Machine Learning Overview

Lucky Mahlangu, Mpinane Mohale, Thulisile Shipyana, Sbusiso Mkhombe

23 August 2018

Gathering and Preparing Data

Data Collection and Preparation

- Since the project is still in its preliminary phase, artifitial data will be generated to keep the project going, however as the project progresses it is worth noting that real data must be used to ensure that the trained machine learning classifier is properly validated.
- It is convenient to work with data that is artificially generated, it saves us from cleaning the data and performing all the data mangling techniques as well as slicing and dicing features.
- Upon generating the datasets, we can take care of the following cases:
  - Ensure that there are no duplicates in the CSV files.
  - Ensure that all the rows have data associated with the corresponding features(*There are no missing values*)
- One primitive hoslistic approach that will be used to generate the datasets is *string manipulation*, *string manipulation* involves processing or transforming string objects into a desired format or structure. One way of achieving this, is using regular expressions.
- To increase cohesion between features, we may consider to not include outliers in the generated datasets, on the other hand we must impose some constraints on the values that are taken by the features.
- To give some sense into our data , we must consider using probability distributions, and focus widely on distributed values.
- New datasets can be derived by merging datasets that are already created some primitive functions that can be used to merge datasets include:
  - Concatenation
  - Join
  - Merge

Data Analysis and Model Selection

- Before choosing any classifier to use as a predictor, we need to carefully analyze the generated datasets and visually represent the data using various graphical methods. We need to carefully inspect each feature.
- As part of our preliminary techniques, we will start brute forcing the problem by using the following machine learning techniques:
  - Logistic Regression
  - K-Means Clustering
- *Logistic Regression*-Logistic Regression is used when the dependent variable(target) is categorical.
- *K-Means Clustering*-K-means clustering aims to partition n observations into k clusters.