**UC Irvine BANA 273 Assignment 2**

**Q1. Classification using Naïve Bayes (7 points)**

For this question, you should be working with the data set called affiliation. This data set includes votes of each of the U.S. House of Representatives (or Congressmen/women) on the 16 key votes identified as 16 different attributes in the data set. We are going to use a portion of this data set (marked for training) to train our classification model. The goal of the classification is simple: given the stands (votes) of an individual congressman/woman, can we predict his/her party affiliation? As to be expected with any real-world dataset, there are several records with NULL values. However, we have taken a subset of the dataset with only those records that do not have a NULL value. The table training-no-NULL should be used to train (build) classification models, and the table testing-no-NULL should be used to test the results. The appropriate data files are provided on Canvas. You can use Excel for parts (a), (b) and (c). Use Weka for part (d).

(a) Prepare a contingency table or frequency (count) chart for the data set and populate it based on the training data. See examples covered in class. A frequency chart shows cross-tab of class variables (i.e., party affiliation) with each of the other attributes. Use Excel Pivot tables for this – you may need several pivot tables.

(b) Prepare a populated probability chart (conditional probability) from the frequency chart in part (a). Again see examples covered in class.

(c) Based on the probability chart, apply Naïve Bayesian classification to predict the party affiliation of the following two congressmen based on their voting records:

(i) y, n, y, n, n, n, y, y, y, n, n, n, n, n, y, y

(ii) n, y, n, y, y, y, n, n, n, n, n, y, y, y, n, y

(d) Use WEKA to run the Naïve Bayes classifier on training-no-NULL.ARFF and set test file as testing-no-NULL.ARFF (use “Supplied test set” option to upload the test set). Report the confusion matrix output by Weka.

**A1. Classification using Naïve Bayes (7 points)**

**a)**

|  |  |  |  |
| --- | --- | --- | --- |
| Handinfan |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 50 | 71 | 121 |
| republican | 83 | 23 | 106 |
| **Grand Total** | **133** | **94** | **227** |
|  |  |  |  |
| water |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 67 | 54 | 121 |
| republican | 57 | 49 | 106 |
| **Grand Total** | **124** | **103** | **227** |
|  |  |  |  |
| budget |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 18 | 103 | 121 |
| republican | 89 | 17 | 106 |
| **Grand Total** | **107** | **120** | **227** |
|  |  |  |  |
| physfee |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 116 | 5 | 121 |
| republican | 1 | 105 | 106 |
| **Grand Total** | **117** | **110** | **227** |
|  |  |  |  |
| Salvador |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 97 | 24 | 121 |
| republican | 5 | 101 | 106 |
| **Grand Total** | **102** | **125** | **227** |
|  |  |  |  |
| relinsch |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 68 | 53 | 121 |
| republican | 14 | 92 | 106 |
| **Grand Total** | **82** | **145** | **227** |
|  |  |  |  |
| sattest |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 28 | 93 | 121 |
| republican | 77 | 29 | 106 |
| **Grand Total** | **105** | **122** | **227** |
|  |  |  |  |
| contras |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 20 | 101 | 121 |
| republican | 90 | 16 | 106 |
| **Grand Total** | **110** | **117** | **227** |
|  |  |  |  |
| missile |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 25 | 96 | 121 |
| republican | 91 | 15 | 106 |
| **Grand Total** | **116** | **111** | **227** |
|  |  |  |  |
| imm |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 55 | 66 | 121 |
| republican | 44 | 62 | 106 |
| **Grand Total** | **99** | **128** | **227** |
|  |  |  |  |
| synfuel |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 60 | 61 | 121 |
| republican | 89 | 17 | 106 |
| **Grand Total** | **149** | **78** | **227** |
|  |  |  |  |
| edu |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 106 | 15 | 121 |
| republican | 15 | 91 | 106 |
| **Grand Total** | **121** | **106** | **227** |
|  |  |  |  |
| rttosue |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 86 | 35 | 121 |
| republican | 17 | 89 | 106 |
| **Grand Total** | **103** | **124** | **227** |
|  |  |  |  |
| crime |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 80 | 41 | 121 |
| republican | 2 | 104 | 106 |
| **Grand Total** | **82** | **145** | **227** |
|  |  |  |  |
| export |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 49 | 72 | 121 |
| republican | 94 | 12 | 106 |
| **Grand Total** | **143** | **84** | **227** |
|  |  |  |  |
| exportsa |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 7 | 114 | 121 |
| republican | 36 | 70 | 106 |
| **Grand Total** | **43** | **184** | **227** |

