1. Describe these concepts:
2. Predictive Accuracy?

Ans: **Predictive 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. What is the key in building a decision tree?

Ans: In order to branch a tree, the **attribute chosen as a root element** is the key to build a decision tree.

1. What is evaluation method?

Ans: **Evaluation methods** are the methods for evaluating the **accuracy and performance** of a machine learning model.

1. Describe ALL evaluation classification methods and their differences?

Ans: The evaluation classification methods are:

**Holdout Sets:** The available data set D is divided into two disjoint subsets.

This is used for large datasets.

**N fold cross validation:** The data is partitioned into n equal sized disjoint subsets.

This is used when available data isn’t large.

**Leave one out cross validation:** This method is used when dataset is very small.

It’s a special case of cross validation.

**Validation Set:** The available data is divided into three subsets i.e. a training set, a validation set and a test set. A validation set is used frequently for estimating parameters in learning algorithms.

1. Scoring and Ranking method, how does it work?

Ans: **Scoring method:** It’s related to classification.

Instead of assigning each test instance a definite class, scoring assigns a probability estimate to indicate a likelihood that the example belongs to the positive class.

**Ranking method:** After each data sample is given a probability estimate score, we can rank all examples according to their probability estimate is called ranking method.

1. Lift Analysis Curve

Ans: A lift curve is a way of **visualising the performance of a classification model**. It is basically:

* group data based on the predicted churn probability (value between 0.0 and 1.0). Typically, you look at deciles, so you'd have 10 groups: 0.0 - 0.1, 0.1 - 0.2, ..., 0.9 - 1.0
* calculate the true churn rate per group. That is, you count how many people in each group churned and divide this by the total number of customers per group.

1. How does Naive Bayes is different from other Evaluation Classification methods?

Ans: Naive Bayes is faster than other evaluation classification methods because it assumes that the presence of a particular feature in a class is unrelated to any other feature.

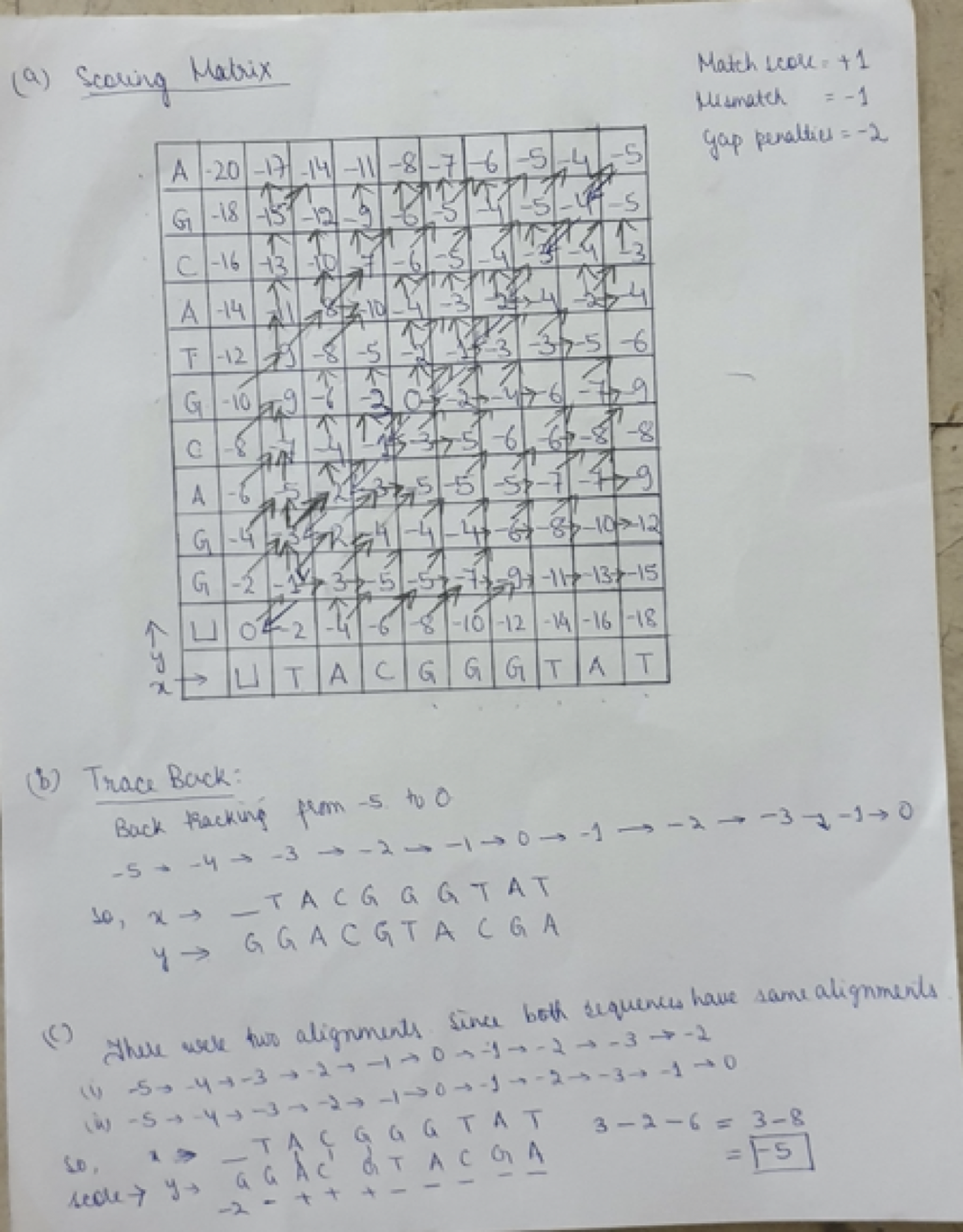
Naive Bayes model is easy to build and particularly useful for very large data sets.

Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

2. Consider the sequences x = TACGGGTAT and y = GGACGTACGA. Assume that the match score is +1, and the mismatch is -1, and gap penalties is -2.

* 1. Fill out the dynamic programming table for a global alignment between x and y.
  2. Draw arrows in the cells to store traceback information.
  3. What is the score of the optimal global alignment and what alignment(s) achieves this score?

Ans:

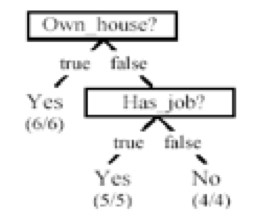


3. Consider this decision tree. Is this an optimal decision tree?

Data: Loan application data

Task: Predict whether a loan should be approved or not.

Performance measure: accuracy



Ans: Yes, this is an optimal decision tree because the root element chosen provides high information gain and high accuracy rate.

4. Consider n-fold cross-validation method:

a) How does algorithm work for Training and Test

<https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f>

Ans: In order to validate a model, we make use of validation. Out of which is n- fold cross validation. Incase of training and test for a large dataset, we divide the dataset randomly into training and test sets in the ratio of 70:30 or 80:20, considering both test and train sets have same distribution. Then later perform model training on training sets and test set for validation.

Algorithm:

* The available data us partitioned into n equal size disjoint subsets.
* Use each subset as the best set and combine the rest n-1 subsets as the training set to leave a classifier.
* The final estimated accuracy of learning is the average of n accuracies.

b) Explain this code. Compile and Run if you can, analyse the results.

<https://github.com/haifengl/smile/blob/master/core/src/main/java/smile/validation/CrossValidation.java>

Ans: The purpose is to validate a machine learning model use cross validation approach.

In Cross-validation we assess how the results of a statistical analysis will generalise an independent data set. It is used in settings where the goal is prediction. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on the training set, and validating the analysis on the 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. Naive Bayes Classification,

<https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c>

Naive Bayes three types of classifiers. What are they, explain.

How does it work? Provide an example showing results

Ans:

Naive Bayes is a classification method to classify based on probabilistic machine learning model. This is based on Bayes Theorem.

Bayes Theorem:

P(A|B) = (p(B|A)P(A)) /P(B)

Types of classifiers

**Bernoulli Naive Bayes**

Bernoulli Naive Bayes 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.

Example-

It is used to check if a word occurs in the text or not.

Consider 1000 fruits which could be either ‘banana’, ‘orange’ or ‘other’. These are the 3 possible classes of the Y variable.  
We have data for the following X variables, all of which are binary (1 or 0).

