**ASSIGNMENT**

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

**engineering in depth.**

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

**methods of function selection?**

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

**approach?**

**4.**

**i. Describe the overall feature selection process.**

**ii. Explain the key underlying principle of feature extraction using an example. What are the most**

**widely used function extraction algorithms?**

**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.**

**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 &quot;high-dimensional data set&quot;? 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?**

**9. Make a few quick notes on:**

**PCA is an acronym for Personal Computer Analysis.**

**2. Use of vectors**

**3. Embedded technique**

**10. Make a comparison between:**

**1. Sequential backward exclusion vs. sequential forward selection**

**2. Function selection methods: filter vs. wrapper**

**3. SMC vs. Jaccard coefficient**

***SOLUTIONS***

1. ***Feature engineering refers to the process of selecting, extracting, and transforming features from raw data to improve machine learning model performance. The aspects of feature engineering include feature selection, feature extraction, feature transformation, and feature scaling.***
2. ***Feature selection is the process of selecting the most relevant features from a dataset to improve model performance and reduce the complexity of the model. The aim of feature selection is to identify the most informative features and remove irrelevant or redundant ones. The various methods of feature selection include filter methods, wrapper methods, and embedded methods.***
3. ***Filter methods and wrapper methods are two approaches to feature selection. Filter methods evaluate features based on statistical properties, while wrapper methods use a model to evaluate the usefulness of features. The pros of filter methods are that they are fast and computationally efficient, while the pros of wrapper methods are that they can capture interactions between features but are computationally expensive. The cons of filter methods are that they can overlook important features, while the cons of wrapper methods are that they can overfit to the training data.***
4. ***The overall feature selection process involves data collection and preprocessing, feature extraction, feature selection, model building and evaluation, and feature evaluation. The key underlying principle of feature extraction is to reduce the dimensionality of the dataset while retaining the most informative features. The most widely used feature extraction algorithms include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA).***
5. ***The feature engineering process in text categorization involves preprocessing the text data, such as removing stop words and stemming, and then transforming the text into numerical features using methods such as bag of words, term frequency-inverse document frequency (TF-IDF), and word embeddings.***
6. ***Cosine similarity is a good metric for text categorization because it measures the similarity between two vectors regardless of their magnitude. The resemblance in cosine between the two document-term matrix rows is 0.77.***

***i. The Hamming distance formula calculates the number of positions where two strings of equal length differ. The Hamming distance between two strings is obtained by counting the number of positions at which the corresponding symbols are different. For example, the Hamming distance between 10001011 and 11001111 is 3, because they differ in three positions (the second, fifth, and seventh).***

***ii. The Jaccard index and similarity matching coefficient are both measures of similarity between sets. The Jaccard index is the size of the intersection of the sets divided by the size of the union of the sets. In this case, the Jaccard index between the two feature sets is 4/6 or 0.67. The similarity matching coefficient is the number of matching pairs divided by the total number of pairs. In this case, the similarity matching coefficient between the two feature sets is 5/8 or 0.63.***

***A high-dimensional data set refers to a data set that has a large number of features or dimensions. Real-life examples of high-dimensional data sets include genomic data, image data, and text data. The difficulties in using machine learning techniques on a data set with many dimensions include the curse of dimensionality, which can lead to overfitting, sparsity, and computational complexity. To address these difficulties, techniques such as dimensionality reduction, feature selection, and regularization can be used.***

***PCA stands for Principal Component Analysis, which is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important information.***

***Vectors are mathematical objects that can represent quantities that have both magnitude and direction. They are widely used in machine learning for representing data, features, and parameters.***

***Embedded techniques refer to machine learning algorithms that perform feature selection and model building simultaneously, rather than treating them as separate steps.***

***10.***

1. ***Sequential backward exclusion and sequential forward selection are two feature selection techniques commonly used in machine learning. Sequential backward exclusion starts with all features and removes one feature at a time until the desired number of features is reached, while sequential forward selection starts with one feature and adds one feature at a time until the desired number of features is reached. The main difference between the two is the direction of the search. Sequential forward selection is generally faster and can be more effective in finding the best subset of features for small datasets, while sequential backward exclusion can be more effective in finding the best subset of features for large datasets.***
2. ***Function selection methods are used to select a subset of features from a larger set of features for use in a machine learning model. Filter methods use statistical techniques to select features that are highly correlated with the target variable and have a significant impact on the model's performance. Wrapper methods use a specific machine learning algorithm to train and test the model with different subsets of features and select the subset that results in the best performance. Wrapper methods are generally more accurate but can be computationally expensive, while filter methods are less accurate but computationally efficient.***
3. ***SMC (Simple Matching Coefficient) and Jaccard coefficient are two measures of similarity between two sets. SMC measures the proportion of matching items between two sets, while Jaccard coefficient measures the proportion of matching items relative to the total number of items in the sets. SMC is more appropriate when the sets are of equal size, while Jaccard coefficient is more appropriate when the sets are of different sizes. Jaccard coefficient is also more suitable for measuring similarity in cases where the items in the sets are binary or categorical, while SMC can be used for any type of data.***