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

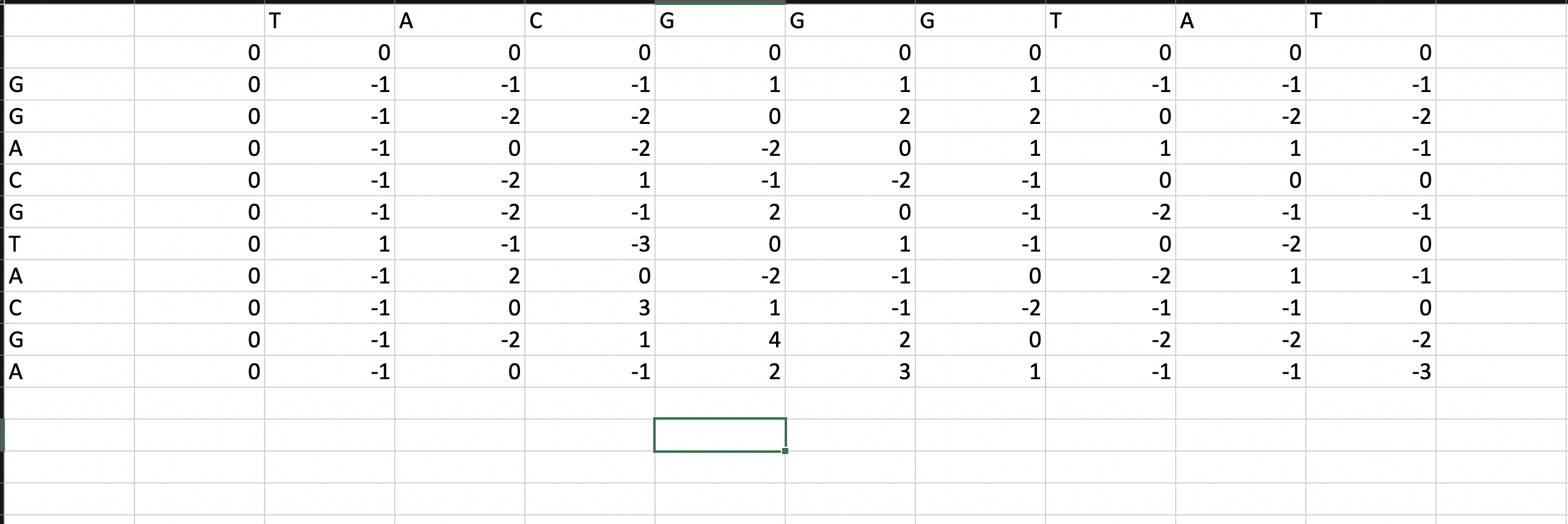
1.5 The ranking of establishments is a comparison of individual establishment scores within their risk level, or in some cases, the type of establishment. Establishment scoring is based on specific violation points assigned to each violation.

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

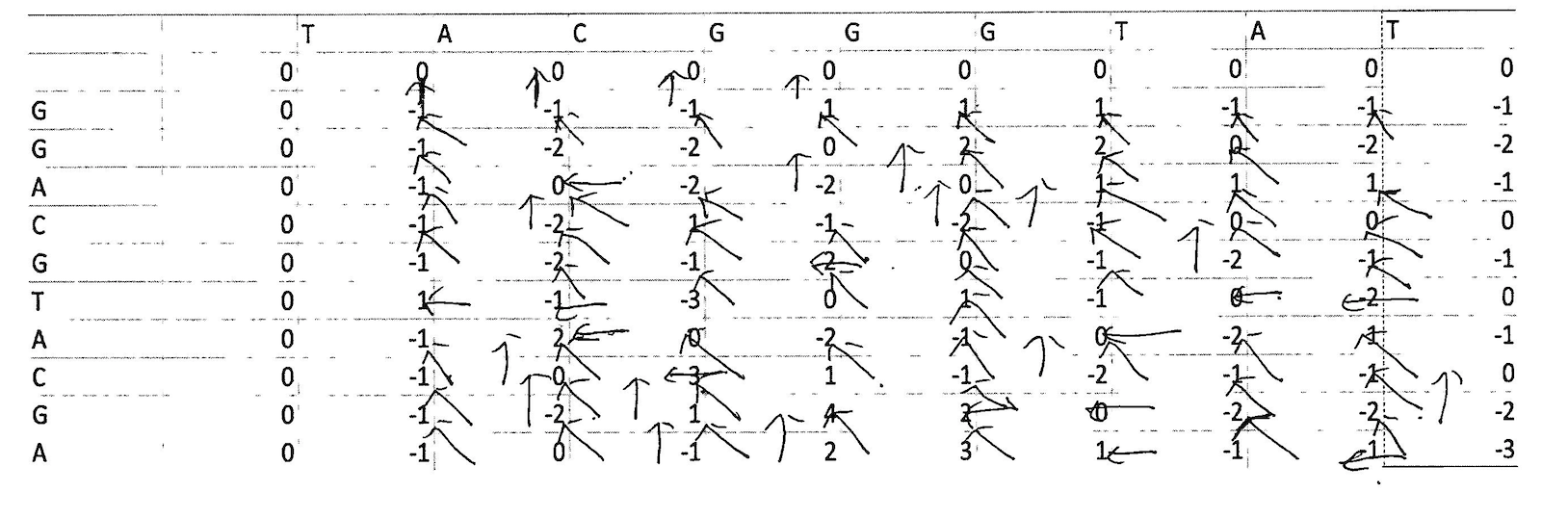
1.7   
It is derived from classical mathematical theory and has a solid mathematical foundation and stable classification efficiency. At the same time, the NBC model requires few parameters to estimate, is less sensitive to missing data, and the algorithm is relatively simple. In theory, the NBC model 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 k folds (value of k shouldn’t be too small or too high, ideally we choose 5 to 10 depending on the data size). The higher value of K leads to less biased model (but large variance might lead to overfit), where as the lower value of K is similar to the train-test split approach we saw before.

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

Repeat this process until every K-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.