·      Long

·      Sweet

·      Yellow

The few rows of the training dataset look like this:

|  |  |  |  |
| --- | --- | --- | --- |
| **Fruit** | **Long (x1)** | **Sweet (x2)** | **Yellow (x3)** |
| Orange | 0 | 1 | 0 |
| Banana | 1 | 0 | 1 |
| Banana | 1 | 1 | 1 |
| Other | 1 | 1 | 0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Type | Long | Not Long | Sweet | Not Sweet | Yellow | Not Yellow | Total |
| Banana | 400 | 100 | 350 | 150 | 450 | 50 | 500 |
| Orange | 0 | 300 | 150 | 150 | 300 | 0 | 300 |
| Other | 100 | 100 | 150 | 50 | 50 | 150 | 200 |
| Total | 500 | 500 | 650 | 350 | 800 | 200 | 1000 |

Compute Prior probabilities for each of the class of fruits.  
P(Y=Banana) = 500 / 1000 = 0.50  
P(Y=Orange) = 300 / 1000 = 0.30  
P(Y=Other) = 200 / 1000 = 0.20  
Compute the probability of evidence that goes in the denominator.  
P(x1=Long) = 500 / 1000 = 0.50  
P(x2=Sweet) = 650 / 1000 = 0.65  
P(x3=Yellow) = 800 / 1000 = 0.80  
Compute the probability of likelihood of evidences that goes in the numerator.  
Probability of Likelihood for Banana  
P(x1=Long | Y=Banana) = 400 / 500 = 0.80  
P(x2=Sweet | Y=Banana) = 350 / 500 = 0.70  
P(x3=Yellow | Y=Banana) = 450 / 500 = 0.90  
So, the overall probability of Likelihood of evidence for Banana = 0.8 \* 0.7 \* 0.9 = 0.504

If a fruit is ‘Long’, ‘Sweet’, and ‘Yellow’. What fruit would it be?

P (Banana | Long, Sweet, Yellow) = 0.252/P(Evidence)  
P (Banana | Long, Sweet, Yellow) = 0.01875/ P(Evidence)  
P (Orange | Long, Sweet, Yellow) = 0  
  
Since, Banana has highest probability it will be predicted class.

**Multinomial Naive Bayes**

Multinomial Naive Bayes estimates the conditional probability of a word given a class as the relative frequency of term in documents belonging to class. The variation considers the number of occurrences of term t in training documents from class, including multiple occurrences.

It is used in document classification based on the frequency of certain words.

Example –

|  |  |
| --- | --- |
| TEXT | REVIEWS |
| I liked the movie | positive |
| It’s a good movie. Nice story | positive |
| Nice songs. But sadly, boring ending. | negative |
| Hero’s acting is bad, but heroine looks good. Overall nice movie | positive |
| Sad, boring movie | negative |

Applying Stemming and stop-words

|  |  |
| --- | --- |
| TEXT | REVIEWS |
| ilikedthemovi | positive |
| itsagoodmovienicestori | positive |
| nicesongsbutsadlyboringend | negative |
| herosactingisbadbutheroinelooksgoodoverallnicemovi | positive |
| sadboringmovi | negative |

Calculating Probabilities

|  |  |  |
| --- | --- | --- |
| Word | P (WORD | POSITIVE) | P (WORD | NEGATIVE) |
| overall | 1+1/17+21 | 0+1/7+21 |
| liked | 1+1/17+21 | 0+1/7+21 |
| the | 2+1/17+21 | 0+1/7+21 |
| movie | 3+1/17+21 | 1+1/7+21 |

P (overall | positive) \* P (liked | positive) \* P (the | positive) \* P (movie | positive) \* P (positive) = 1.38 \* 10^ {-5} = 0.0000138

