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.

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).

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: 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.

The filter method has the fastest running time; however, it does not consider feature dependencies and tends to each feature separately when univariate techniques are used. The wrapper method has the advantages of better generalization and robust interaction with the classifier used for feature selection

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.

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.

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).

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: Thus 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`.

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 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.