Homework9

1.

1.1 Predictive accuracy: Prediction accuracy is expressed as the correlation between the AMS prediction and the actual score. Accuracy of 1 indicates a perfect accuracy, whereas the accuracy of 0 indicates a random guess.

1.2 The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. Place the best attribute of the dataset at the root of the tree. Split the training set into subsets.

## 1.3 Evaluation methods are the criteria for evaluating the success of a program or project.

The three main types of evaluation methods are goal based, process based and outcomes based. Goal based evaluations measure if objectives have been achieved. Process based evaluations analyze strengths and weaknesses. Outcomes based evaluations examine broader impacts and often investigate what greater good was served as a result of the program or project.

1.4

classification trees, logistic regression, discriminant analysis, neural networks, boosted trees, random forests, deep learning methods, nearest neighbors, support vector machines, etc.

Classification Algorithms could be broadly classified as the following:

* Linear Classifiers
  + Logistic regression
  + Naive Bayes classifier
  + Fisher’s linear discriminant
* Support vector machines
  + Least squares support vector machines
* Quadratic classifiers
* Kernel estimation
  + k-nearest neighbor
* Decision trees
  + Random forests
* Neural networks
* Learning vector quantization

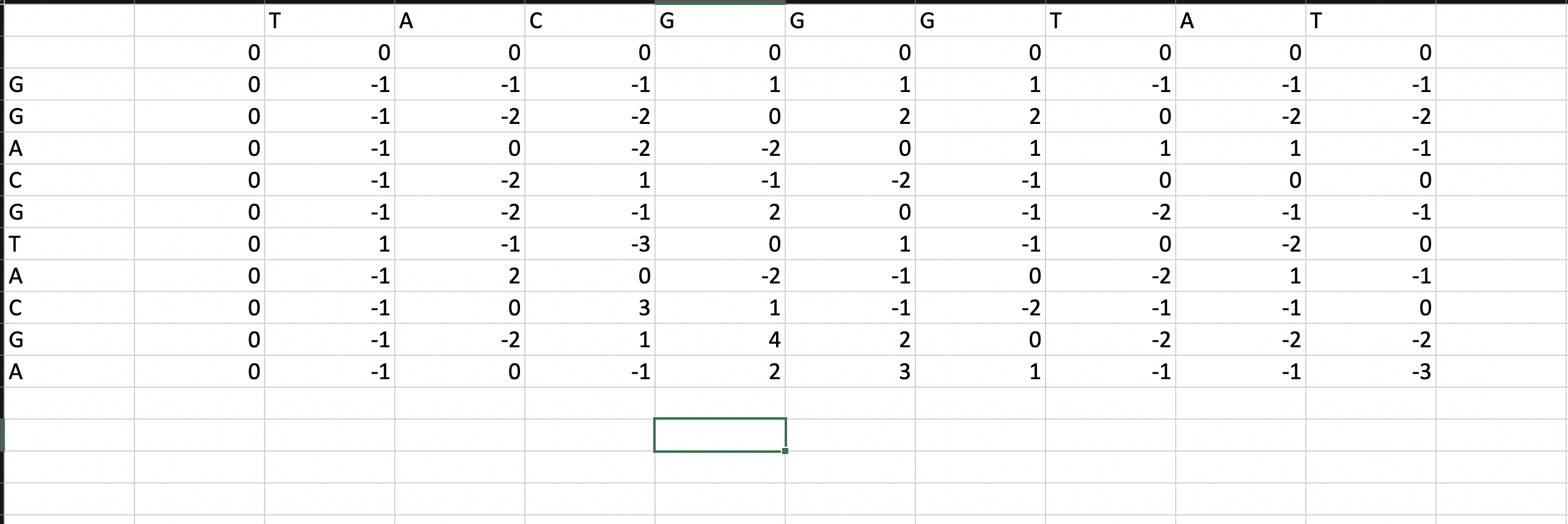
1.5 The ranking of establishments is a comparison of individual establishment scores within their risk level, or in some cases, the type of establishment. Establishment scoring is based on specific violation points assigned to each violation.