**Using Weka:**

**Class**

**Attribute republican democrat**

**(0.47) (0.53)**

**=====================================**

**handinfan**

**y 24.0 72.0**

**n 84.0 51.0**

**[total] 108.0 123.0**

**water**

**y 50.0 55.0**

**n 58.0 68.0**

**[total] 108.0 123.0**

**budget**

**y 18.0 104.0**

**n 90.0 19.0**

**[total] 108.0 123.0**

**physfee**

**y 106.0 6.0**

**n 2.0 117.0**

**[total] 108.0 123.0**

**salvador**

**y 102.0 25.0**

**n 6.0 98.0**

**[total] 108.0 123.0**

**relinsch**

**y 93.0 54.0**

**n 15.0 69.0**

**[total] 108.0 123.0**

**sattest**

**y 30.0 94.0**

**n 78.0 29.0**

**[total] 108.0 123.0**

**contras**

**y 17.0 102.0**

**n 91.0 21.0**

**[total] 108.0 123.0**

**missile**

**y 16.0 97.0**

**n 92.0 26.0**

**[total] 108.0 123.0**

**imm**

**y 63.0 67.0**

**n 45.0 56.0**

**[total] 108.0 123.0**

**synfuel**

**y 18.0 62.0**

**n 90.0 61.0**

**[total] 108.0 123.0**

**edu**

**y 92.0 16.0**

**n 16.0 107.0**

**[total] 108.0 123.0**

**rttosue**

**y 90.0 36.0**

**n 18.0 87.0**

**[total] 108.0 123.0**

**crime**

**y 105.0 42.0**

**n 3.0 81.0**

**[total] 108.0 123.0**

**export**

**y 13.0 73.0**

**n 95.0 50.0**

**[total] 108.0 123.0**

**exportsa**

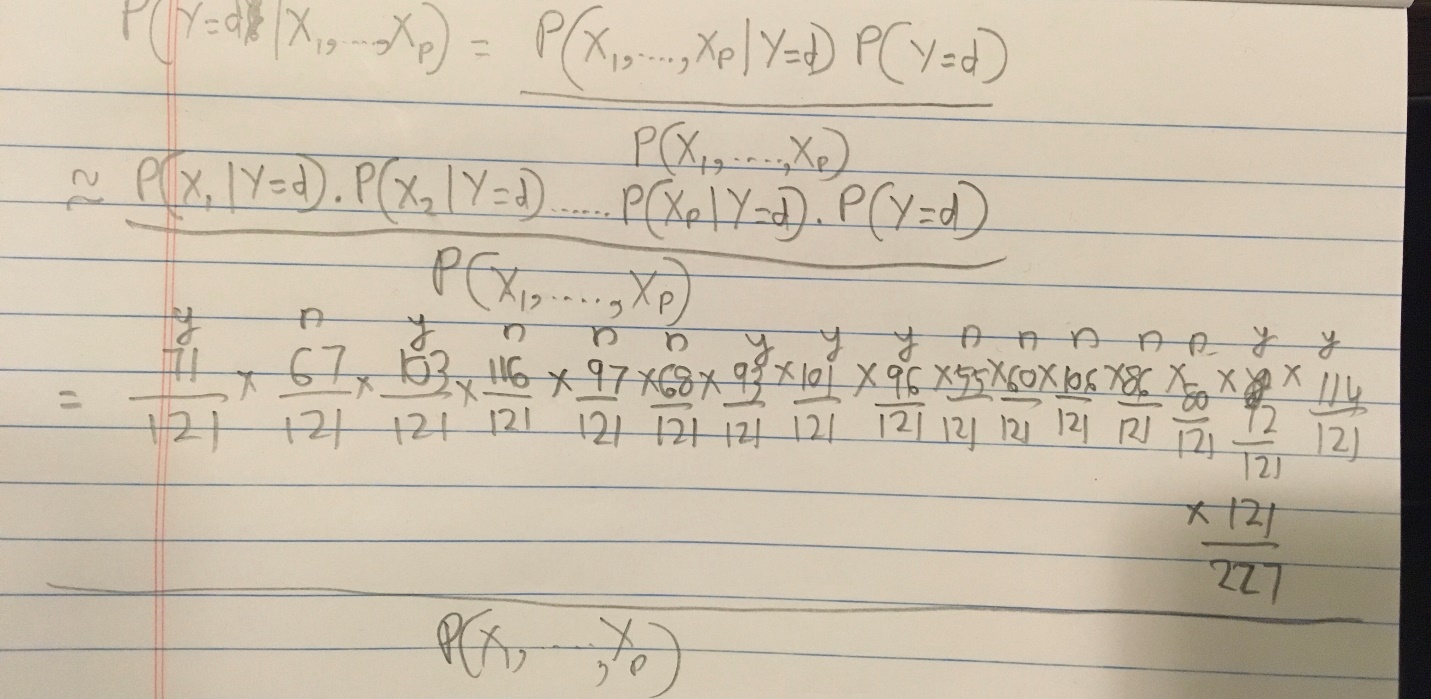
**y 71.0 115.0**

**n 37.0 8.0**

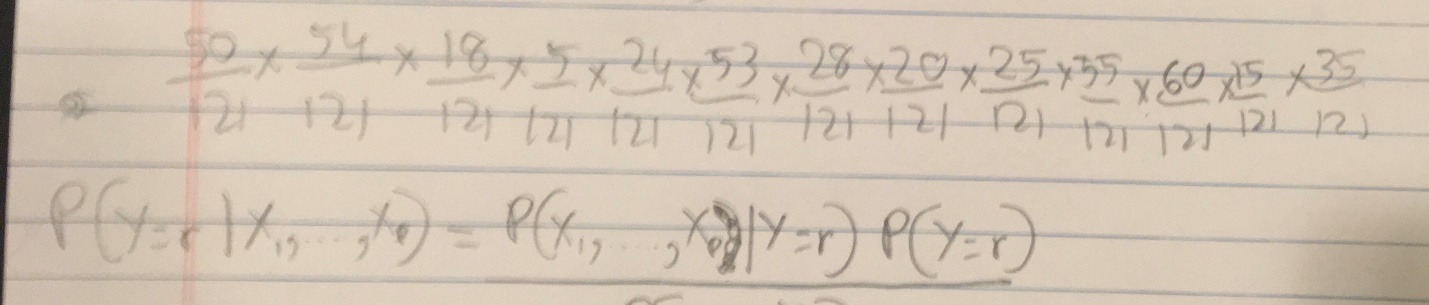
**[total] 108.0 123.0**

**b)**

|  |  |  |  |
| --- | --- | --- | --- |
| Handinfan |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 37.59% | 75.53% | 53.30% |
| republican | 62.41% | 24.47% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| water |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 54.03% | 52.43% | 53.30% |
| republican | 45.97% | 47.57% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| budget |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 16.82% | 85.83% | 53.30% |
| republican | 83.18% | 14.17% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| physfee |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 99.15% | 4.55% | 53.30% |
| republican | 0.85% | 95.45% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| Salvador |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 95.10% | 19.20% | 53.30% |
| republican | 4.90% | 80.80% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| relinsch |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 82.93% | 36.55% | 53.30% |
| republican | 17.07% | 63.45% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| sattest |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 26.67% | 76.23% | 53.30% |
| republican | 73.33% | 23.77% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| contras |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 18.18% | 86.32% | 53.30% |
| republican | 81.82% | 13.68% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| missile |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 21.55% | 86.49% | 53.30% |
| republican | 78.45% | 13.51% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| imm |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 55.56% | 51.56% | 53.30% |
