Assignment 09 Solutions

1. 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 selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

Feature engineering in ML consists of four main steps: Feature Creation, Transformations, Feature Extraction, and Feature Selection. Feature engineering consists of creation, transformation, extraction, and selection of features, also known as variables, that are most conducive to creating an accurate ML algorithm.

Lets take a look at some Feature Engineering Types:

1. Imputation: Imputation deals with handling missing values in data.
2. Discretization/Binning: Discretization involves essentially taking a set of values of data and grouping sets of them together in some logical fashion into bins (or buckets).
3. Caategorical Encoding: Categorical encoding is the technique used to encode categorical features into numerical values which are usually simpler for an algorithm to understand. One hot encoding(OHE) is a popularly used technique of categorical encoding. Here, categorical values are converted into simple numerical 1’s and 0’s without the loss of information.
4. Feature Splitting: Splitting features into parts can sometimes improve the value of the features toward the target to be learned. For instance, Date might better contribute to the target function than Date and Time.
5. Handling Outliers: Outliers are unusually high or low values in the dataset which are unlikely to occur in normal scenarios. Since these outliers could adversely affect your prediction they must be handled appropriately. The various methods of handling outliers include: a) Removal: The records containing outliers are removed from the distribution. However, the presence of outliers over multiple variables could result in losing out on a large portion of the datasheet with this method. b) Replacing values: The outliers could alternatively bed treated as missing values and replaced by using appropriate imputation. c) Capping: Capping the maximum and minimum values and replacing them with an arbitrary value or a value from a variable distribution. d) Discretization
6. Scaling: Feature scaling is done owing to the sensitivity of some machine learning algorithms to the scale of the input values. This technique of feature scaling is sometimes referred to as feature normalization. The commonly used processes of scaling include: a) Min-Max Scaling: This process involves the rescaling of all values in a feature in the range 0 to 1. In other words, the minimum value in the original range will take the value 0, the maximum value will take 1 and the rest of the values in between the two extremes will be appropriately scaled. b) Standardization/Variance scaling: All the data points are subtracted by their mean and the result divided by the distribution's variance to arrive at a distribution with a 0 mean and variance of 1.
7. Variable Transformation: Variable transformation techniques could help with normalizing skewed data. One such popularly used transformation is the logarithmic transformation. Logarithmic transformations operate to compress the larger numbers and relatively expand the smaller numbers. This in turn results in less skewed values especially in the case of heavy-tailed distributions. Other variable transformations used include Square root transformation and Box cox transformation which is a generalization of the former two.

8. Feature Creation: Feature creation involves deriving new features from existing ones. This can be done by simple mathematical operations such as aggregations to obtain the mean, median, mode, sum, or difference and even product of two values. These features, although derived directly from the given

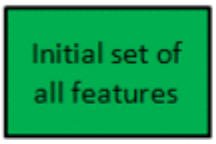
2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

Ans: Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

There are three types of feature selection:

- **Wrapper methods** (forward, backward, and stepwise selection)
- **Filter methods** (ANOVA, Pearson correlation, variance thresholding)
- **Embedded methods** (Lasso, Ridge, Decision Tree).

"



\""\n",

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

Ans: The main differences between the filter and wrapper methods for feature selection are:

1. 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.
2. Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
3. Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
4. Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.
5. Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

4. Please Answer the following Questions :

1. Describe the overall feature selection process.
2. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?

Ans: 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. There are three types of feature selection:

- **Wrapper methods** (forward, backward, and stepwise selection): 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.
- **Filter methods** (ANOVA, Pearson correlation, variance thresholding): Filter methods are generally used as a preprocessing 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.
- **Embedded methods** (Lasso, Ridge, Decision Tree): Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods. Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting. Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients. Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

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

Ans: 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. Example: Text cleaning steps vary according to the type of data and the required task. Generally, the string is converted to lowercase and punctuation is removed before text gets tokenized. Tokenization is the process of splitting a string into a list of strings (or "tokens"). We can create a list of generic stop words for the English vocabulary with NLTK (the Natural Language Toolkit), which is a suite of libraries and programs for symbolic and statistical natural language processing. Then we can remove these stop words. We need to be very careful with stop words because if you remove the wrong token you may lose important information. For example, the word "will" was removed and we lost the information that the person is Will Smith. With this in mind, it can be useful to do some manual modification to the raw text before removing stop words (for example, replacing "Will Smith" with "Will_Smith"). Stemming and Lemmatization both generate the root form of words. The difference is that stem might not be an actual word whereas lemma is an actual language word (also stemming is usually faster). Those algorithms are both provided by NLTK.

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.

Ans: Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together.

Cosine similarity is the cosine of the angle between two n-dimensional vectors in an n-dimensional space. It is the dot product of the two vectors divided by the product of the two vectors' lengths (or magnitudes).

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

In our Question cos(x,y) = 23/(root 40 * root 29) = 0.675

7. Explain the following:

1. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.
2. 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:

1. The Hamming distance between two vectors is the number of bits we must change to change one into the other. Example Find the distance between the vectors `01101010` and `11011011`. They differ in four places, so the Hamming distance `d(01101010,11011011) = 4`. In question mentioned between 10001011 and 11001111, hamming distance will be 2 as two character are different.
2. Jaccard Index = (the number in both sets) / (the number in either set) * 100 For Question given, Jaccard Index = 2/2 *100 = 100%

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?

Ans: High dimension is when variable numbers p is higher than the sample sizes n i.e. p>n, cases. High dimensional data is referred to a data of n samples with p features, where p is larger than n.

For example, tomographic imaging data, ECG data, and MEG data. One example of high dimensional data is microarray gene expression data.

9. Make a few quick notes on:

1. PCA is an acronym for Personal Computer Analysis.
2. Use of vectors
3. Embedded technique

Ans: The Principal component analysis (PCA) is a technique used for identification of a smaller number of uncorrelated variables known as principal components from a larger set of data. The technique is widely used to emphasize variation and capture strong patterns in a data set.

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

In the context of machine learning, an embedding is a low-dimensional, learned continuous vector representation of discrete variables into which you can translate high-dimensional vectors. Generally, embeddings make ML models more efficient and easier to work with, and can be used with other models as well

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

Ans: Sequential floating forward selection (SFFS) starts from the empty set. After each forward step, SFFS performs backward steps as long as the objective function increases. Sequential floating backward selection (SFBS) starts from the full set.

The main differences between the filter and wrapper methods for feature selection are:

1. 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.
2. Filter methods are much faster compared to wrapper methods as they do not involve training the models. On the other hand, wrapper methods are computationally very expensive as well.
3. Filter methods use statistical methods for evaluation of a subset of features while wrapper methods use cross validation.
4. Filter methods might fail to find the best subset of features in many occasions but wrapper methods can always provide the best subset of features.
5. Using the subset of features from the wrapper methods make the model more prone to overfitting as compared to using subset of features from the filter methods.

The Jaccard coefficient is a measure of the percentage of overlap between sets defined as: (5.1) where W1 and W2 are two sets, in our case the 1-year windows of the ego networks. The Jaccard coefficient can be a value between 0 and 1, with 0 indicating no overlap and 1 complete overlap between the sets.

An Example is given below:

"

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0