**FSDS MAY BATCH 2022(ML Assignment -9)**

**Submitted by: Shubham Tiwari**

Q1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

Ans: Feature engineering is the process of transforming raw data into features that can be used in machine learning models. It involves selecting relevant features from the raw data, transforming them into a format that the model can understand, and creating new features through the combination or manipulation of existing ones.

There are several aspects of feature engineering, including:

1. Feature Selection: This involves identifying and selecting the most relevant features from the raw data. This can be done using techniques such as correlation analysis, mutual information, or feature importances from tree-based models.
2. Feature Extraction: This involves creating new features from the raw data by applying mathematical operations or transformations. Common techniques include principal component analysis (PCA), singular value decomposition (SVD), and feature scaling.
3. Feature Creation: This involves creating new features through the combination or manipulation of existing features. For example, creating a new feature that is the sum of two other features, or creating a categorical feature from a numerical one by binning the values.
4. Feature Encoding: This involves converting categorical variables into numerical ones so that they can be used in a machine learning model. Common encoding techniques include one-hot encoding, ordinal encoding, and binary encoding.
5. Feature Interaction: This involves creating new features by combining multiple features together. This can be done by taking the product, sum, or difference of two or more features.
6. Overall feature engineering is an iterative process and requires domain knowledge and experimentation to achieve the best results.

Q2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Ans: Feature selection is the process of identifying and selecting a subset of the most relevant features from a larger set of features. The goal of feature selection is to improve the performance and interpretability of machine learning models by reducing the dimensionality of the data, removing irrelevant or redundant features, and avoiding overfitting.

There are several methods of feature selection, including:

1. Filter Methods: These methods use statistical measures to evaluate the relevance of each feature independently and select the ones that have the highest scores. Common measures include correlation coefficient, chi-squared test, and mutual information.
2. Wrapper Methods: These methods use a machine learning model to evaluate the performance of different feature subsets and select the ones that result in the best performance. Common methods include forward selection, backward elimination, and recursive feature elimination.
3. Embedded Methods: These methods use the structure of the machine learning model to perform feature selection as part of the model training process. Examples include lasso regression and decision trees with feature importances.
4. Hybrid Methods: These methods combine two or more feature selection methods to take advantage of their strengths and overcome their weaknesses.

It's important to note that feature selection is an iterative process and is often done in conjunction with other feature engineering techniques such as feature extraction, feature creation, and feature scaling. Additionally, it's important to use appropriate feature selection method depending on the problem and data type.

Q3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach?

Ans: Filter methods and wrapper methods are two commonly used approaches for feature selection.

**Filter methods:**

* Filter methods evaluate the relevance of each feature independently, using statistical measures such as correlation coefficient, chi-squared test, or mutual information.
* These methods are computationally efficient and easy to implement, and they can be used with any type of machine learning model.
* Filter methods do not consider the interaction between features and the specific machine learning model being used, which may result in suboptimal feature subsets.

In addition, filter methods may not account for the specific characteristics of the data, such as class imbalance or non-linear relationships between features and the target variable.

**Wrapper methods:**

* Wrapper methods use a machine learning model to evaluate the performance of different feature subsets and select the ones that result in the best performance.
* These methods are more computationally intensive than filter methods, as they require training the model multiple times with different feature subsets.
* Wrapper methods are more specific to the machine learning model being used and can account for the interaction between features.
* However, they may be sensitive to the specific choice of the model and the initial feature subset, and they may also be prone to overfitting.

In summary, filter methods are computationally efficient and easy to implement, but they may not be as specific to the machine learning model or the data as wrapper methods. Wrapper methods are more specific to the machine learning model and the data, but they are more computationally intensive and may be prone to overfitting.

Q4.

i. Describe the overall feature selection process.

Ans: Feature selection is the process of identifying and selecting a subset of the most relevant features from a larger set of features. The goal of feature selection is to improve the performance and interpretability of machine learning models by reducing the dimensionality of the data, removing irrelevant or redundant features, and avoiding overfitting.

There are several methods of feature selection, including:

1. Filter Methods: These methods use statistical measures to evaluate the relevance of each feature independently and select the ones that have the highest scores. Common measures include correlation coefficient, chi-squared test, and mutual information.
2. Wrapper Methods: These methods use a machine learning model to evaluate the performance of different feature subsets and select the ones that result in the best performance. Common methods include forward selection, backward elimination, and recursive feature elimination.
3. Embedded Methods: These methods use the structure of the machine learning model to perform feature selection as part of the model training process. Examples include lasso regression and decision trees with feature importances.
4. Hybrid Methods: These methods combine two or more feature selection methods to take advantage of their strengths and overcome their weaknesses.

It's important to note that feature selection is an iterative process and is often done in conjunction with other feature engineering techniques such as feature extraction, feature creation, and feature scaling. Additionally, it's important to use appropriate feature selection method depending on the problem and data type.

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

Ans: The key underlying principle of feature extraction is to transform the high-dimensional data into a lower-dimensional space, while preserving as much of the information as possible. This is done by identifying the most relevant features for the problem at hand and discarding the rest.

