1 Dimensionality Reduction

To understand Dimensionality Reduction, First we should understand Curse Of Dimensionality. It refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

In machine learning and statistics, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables.

1.1 Why Dimensionality Reduction is important?

Nowadays, Data comes in all forms video, audio, images, texts etc., with huge number of features. Is it that all features are relevant ?, NO, not all feature are important or relevant. Based on business requirement or redundancy nature of the data captured we have to reduce the feature size through Feature selection and Feature Extraction. These techniques not only reduce computation cost but it also helps in avoiding the misclassification because of highly correlated variable.

To overcome the above problem, we do dimensionality reduction. There are number of ways of Dimensionality reduction such as feature selection and Feature Extraction.

2 Feature Selection

Feature Selection is a very critical component in a Data Scientist's workflow. When presented data with very high dimensionality, models usually choke because

1. *Training time* increases exponentially with number of features.
2. Models have increasing risk of *overfitting* with increasing number of features.

Feature Selection methods helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.

There are three feature selection techniques. They are:

1. Filter Methods
2. Wrapper Methods and
3. Embedded Methods

2.1 Filter Methods

Filter Methods considers the relationship between features and the target variable to compute the importance of features.

F Test

F Test is a statistical test used to compare between models and check if the difference is significant between the model.

F-Test does a hypothesis testing model X and Y where X is a model created by just a constant and Y is the model created by a constant and a feature.

The least square errors in both the models are compared and checks if the difference in errors between model X and Y are significant or introduced by chance.

Mutual Information

Mutual Information between two variables measures the dependence of one variable to another. If X and Y are two variables, and

1. If X and Y are independent, then no information about Y can be obtained by knowing X or vice versa. Hence their mutual information is 0.

2. If X is a deterministic function of Y, then we can determine X from Y and Y from X with mutual information 1.

3. When we have Y = f(X,Z,M,N), $0 <$ mutual information < 1

We can select our features from feature space by ranking their mutual information with the target variable.

Advantage of using mutual information over F-Test is, it does well with the non-linear relationship between feature and target variable.

Variance Threshold

This method removes features with variation below a certain cutoff.

The idea is when a feature doesn't vary much within itself, it generally has very little predictive power.

2.2 Wrapper Methods

Wrapper Methods generate models with a subsets of feature and gauge their model performances.

Forward Search

This method allows you to search for the best feature w.r.t model performance and add them to your feature subset one after the other.

For data with n features,

1. On first round 'n' models are created with individual feature and the best predictive feature is selected.

2. On second round, 'n-1' models are created with each feature and the previously selected feature.

3. This is repeated till a best subset of 'm' features are selected.

Recursive Feature Elimination

As the name suggests, this method eliminates worst performing features on a particular model one after the other until the best subset of features are known.

For data with n features,

1. On first round 'n-1' models are created with combination of all features except one. The least performing feature is removed.

2. On second round 'n-2' models are created by removing another feature.

Wrapper Methods promises you a best set of features with a extensive greedy search.

2.3 Embedded Methods

Feature selection can also be acheived by the insights provided by some Machine Learning models.

LASSO Linear Regression can be used for feature selections. Lasso Regression is performed by adding an extra term to the cost function of Linear Regression. This apart from preventing overfitting also reduces the coefficients of less important features to zero.

Tree based models calculates feature importance for they need to keep the best performing features as close to the root of the tree. Constructing a decision tree involves calculating the best predictive feature.

3 Feature Extraction

It is a transformation of raw data into features suitable for modeling. Say we have ten independent variables, in feature extraction, we create ten "new" independent variables, where each "new" independent variable is a combination of each of the ten "old" independent variables. However, we create these new independent variables in a specific way and order these new variables by how well they predict our dependent variable.

You might say, "Where does the dimensionality reduction come into play?" Well, we keep as many of the new independent variables as we want, but we drop the "least important ones". Because we ordered the new variables by how well they predict our dependent variable, we know which variable is the most important and least important because, these new independent variables are combinations of our old ones, we're still keeping the most valuable parts of our old variables, even when we drop one or more of these "new" variables

Principal component analysis is a technique for feature extraction-so it combines our input variables in a specific way, then we can drop the "least important" variables while still retaining the most valuable parts of all of the variables! As an added benefit, each of the "new" variables after PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another. we will discuss this topic in further lecture in detail.