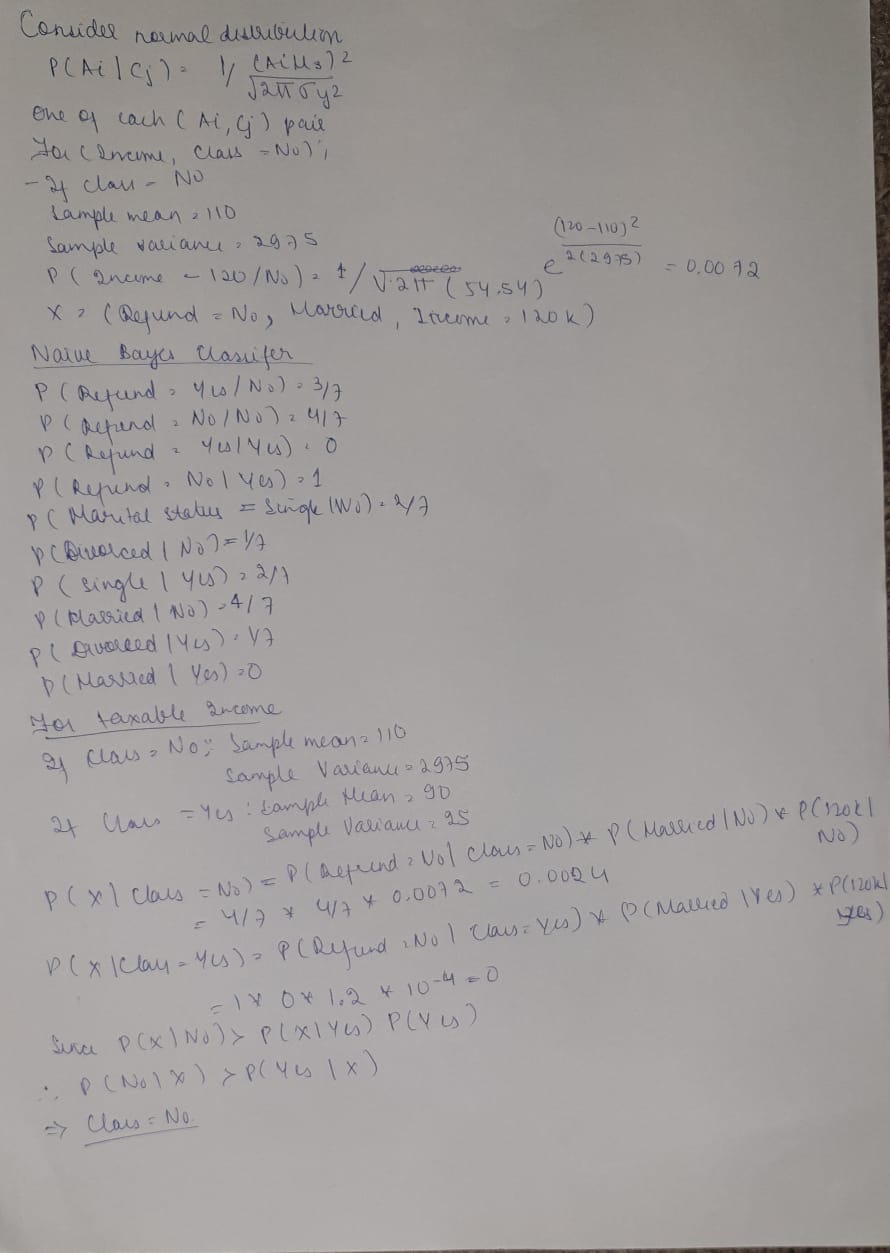
P (overall | negative) \* P (liked | negative) \* P (the | negative) \* P (movie | negative) \* P(negative) = 0.13 \* 10^ {-5} = 0.0000013

Multinomial Naive Bayes classifier gives “overall liked the movie” as the positive tag.

**Gaussian Naive Bayes**

Gaussian Naive Bayes is used when working with continuous values and not discrete in nature. we assume that these values are sampled from a gaussian distribution.

Example- It is used in image classification.

Consider the following data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tid | Refund | Marital Status | Taxable Amount | Evade |
| 1 | Yes | Single | 125 | No |
| 2 | No | Married | 100 | No |
| 3 | No | Single | 70 | No |
| 4 | Yes | Married | 120 | No |
| 5 | No | Divorced | 95 | Yes |
| 6 | No | Married | 60 | No |
| 7 | Yes | Divorced | 220 | No |
| 8 | No | Single | 85 | Yes |
| 9 | No | Married | 75 | No |
| 10 | No | Single | 90 | Yes |

6. Consider the following example using Naive Bayes classifier:

<https://www.codingame.com/playgrounds/6734/machine-learning-with-java---part-5-naive-bayes>

1. Describe Example

Ans: Based on Naive Bayes Model, this example walks us through as to how to determine whether a given sentence is a Spam or a Ham. By applying Bayes theorem and calculating the conditional probability of each word based on the word frequency we can determine whether a sentence is Spam or Ham.

1. Run program, describe outputs

Ans: Output of the program

\* 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 algorithm 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

1. Take the Java code and build it in your environment

Ans: Java code available.