1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy.

2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach?

The main differences between the filter and wrapper methods for feature selection are: Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it.

4.

i. Describe the overall feature selection process.

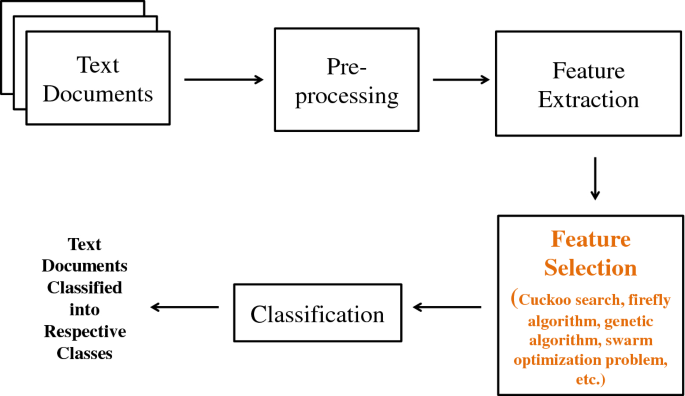
Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data.

ii. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones and then discarding the original features.

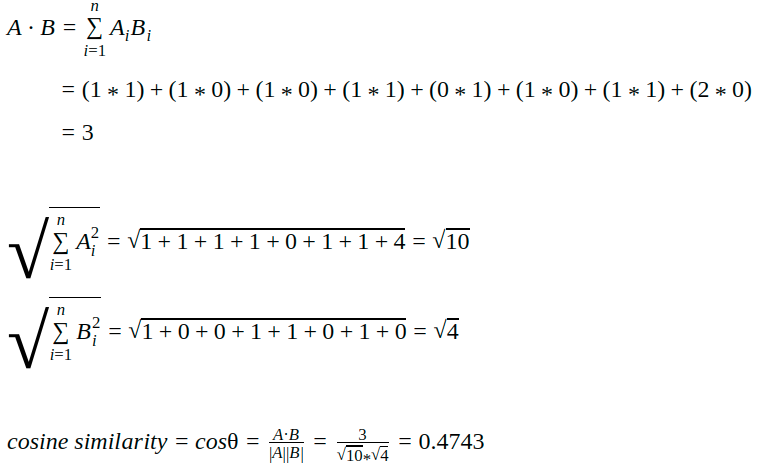
An example of a simple feature is the mean of a window in a signal. Automated feature extraction uses specialized algorithms or deep networks to extract features automatically from signals or images without the need for human intervention.

5. Describe the feature engineering process in the sense of a text categorization issue.



6. What makes cosine similarity a good metric for text categorization? A document-term matrix has two rows with values of (2, 3, 2, 0, 2, 3, 3, 0, 1) and (2, 1, 0, 0, 3, 2, 1, 3, 1). Find the resemblance in cosine.

The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity.

0.759

7.

i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

ii. Compare the Jaccard index and similarity matching coefficient of two features with values (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), respectively (1, 0, 0, 1, 1, 0, 0, 1).

8. State what is meant by "high-dimensional data set"? Could you offer a few real-life examples? What are the difficulties in using machine learning techniques on a data set with many dimensions? What can be done about it?

High-dimensional data are defined as data in which the number of features (variables observed), p, are close to or larger than the number of observations (or data points), n. Subjects like genomics and medical sciences often use both tall (in terms of n) and wide (in terms of p) datasets that can be difficult to analyse or visualise using standard statistical tools. An example of high-dimensional data in biological sciences may include data collected from hospital patients recording symptoms, blood test results, behaviours, and general health, resulting in datasets with large numbers of features. Researchers often want to relate these features to specific patient outcomes (e.g. survival, length of time spent in hospital).

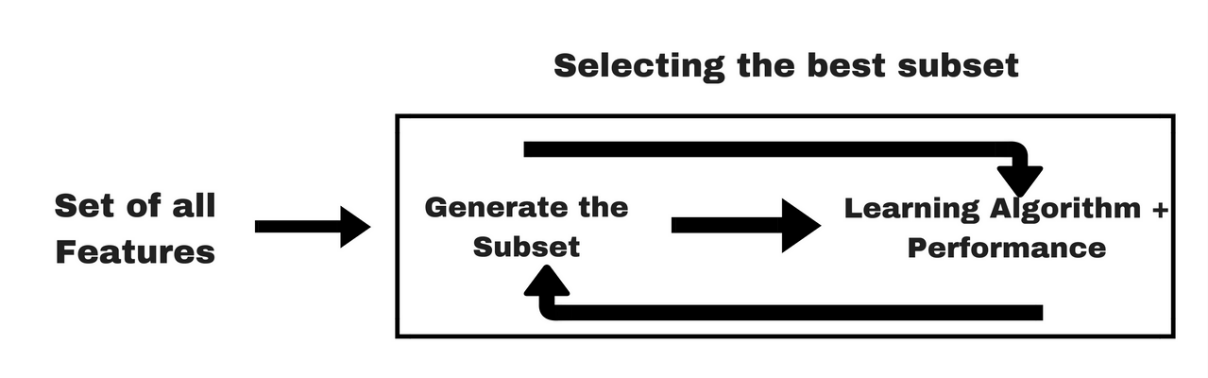
9. Make a few quick notes on:

PCA is an acronym for Personal Computer Analysis.

2. Use of vectors

Vectors are commonly used in machine learning as they lend a convenient way to organize data. Often one of the very first steps in making a machine learning model is vectorizing the data. They are also relied upon heavily to make up the basis for some machine learning techniques as well.

3. Embedded technique



10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection

In forward selection you start with your null model and add predictors. In backward selection you start with a full model including all your variables and then you drop those you do not need/ are not significant 1 at a time.

2. Function selection methods: filter vs. wrapper

Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it.

3. SMC vs. Jaccard coefficient

The SMC is very similar to the more popular Jaccard index. The main difference is that the SMC has the term in its numerator and denominator, whereas the Jaccard index does not. Thus, the SMC counts both mutual presences (when an attribute is present in both sets) and mutual absence (when an attribute is absent in both sets) as matches and compares it to the total number of attributes in the universe, whereas the Jaccard index only counts mutual presence as matches and compares it to the number of attributes that have been chosen by at least one of the two sets.