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

Answer: Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data. The features in your data will directly influence the predictive models you use and the results you can achieve. Domain knowledge of data is key to the process. Along with domain knowledge, both programming and math skills are required to perform feature engineering. The correct features make machine learning algorithms successful.

Features influence the results of predictive models:

1. Reduced complexity
2. Increased accuracy

Three general types of features:

1. Categorical features
2. Text features
3. Image features

5 Feature Engineering Techniques:

1. Imputation
2. Outliers
3. Log Transform
4. Binning
5. Feature Split

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

Answer: Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

As such, it can be challenging for a machine learning practitioner to select an appropriate statistical measure for a dataset when performing filter-based feature selection.

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. Feature selection is primarily focused on removing non-informative or redundant predictors from the model.

Some predictive modeling problems have a large number of variables that can slow the development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

Feature selection as follows:

* Feature Selection: Select a subset of input features from the dataset.
  + Unsupervised: Do not use the target variable (e.g. remove redundant variables).
    - Correlation
  + Supervised: Use the target variable (e.g. remove irrelevant variables).
    - Wrapper: Search for well-performing subsets of features.
      * RFE
    - Filter: Select subsets of features based on their relationship with the target.
      * Statistical Methods
      * Feature Importance Methods
    - Intrinsic: Algorithms that perform automatic feature selection during training.
      * Decision Trees
* Dimensionality Reduction: Project input data into a lower-dimensional feature space.

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

Answer:

Filter methods are generally used as a pre-processing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here. For basic guidance, you can refer to the following table for defining correlation co-efficient.

|  |  |  |
| --- | --- | --- |
| Feature | Continuous | Categorical |
| Continuous | Pearson’s Correlation | LDA |
| Categorical | Anova | Chi-Square |

* Pearson’s Correlation: It is used as a measure for quantifying linear dependence between two continuous variables X and Y. Its value varies from -1 to +1.
* LDA: Linear discriminant analysis is used to find a linear combination of features that characterizes or separates two or more classes (or levels) of a categorical variable.
* ANOVA: ANOVA stands for Analysis of variance. It is similar to LDA except for the fact that it is operated using one or more categorical independent features and one continuous dependent feature. It provides a statistical test of whether the means of several groups are equal or not.
* Chi-Square: It is a is a statistical test applied to the groups of categorical features to evaluate the likelihood of correlation or association between them using their frequency distribution.

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

* Forward Selection: Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
* Backward Elimination: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.
* Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

|  |  |  |
| --- | --- | --- |
| Feature Selection Method | Merits | Demerits |
| Filter | Computationally cheaper as compared to wrapper and embedded. Fastest running time. Lower risk of over fitting. Ability of good generalization. Easily scale to high-dimensional datasets. | No interaction with classification model for feature selection. Mostly ignores feature dependencies and considers each feature separately in case of univariate techniques, which may lead to low computational performance as compared to other techniques of feature selection. |
| Wrapper | Interacts with the classifier for feature selection. More comprehensive search of feature-set space. Considers feature dependencies. Better generalization than filter approach | High computational cost. Longer running time. Higher risk of over fitting as compared to filter and embedded. More computationally unfeasible with increased number of features No guarantee of optimality of the solution if predicted with another classifier. |

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?

Answer:

1. Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

As such, it can be challenging for a machine learning practitioner to select an appropriate statistical measure for a dataset when performing filter-based feature selection.

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. Feature selection is primarily focused on removing non-informative or redundant predictors from the model.

Some predictive modeling problems have a large number of variables that can slow the development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

Feature selection as follows:

* Feature Selection: Select a subset of input features from the dataset.
  + Unsupervised: Do not use the target variable (e.g. remove redundant variables).
    - Correlation
  + Supervised: Use the target variable (e.g. remove irrelevant variables).
    - Wrapper: Search for well-performing subsets of features.
      * RFE
    - Filter: Select subsets of features based on their relationship with the target.
      * Statistical Methods
      * Feature Importance Methods
    - Intrinsic: Algorithms that perform automatic feature selection during training.
      * Decision Trees
* Dimensionality Reduction: Project input data into a lower-dimensional feature space.

