no","no","no","cellular","apr","failure","yes"]
  Probability log value for class “no”: 3.56164779388
  Probability log value for class “yes”:2.9350996696

- ["blue-collar","married","secondary","no","yes","no","cellular","jun","other","yes"]
  Probability log value for class “no”: 3.29727042259
  Probability log value for class “yes”:2.42947375208

- ["self-employed","married","unknown","no","yes","no","cellular","may","unknown","yes"]
  Probability log value for class “no”: 2.09260512018
  Probability log value for class “yes”: 2.71557708774

Dataset2:

- [555977,5,6,6,8,6,10,4,10,4,4] Malignant classified as benign.
  The probability values of benign: 10^7.59756673253

The probability values of malignant: 10^9.98954922206

- [1017023,4,1,1,3,2,1,3,1,1,2] benign as malignant
  The probability values of benign: 10^16.4073116499
  The probability values of malignant: 10^8.85222322607

- [1182404,5,1,4,1,2,1,3,2,1,2] benign as malignant
  The probability values of benign: 10^14.7672756483
  The probability values of malignant: 10^9.00990137652

2. What is the role of the following Laplacian smoothing used in the pseudocode for estimating posterior probabilities?

Sol: For estimating the posterior probability for a record with n features, we calculate the probability of each of the n attributes value given the class i.e.

P(C | x) = (P(C) / P(x)) * (P (x1 = v1 | C) * P (x2 = v2 | C) ** P (xn = vn | C)).

The value that we get is directly proportional to the multiplication of the probabilities of the value vi being present in attribute xi in the class C. So if any one of the value in above equation becomes 0 then the complete posterior probability will evaluate to 0. This case will emerge when some value vk is not present in attribute xk. In this case the P (xk = vk | C) = 0. This will make P(C | x) = 0 i.e. posterior probability = 0.

To fix this Laplace smoothing is used where in instead of counting the non-existent value vk in the attribute xk for class C as 0; this value is smoothed and given very small value. This small value is given by the formula

This is done with an assumption that any unknown instance will have probability = 1 / (n + |dataset|). Instead of giving straight probability = 0 to such instance we give them the least probability 1 / (n + |dataset|).

3. Briefly explain what modifications you would suggest in order to build an NB classifier dealing with mixed data (consisting of both continuous and discrete features) in the first dataset (Adult Dataset).

Sol: To handle both discrete and continuous attributes is to convert continuous data into discrete data. To accomplish this we create various bins (each bounded by a range) and then depending on the value of the continuous data, a bin gets associated with it.And then apply the discrete probability function of Naïve Baye.

4. What procedure would you suggest for considering missing values (and not discarding them!)?

Sol: Use most frequent term: We can replace the missing fields with the most frequent term of the attribute.This approach is suitable because the probability of most frequent item occurring in place of missing field is higher than any other value.

Use mean: We can replace the missing fields with the mean.The disadvantage in this approach is in case of discrete values, if the mean is such a value that does not exist in the dataset for that column, then the assumption of missing value being mean will affect results.