Assignment 08 Solutions

1. What exactly is a feature? Give an example to illustrate your point ?

Ans: Features are the basic building blocks of datasets. The quality of the features in your dataset has a major impact on the quality of the insights you will gain when you use that dataset for machine learning.

Additionally, different business problems within the same industry do not necessarily require the same features, which is why it is important to have a strong understanding of the business goals of your data science project.

In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon. Choosing informative, discriminating and independent features is a crucial element of effective algorithms in pattern recognition, classification and regression.

2. What are the various circumstances in which feature construction is required ?

Ans: The features in your data will directly influence the predictive models you use and the results you can achieve. Our results are dependent on many inter-dependent properties. We need great features that describe the structures inherent in your data. Better features means flexibility. The process of generating new variables (features) based on already existing variables is known as feature construction.

Feature Construction is a useful process as it can add more information and give more insights of the data we are dealing with. It is done by transforming the numerical features into categorical features which is done while performing Binning. Also, feature construction is done by decomposing variables so that these new variables can be used in various machine learning algorithms such as the creation of Dummy Variables by performing Encoding. Other ways of constructing include deriving features from the pre-existing features and coming up with more meaningful features.

3. Describe how nominal variables are encoded ?

Ans: Nominal data is made of discrete values with no numerical relationship between the different categories — mean and median are meaningless. Animal species is one example. For example, pig is not higher than bird and lower than fish. Ordinal or Label Encoding can be used to transform non-numerical labels into numerical labels (or nominal categorical variables). Numerical labels are always between 1 and the number of classes. The labels chosen for the categories have no relationship. So categories that have some ties or are close to each other lose such information after encoding. The first unique value in your column becomes 1, the second becomes 2, the third becomes 3, and so on.

4. Describe how numeric features are converted to categorical features ?

Ans: Numeric Features can be converted to Categorical Features using Binning. Discretization: It is the process of transforming continuous variables into categorical variables by creating a set of intervals, which are contiguous, that span over the range of the variable’s values. It is also known as “Binning”, where the

bin is an analogous name for an interval.

Benefits of this method are:

1. Handles the Outliers in a better way.
2. Improves the value spread.
3. Minimize the effects of small observation errors.

ables.

Techniques to Encode Numerical Columns:

(a) Equal width binning: It is also known as “Uniform Binning” since the width of all the intervals is the same. The algorithm divides the data into N intervals of equal size. The width of intervals is:

w=(max-min)/N

Therefore, the interval boundaries are:[min+w], [min+2w], [min+3w],..., [min+(N-1)w] where, min and max are the minimum and maximum value from the data respectively. This technique does not changes the spread of the data but does handle the outliers.

(b) Equal frequency binning: It is also known as “Quantile Binning”. The algorithm divides the data into N groups where each group contains approximately the same number of values.

Consider, we want 10 bins, that is each interval contains 10% of the total observations. Here the width of the interval need not necessarily be equal. Handles outliers better than the previous method and makes the value spread approximately uniform(each interval contains almost the same number of values).

(c) K-means binning: This technique uses the clustering algorithm namely ” K-Means Algorithm”. This technique is mostly used when our data is in the form of clusters

5. Describe the feature selection wrapper approach. State the advantages and disadvantages of this approach ?

Ans: Wrapper methods measure the “usefulness” of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the “relevance” of the features) measured via univariate statistics instead of cross-validation performance.

The wrapper classification algorithms with joint dimensionality reduction and classification can also be used but these methods have high computation cost, lower discriminative power. Moreover, these methods depend on the efficient selection of classifiers for obtaining high accuracy.

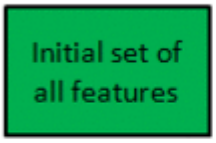
Most commonly used techniques under wrapper methods are:

1.**Forward selection**: In forward selection, we start with a null model and then start fitting the model with each individual feature one at a time and select the feature with the minimum p-value. Now fit a model with two features by trying combinations of the earlier selected feature with all other remaining features. Again select the feature with the minimum p-value. Now fit a model with three features by trying combinations of two previously selected features with other remaining features. Repeat this process until we have a set of selected features with a p-value of individual features less than the significance level.