1.6 A lift curve is a way of visualizing the performance of a classification model. Lift curves are closely related to, and frequently confused with, cumulative gains charts. A cumulative gains chart shows the total number of events captured by a model over a given number of samples

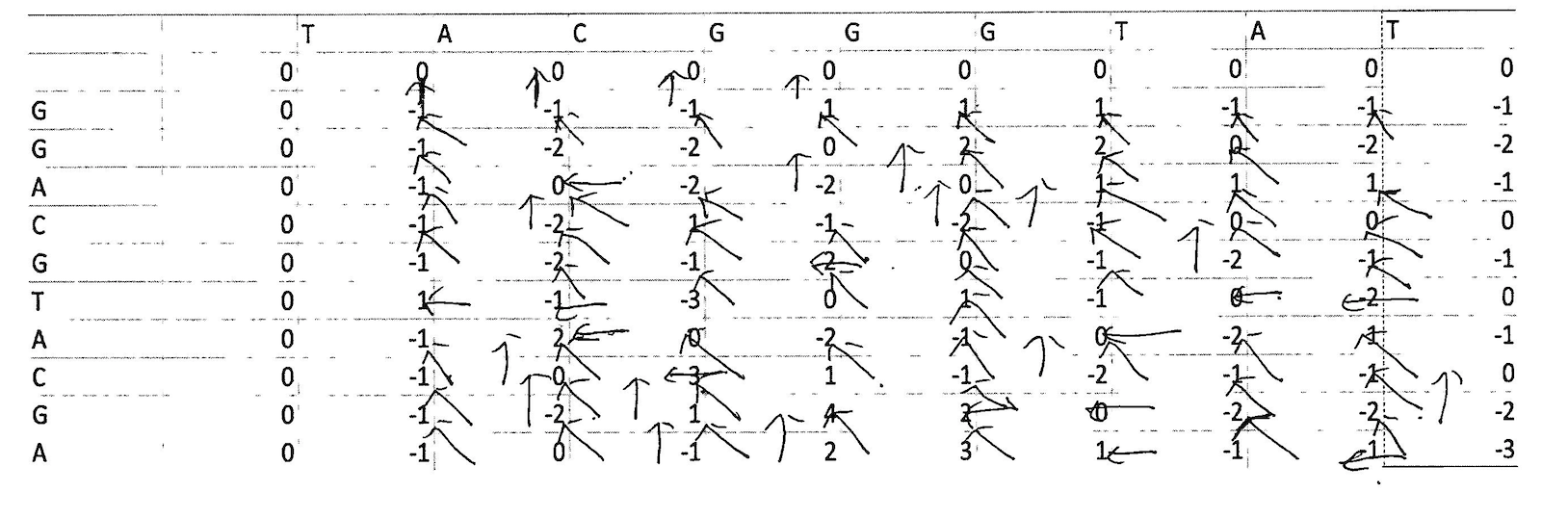
1.7   
It is derived from classical mathematical theory and has a solid mathematical foundation and stable classification efficiency. At the same time, the NBC model requires few parameters to estimate, is less sensitive to missing data, and the algorithm is relatively simple. In theory, the NBC model has the smallest error rate compared to other classification methods.

2.

A):



B):



C): -----TACGGGTAT

GGACGTACGA----

3.Yes

4.

a:

Split the entire data randomly into 4 folds

Then fit the model using the 3 (K minus 1) folds and validate the model using the remaining Kth fold. Note down the scores/errors.

Repeat this process until every 4-fold serve as the test set. Then take the average of your recorded scores. That will be the performance metric for the model.

b):

Cross-validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

5. A classifier is a machine learning model that is used to discriminate different objects based on certain features.

## Multinomial Naive Bayes:

This is mostly used for document classification problem, i.e whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