For example, imagine we have a data set containing information on 1,000 different variables, such as age, income, education level, and so on, for a group of individuals. Through feature extraction, we can identify the most relevant variables that are most predictive of an individual's job performance, such as education level and years of experience. By extracting these features, we are able to reduce the dimensionality of the data and make it more manageable for machine learning algorithms.

The most widely used feature extraction algorithms include:

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Independent Component Analysis (ICA)
* Singular Value Decomposition (SVD)
* Non-negative Matrix Factorization (NMF)
* t-distributed Stochastic Neighbor Embedding (t-SNE)
* Autoencoder neural networks

PCA is one of the most widely used feature extraction algorithms, it is a linear technique that finds the most important principal components of the data, which are linear combinations of the original features. LDA is also a linear technique that finds the linear combinations of the features that maximally separate different classes. ICA is a technique that finds the independent components of the data, which are linear combinations of the original features that are as independent as possible. SVD is a technique that factorizes a matrix into three matrices, which can be used for dimensionality reduction. NMF is a technique that factorizes a non-negative matrix into two non-negative matrices, which can be used for dimensionality reduction. t-SNE is a non-linear technique that maps high-dimensional data to a low-dimensional space while preserving the structure of the data. Autoencoder neural networks are also used for feature extraction, where an encoder network reduces the dimensionality of the data, and a decoder network reconstructs the data.Top of Form

Q5. Describe the feature engineering process in the sense of a text categorization issue.

Ans: In text categorization, feature engineering is the process of extracting relevant features from raw text data to be used in a machine learning model for classification. This can include techniques such as tokenization, stemming, and stop word removal to preprocess the text, and then extracting features such as term frequency-inverse document frequency (TF-IDF) or bag of words representations to represent the text numerically. Once the features have been extracted, they can be used as input to a machine learning model, such as a support vector machine or a neural network, to train a classifier that can predict the category of new text instances.

Q6. 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.

Ans: Cosine similarity is a good metric for text categorization because it measures the similarity between two vectors by taking into account both the magnitude and direction of the vectors. In text classification, we are often comparing the similarity of two documents represented as vectors of term frequencies. The cosine similarity metric is commonly used because it is able to account for the fact that a document with high term frequencies may not necessarily be more similar to another document than a document with lower term frequencies, but similar distribution of term frequencies.

To calculate the cosine similarity of the two document-term matrix rows provided, we would first have to find the dot product of the two vectors, which is (22) + (31) + (20) + (00) + (23) + (32) + (31) + (03) + (11) = 16. Then, we would find the magnitude of each vector, which is square root of (22 + 33 + 22 + 00 + 22 + 33 + 33 + 00 + 11) = sqrt(48) = 6.83 and square root of (22 + 11 + 00 + 00 + 33 + 22 + 11 + 33 + 11) = sqrt(32) = 5.66. Finally, we can calculate the cosine similarity by dividing the dot product by the product of the magnitudes, which is (16)/(6.835.66) = 0.863

It's important to notice that this value ranges from -1 to 1, where 1 means the documents are exactly the same, 0 means they are completely different and -1 means they are exactly opposite.

Q7.

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

Ans: **The formula for calculating Hamming distance is: Hamming distance = the number of positions at which the corresponding bits are different.**

To calculate the Hamming distance between the two binary numbers 10001011 and 11001111, we compare the bits in each position and count the number of positions at which the bits are different.

10001011 11001111

In this example, there are 4 positions at which the corresponding bits are different (the first, third, fourth and seventh positions). So the Hamming distance between the two binary numbers is 4.

so Hamming gap in this example is 4.

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).

Ans: The Jaccard index and similarity matching coefficient are two different measures of similarity between two sets of data.

The Jaccard index is defined as the size of the intersection of two sets divided by the size of the union of the two sets.

Jaccard index = (size of intersection) / (size of union)

For the two sets of data (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), the size of the intersection is 6 and the size of the union is 7. So the Jaccard index for these two sets of data is 6/7.

The Similarity Matching Coefficient (SMC) measure the similarity between two binary data sets by the number of positions that are the same in the two data sets divided by the number of positions that are different.

SMC = (size of intersection) / (size of intersection + size of difference)

For the two sets of data (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), the size of the intersection is 6 and the size of the difference is 2. So the SMC for these two sets of data is 6/(6+2) = 6/8 = 3/4

So, Jaccard index is 6/7 and SMC is 3/4.

Top of Form

Q8. 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?

Ans: A high-dimensional data set refers to a data set with a large number of features or dimensions. For example, a data set containing information on 1,000 different variables would be considered high-dimensional. Real-life examples of high-dimensional data sets include image data (such as photographs or videos), text data (such as articles or social media posts), and genetic data (such as DNA sequences).

The difficulties in using machine learning techniques on high-dimensional data sets include the "curse of dimensionality," which refers to the fact that as the number of dimensions increases, the amount of data required to accurately model the data also increases exponentially. Additionally, high-dimensional data can make it more difficult to visualize and understand the relationships between variables.

