Week 3 Data-Exploration

1. What is meant by data exploration and why is it important?

- Data exploration is concerned with understanding our dataset, to gain deep insights into its content and suitability for using in a machine learning application.
- The instances can come from a variety of sources such as electronic sensors, human observations, survey results, data logs and so on.

2. What is the process of data cleaning and preparation? When and why do we carry out this activity?

- Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

- After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

3. What does covariance measure and why do we use it? What is an alternative measure and why might we use that instead?

- We ues covariance is to complement the plot visualisations of potential feature correlations of continuous variables
- Covariance is a function of the sample means of the variables in question and is defined as the weighted sum of the differences between the observed values for the feature pair and their respective population mean.
- The alternative measure is a normalized measure such as correlation.
- If the features being compared are measured in the different units.

4. What is a scatter plot matrix and what does it tell you about the features in a dataset?

- a scatter plot matrix is the total set of pairwise plots of our dataset presented as one graph.
- If we see any such suspected correlations, this would warrant further investigate and analysis of the impacted features such as calculating formal measures of this relationship.

5. What is the problem with including correlated features in a dataset when training a model?

- Strongly correlated features can impact on many algorithms' performance. Having two features contributing to the same information and learning process is both redundant and potentially error prone, especially if they are negatively correlated.
- Another reason to eliminate redundancy in our data is that feature space reduction will allow our learning algorithms to train more efficiently by converging faster and using fewer computing resources.

---

6. What is meant by dimensionality reduction?

- In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of **reducing the number** of random variables under consideration by **obtaining a set of principal variables**. Approaches can be divided into feature selection and feature extraction.

---

7. Why might it be recommended to normalize a continuous feature?

- Normalization is an way of transforming a feature value range into another range with linear scaling preserved.
- The standard score of a variable range measures how many standard deviations from the mean a variable value is. This squeezes the feature distribution into a scaled distribution having a mean of zero and a standard deviation of one.

---

8. What is the purpose of binning? Explain how it works with an example

- Binning is one way to solve the problem of a large cardinality for a feature may not be useful for some learning approaches, for example decision trees.
- There are two kinds of binning commonly used. These are **equal-width binning** and **equal-frequency binning**.
- Binning requires that we specify the number of bins for our continuous feature values. Too low a number of bins results in a loss of resolution of our feature values. Too high a number of bins may result in some empty bins.
- **Equal-width binning** means dividing the feature range onto k, equally sized intervals and categorizing any variable values falling in that interval.
- **Equal-frequency binning** means dividing the feature value range having into k intervals having approximately the same number of data points.