## Bernoulli Naive Bayes:

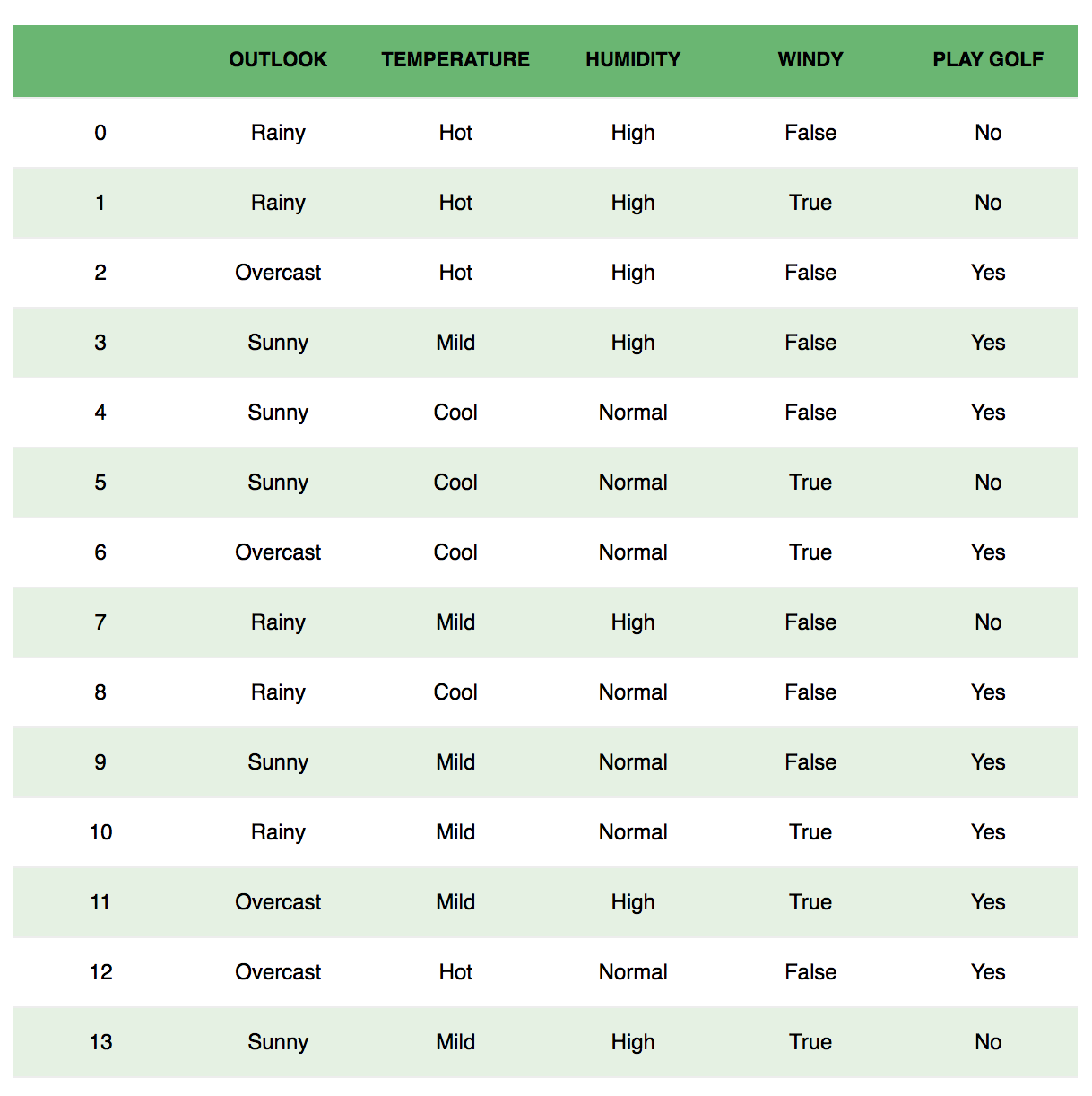
This is similar to the multinomial naive bayes but the predictors are boolean variables. The parameters that we use to predict the class variable take up only values yes or no, for example if a word occurs in the text or not.

## Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution.

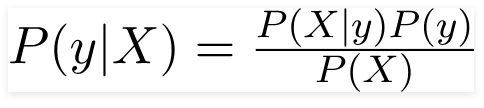
# Example:

Let us take an example to get some better intuition. Consider the problem of playing golf. The dataset is represented as below.



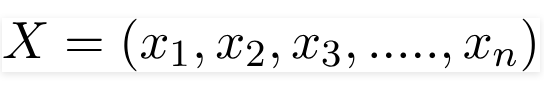
We classify whether the day is suitable for playing golf, given the features of the day. The columns represent these features and the rows represent individual entries. If we take the first row of the dataset, we can observe that is not suitable for playing golf if the outlook is rainy, temperature is hot, humidity is high and it is not windy. We make two assumptions here, one as stated above we consider that these predictors are independent. That is, if the temperature is hot, it does not necessarily mean that the humidity is high. Another assumption made here is that all the predictors have an equal effect on the outcome. That is, the day being windy does not have more importance in deciding to play golf or not.