To address these difficulties, various dimensionality reduction techniques can be used, such as principal component analysis (PCA) or linear discriminant analysis (LDA). These techniques transform the high-dimensional data into a lower-dimensional space while preserving as much of the information as possible. Other technique is feature selection which tries to identify the most relevant features for the problem at hand and discard the rest of the features.

Q9. Make a few quick notes on:

1.PCA is an acronym for Personal Computer Analysis.

Ans: No, PCA is an acronym for Principal Component Analysis, a statistical technique used for dimensionality reduction and data visualization. It is used in various fields such as image processing, signal processing, and data mining. It works by transforming the original data into a new set of uncorrelated variables called principal components. These principal components can then be used to represent the original data in a lower-dimensional space while retaining as much information as possible.

2. Use of vectors.

Ans: Vectors are mathematical objects that can be used to represent quantities with both magnitude and direction. They are widely used in various fields such as physics, engineering, computer graphics, and machine learning.

In physics, vectors are used to describe physical quantities such as velocity, acceleration, force, and momentum. They can be used to model the motion of objects, the behavior of fluids, and the properties of electric and magnetic fields.

In engineering, vectors are used to represent engineering quantities such as displacement, velocity, and acceleration. They can be used to model the behavior of mechanical systems, electrical systems, and control systems.

In computer graphics, vectors are used to represent images and graphics. They can be used to represent the position, size, and rotation of graphical objects, and can be used to perform transformations such as scaling, rotation, and translation.

In machine learning, vectors are used to represent data. They can be used to represent images, text, and audio, and can be used to perform machine learning tasks such as classification, clustering, and dimensionality reduction.

3. Embedded technique.

Ans: Embedded technique refers to a method of integrating one system into another, typically a computer system into a larger device or system. In embedded systems, the computer is integrated into the device or system and is specifically designed to perform a specific task or set of tasks. These tasks may include real-time control, data acquisition, and communication. Embedded systems can be found in a wide range of devices, from small devices such as smartphones and smartwatches to larger systems such as automobiles and industrial control systems. They are characterized by their ability to operate in harsh environments and their ability to interact with the physical world in real-time.

Q10. Make a comparison between:

1. Sequential backward exclusion vs. sequential forward selection.

Ans: Sequential backward elimination and sequential forward selection are both wrapper methods used for feature selection in machine learning.

**Sequential backward elimination (SBE) starts** with a full set of features and iteratively removes the feature that causes the smallest decrease in performance of the model. The process stops when no further improvement is observed or when a pre-determined number of features is reached. SBE is useful when the number of features is large and the goal is to reduce the feature set to a smaller size.

**Sequential forward selection (SFS), on the other hand, starts** with an empty set of features and iteratively adds the feature that causes the greatest increase in performance of the model. The process stops when no further improvement is observed or when a pre-determined number of features is reached. SFS is useful when the number of features is small or when the goal is to increase the feature set to a larger size.

In summary, SBE starts with a full set of features and iteratively removes features, while SFS starts with an empty set of features and iteratively adds features. SBE is useful when the number of features is large and the goal is to reduce the feature set to a smaller size, while SFS is useful when the number of features is small or when the goal is to increase the feature set to a larger size.Top of Form

2. Function selection methods: filter vs. wrapper.

Ans: Function selection methods are used to select a subset of features or variables from a larger set for use in a machine learning model. There are two main types of function selection methods: filter methods and wrapper methods.

**Filter methods** are based on the properties of the features themselves, such as their correlation with the target variable or their mutual information. These methods are computationally efficient and can be applied before building the model, but they do not take into account the performance of the model on the selected features. Examples of filter methods include chi-square test, mutual information, and correlation coefficient.

**Wrapper methods**, on the other hand, consider the performance of the model on the selected features. These methods use a specific machine learning algorithm as a "wrapper" to evaluate the feature subset. These methods are computationally more expensive and can be applied after building the model. Examples of wrapper methods include Recursive Feature Elimination (RFE), forward selection, and backward elimination.

In summary, filter methods are computationally efficient, but only consider the properties of the features themselves. Wrapper methods consider the performance of the model on the selected features but are computationally more expensive.

3. SMC vs. Jaccard coefficient.

Ans: The Jaccard coefficient and the Soft Jaccard coefficient (SMC) are both measures of similarity between two sets.

The Jaccard coefficient is defined as the size of the intersection of two sets divided by the size of the union of the two sets. It ranges from 0 (no similarity) to 1 (sets are identical).

SMC is a variation of Jaccard coefficient, which is defined as the size of the intersection of two sets divided by the sum of the sizes of the two sets minus the size of the intersection. It ranges from -1 (sets are completely dissimilar) to 1 (sets are identical).

In summary, both SMC and Jaccard coefficients are measures of similarity of sets, but Jaccard coefficient range is only between 0 to 1, while SMC coefficient range is between -1 to 1.