2.**Backward elimination**: In backward elimination, we start with the full model (including all the independent variables) and then remove the insignificant feature with the highest p-value(> significance level). This process repeats again and again until we have the final set of significant features

3.**Bi-directional elimination(Stepwise Selection):** It is similar to forward selection but the difference is while adding a new feature it also checks the significance of already added features and if it finds any of the already selected features insignificant then it simply removes that particular feature through backward elimination. Hence, It is a combination of forward selection and backward elimination.

""



6. When is a feature considered irrelevant? What can be said to quantify it ?

Ans: Features are considered relevant if they are either strongly or weakly relevant, and are considered irrelevant otherwise.

Irrelevant features can never contribute to prediction accuracy, by definition. Also to quantify it we need to first check the list of 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).

p-value greater than 0.05 means that the feature is insignificant.

7. When is a function considered redundant? What criteria are used to identify features that could be redundant ?

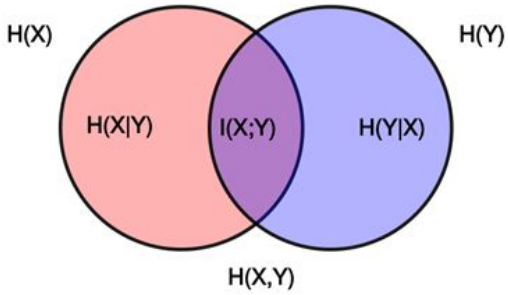
Ans: If two features `{X1, X2}` are highly correlated, then the two features become redundant features since they have same information in terms of correlation measure. In other words, the correlation measure provides statistical association between any given a pair of features.

Minimum redundancy feature selection is an algorithm frequently used in a method to accurately identify characteristics of genes and phenotypes

""

Background

- Relevance between features
  - Correlation
  - F-statistic
  - Mutual information



$$I(x,y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy$$

8. What are the various distance measurements used to determine feature similarity ?

Ans: Four of the most commonly used distance measures in machine learning are as follows:

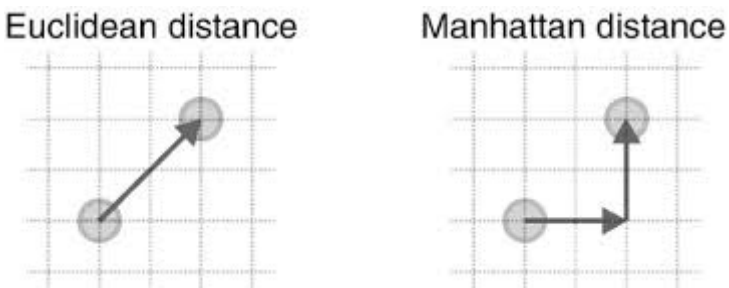
- Hamming Distance: Hamming distance calculates the distance between two binary vectors, also referred to as binary strings or bitstrings for short.
- Euclidean Distance: Calculates the distance between two real-valued vectors.
- Manhattan Distance: Also called the Taxicab distance or the City Block distance, calculates the distance between two real-valued vectors.
- Minkowski Distance: Minkowski distance calculates the distance between two real-valued vectors. It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the “order” or “p“, that allows different distance measures to be calculated.

9. State difference between Euclidean and Manhattan distances ?

Ans: Euclidean & Hamming distances are used to measure similarity or dissimilarity between two sequences. Euclidean distance is extensively applied in analysis of convolutional codes and Trellis codes.

Euclidean distance is the shortest path between source and destination which is a straight line as shown in Figure 1.3. but Manhattan distance is sum of all the real distances between source(s) and destination(d) and each distance are always the straight lines

""



""

10. Distinguish between feature transformation and feature selection ?

Ans: Feature selection is for filtering irrelevant or redundant features from your dataset. The key difference between feature selection and extraction is that feature selection keeps a subset of the original features while feature extraction creates brand new ones.

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.

Feature Transformation is a technique by which we can boost our model performance. Feature transformation is a mathematical transformation in which we apply a mathematical formula to a particular column(feature) and transform the values which are useful for our further analysis. It is also known as Feature Engineering, which is creating new features from existing features that may help in improving the model performance. It refers to the family of algorithms that create new features using the existing features. These new features may not have the same interpretation as the original features, but they may have more explanatory power in a different space rather than in the original space. This can also be used for Feature Reduction. It can be done in many ways, by linear combinations of original features or by using non-linear functions. It helps machine learning algorithms to converge faster.