According to this example, Bayes theorem can be rewritten as:

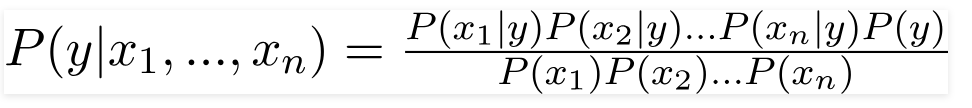


The variable **y** is the class variable(play golf), which represents if it is suitable to play golf or not given the conditions. Variable **X**represent the parameters/features.

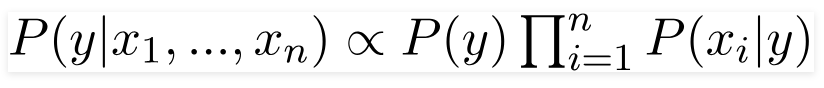
**X** is given as,



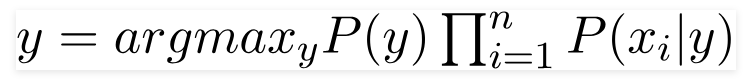
Here x\_1,x\_2….x\_n represent the features, i.e they can be mapped to outlook, temperature, humidity and windy. By substituting for **X**and expanding using the chain rule we get,



Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all entries in the dataset, the denominator does not change, it remain static. Therefore, the denominator can be removed and a proportionality can be introduced.



In our case, the class variable(**y**) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we need to find the class **y** with maximum probability.



Using the above function, we can obtain the class, given the predictors.

6.

b) \*\* Naive Bayes Evaluation with Datasets \*\*

Correctly Classified Instances 7 100 %

Incorrectly Classified Instances 0 0 %

Kappa statistic 1

Mean absolute error 0.1378

Root mean squared error 0.1444

Relative absolute error 28.0006 %

Root relative squared error 29.1716 %

Total Number of Instances 7

the expression for the input data as per alogorithm is The independent probability of a class

--------------------------------------

spam 0.5555555555555556

ham 0.4444444444444444

The probability of a word given the class

-----------------------------------------

spam ham

Congrats 0.07407407407407407 0.043478260869565216

cards 0.07407407407407407 0.043478260869565216

credit 0.07407407407407407 0.043478260869565216

for 0.11111111111111109 0.043478260869565216

free 0.07407407407407407 0.043478260869565216

lottery 0.11111111111111109 0.043478260869565216

selected 0.07407407407407407 0.08695652173913045

travel 0.07407407407407407 0.043478260869565216

won 0.07407407407407407 0.043478260869565216

you 0.07407407407407407 0.08695652173913045

Congratulation 0.037037037037037035 0.08695652173913045

Good 0.037037037037037035 0.13043478260869565

are 0.037037037037037035 0.08695652173913045

night 0.037037037037037035 0.08695652173913045

very 0.037037037037037035 0.08695652173913045

{0 ?,1 1,2 1,3 1}

spam

The probabilities are calculated using the word frequencies. It is calculated using some basic properties of probabilities and the Bayes theorem. The conditional probabilities like the one discussed here will suits for the Bayes theorem.

P (A | B) = (P (B | A) \* P(A))/ P(B)

Let us do one sample example to understand the probability calculation.

P (ham| you won lottery) = (P (you won lottery | ham) \* P(ham))/ P (you won lottery)

In our classifier, we are trying to find out the category which has the bigger probability.so we can discard the divisor and perform how many times the sentence "you won lottery" appears in ham category divide it by the total and obtain P (you won lottery | ham)

we here assume that every word in a sentence is independent of the other ones.so, we are not considering entire sentences but rather at individual’s words. Example: "Good things are “and "things are Good" are same.

P (you won lottery | ham) = P (you | ham) \* P (won | ham) \* P (lottery | ham)

P (you won lottery | spam) = P (you | spam) \* P (won | spam) \* P (lottery | spam)

with this above formula, we will find which one this sentence belongs to.

There are cases, some of the words will not be in training sets and the value is 0. The multiplication of 0 with others will be 0. To avoid this, we are going to use Laplace smoothing. i.e., we will add 1 to every count so it will not be zero

The words are listed in the above table. The total number of words: 15 The total number of ham words: 7 The total number of spam words: 10

P (you | ham) = (1 + 1)/ (7 + 15)

P (you | spam) = (1 + 1)/ (10 + 15)

Like the above we will calculate for every word and multiply using the above probability. The results for the one who have higher probability won.