1. Feature Extraction: By finding a smaller set of new variables, each being a combination of the input variables, containing basically the same information as the input variables.

i. PCA (Principal Component Analysis)

ii. LDA (Linear Discriminant Analysis)

The feature Extraction technique gives us new features which are a linear combination of the existing features. The new set of features will have different values as compared to the original feature values. The main aim is that fewer features will be required to capture the same information. We might think that choosing fewer features might lead to underfitting but in the case of the Feature Extraction technique, the extra data is generally noise.

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

Answer: Text classification is the problem of assigning categories to text data according to its content. The most important part of text classification is feature engineering: the process of creating features for a machine learning model from raw text data. Text classification is the task of automatically placing

pre-defined labels on previously unseen documents. Used in document indexing, e-mail filtering, web browsing, and personal information agents, text classification is an active and important area of research where machine learning and information retrieval (IR) intersect.

Steps:

1. Environment setup: import packages and read data.
2. Language detection: understand which natural language data is in.
3. Text pre-processing: text cleaning and transformation.
4. Length analysis: measured with different metrics.
5. Sentiment analysis: determine whether a text is positive or negative.
6. Named-Entity recognition: tag text with pre-defined categories such as person names, organizations, locations.
7. Word frequency: find the most important n-grams.
8. Word vectors: transform a word into numbers.
9. Topic modeling: extract the main topics from corpus.

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.

Answer:

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis.

x= (2, 3, 2, 0, 2, 3, 3, 0, 1)

y= (2, 1, 0, 0, 3, 2, 1, 3, 1)

x . y = 2\*2 + 3\*1 + 2\*0 + 0\*0 + 2\*3 + 3\*2 + 3\*1 + 0\*3 + 1\*1 = 23

||x|| = √ (2)^2 + (3)^2 + (2)^2 + (0)^2 + (2)^2+ (3)^2+ (3)^2+ (0)^2+ (1)^2 = √ 40

||y|| = √ (2)^2 + (1)^2 + (0)^2 + (0)^2 + (3)^2+ (2)^2+ (1)^2+ (3)^2+ (1)^2 = √ 29

The formula for calculating the cosine similarity is : Cos(x, y) = x . y / ||x|| \* ||y||

Cos(x, y) = 23 / 6.3245 \* 5.385 = 23 / 34.05 = 0.675

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).

Answer:

1. Hamming distance between two vectors is the number of bits we must change to change

one into the other. d=min {d(x,y):x,y∈C,x≠y}. The Hamming distance of two given lines of code is the number of points at which the lines' binary code values are different (assuming that the two lines of code are the same length).

Hamming gap between 10001011 and 11001111 is 4.

1. Simple matching coefficient 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.

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?

Answer:

High Dimensional means that the number of dimensions are staggeringly high — so high that calculations become extremely difficult. With high dimensional data, the number of features can exceed the number of observations

Dimensionally cursed phenomena occur in domains such as numerical analysis, sampling, combinatorics, machine learning, data mining and databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse.

High dimension is when variable numbers p is higher than the sample sizes n i.e. p>n, cases. For example, tomographic imaging data, ECG data, and MEG data. One example of high dimensional data is microarray gene expression data.

Seven Techniques for Data Dimensionality Reduction

1. Missing Values Ratio.
2. Low Variance Filter.
3. High Correlation Filter.
4. Random Forests / Ensemble Trees.
5. Principal Component Analysis (PCA).
6. Backward Feature Elimination.
7. Forward Feature Construction.

9. Make a few quick notes on:

1. PCA is an acronym for Personal Computer Analysis.

2. Use of vectors

3. Embedded technique

Answer:

1. Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
2. Vectors can be used to represent physical quantities. Most commonly in physics, vectors are used to represent displacement, velocity, and acceleration. Vectors are a combination of magnitude and direction and are drawn as arrows.
3. Embedded methods complete the feature selection process within the construction of the machine learning algorithm itself. In other words, they perform feature selection during the model training, which is why we call them embedded methods.

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

Answer:

1. Sequential Forward Selection (SFS)

Start with empty set, X=0

Repeatedly add most significant feature with respect to X

Disadvantage: Once a feature is retained, it cannot be discarded; nesting problem

Sequential Backward Selection (SBS)

Start with full set, X=Y

Repeatedly delete least significant feature in X

Disadvantage: SBS requires more computation than SFS; Nesting problem

1. On one hand, filter methods use an evaluation function that relies solely on properties of the data, thus is independent on any particular algorithm. On the other hand, wrapper methods use the inductive algorithm to estimate the value of a given subset.
2. Simple matching coefficient 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.