| republican | 44.44% | 48.44% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| synfuel |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 40.27% | 78.21% | 53.30% |
| republican | 59.73% | 21.79% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| edu |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 87.60% | 14.15% | 53.30% |
| republican | 12.40% | 85.85% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| rttosue |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 83.50% | 28.23% | 53.30% |
| republican | 16.50% | 71.77% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| crime |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 97.56% | 28.28% | 53.30% |
| republican | 2.44% | 71.72% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| export |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 34.27% | 85.71% | 53.30% |
| republican | 65.73% | 14.29% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |
|  |  |  |  |
| exportsa |  |  |  |
| **Count of party** | **Column Labels** |  |  |
| **Row Labels** | **n** | **y** | **Grand Total** |
| democrat | 16.28% | 61.96% | 53.30% |
| republican | 83.72% | 38.04% | 46.70% |
| **Grand Total** | **100.00%** | **100.00%** | **100.00%** |

**c) 1) **

It is evident for the first example that the congressman is a democrat. Most of the values of P(X1,2,3……p| Y=d) are greater than 0.5 and for P(X1,2,3……p| Y=r) are lesser than 0.5. This shows that the probability of the congressman being a democrat is higher.

**2) **

It is evident for the second example that the congressman is a republican. Most of the values of P(X1,2,3……p| Y=r) are greater than 0.5 and for P(X1,2,3……p| Y=d) are lesser than 0.5. This shows that the probability of the congressman being a republican is higher.

**d) === Confusion Matrix ===**

**a b <-- classified as**

**2 0 | a = republican**

**1 2 | b = democrat**

**Q2. Model Testing and Evaluation (3 points)**

After running a classifier in WEKA on some dataset, the following confusion matrix was obtained:

==========Confusion Matrix=========

a b ← classified as

921 28 |a=yes

17 374 |b=no

(a) Based on this confusion matrix, estimate the overall accuracy of the classifier.

(b) Estimate the stratified accuracies of the classifier.

(c) Consider the following cost/benefit scenario: The company gains $80 from a correctly classified class-a instance, but loses $5 from an incorrectly classified class-a instance (i.e., a class-b instance incorrectly classified as class-a). The company incurs no benefit or loss from an instance classified as class b. What expected value per instance (e.g., per customer) would this classifier create for the company?

**A2. Model Testing and Evaluation (3 points)**

1. Accuracy = (921 + 374)/(921+374+28+17) = 0.96641791044
2. Stratified Accuracy

– Accuracy for “a” = 921/(921+28) = 97.05%  
– Accuracy for “b” = 374/(17+374) = 95.65%

1. Expected value = (921\*80-28\*5)/(921+28+17+374) = 54.88