First-order Theorem Proving
Which Features Impact the Most ?

Abhinandan Dubey
Stony Brook University

Abstract

The aim of this project was to find out simple feature measurements of a conjecture and axioms that sufficiently provide information to determine a good choice of heuristic. Our dataset consists a set of 5 heuristics, each having results from 14 static feature measurements and 39 dynamic feature measurements. The basic units affecting the features include the set of processed clauses, the set of unprocessed clauses and the axioms. We use the D3.js library, d3-v3.js for visualization and Python for the backend.

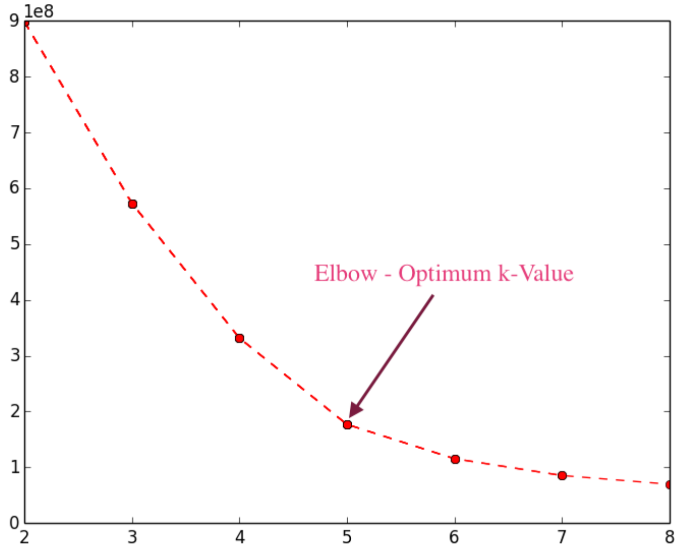
Introduction

The task is to perform Principal Component Analysis, over data sampled using random and stratified sampling, and find the number of principal components. We also perform k-means clustering with an elbow method to find an optimum value of k, and use it to find out similar cases. We also perform Multi-dimensional scaling to visualize the similarity between individual cases in the dataset. The first-order theorem proving dataset is used to analyze the top factors which contribute in time taken to prove a theorem using predicate calculus. The dataset has been taken from the UCI Machine Learning Repository.

We also present a dashboard which briefly summarizes the results from our analyses. The dashboard includes a title indicates the original research paper and attributes the names. The main area holds a section menu for the user to choose which part to visualize. A link to feature set description has also been provided for further clarity.

Analysis

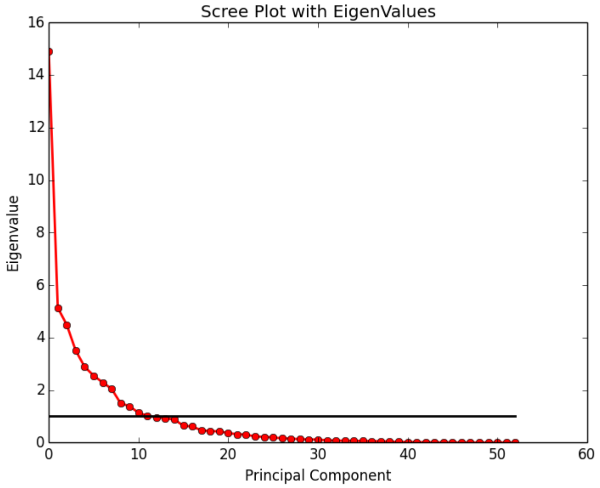
Our dataset has around 6118 problems. We used both kinds of sampling to compare the results (random and stratified.). For stratified sampling, we performed clustering using K-Means Clustering algorithm, and then randomly picked 20% of cases from each cluster. We performed optimization using elbow method for finding an optimum value of k.



The figure above shows the Elbow Plot generated from the optimize_elbow method above.

Finding the intrinsic dimensionality of the data using PCA

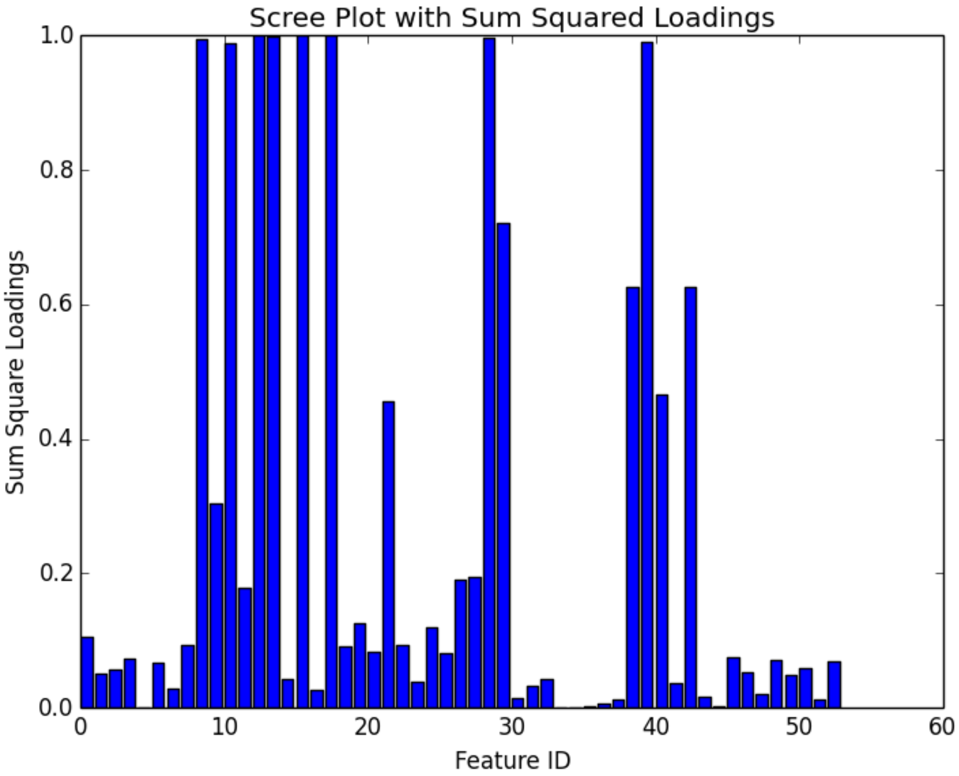
In our case the intrinsic dimensionality of the data comes out to be 10. There are a total of 53 features. Figure below shows a Scree Plot with Eigen Values.



The final task was to perform PCA using that number of components. Here is a Scree Plot using Sum Square Loadings, also showing the top 3 feature loadings:

The following were the attributes with Top 3 Loadings:

- **D2 : Sharing Factor (The number of shared terms)**
- **D4 : |U|/|A| (The ratio of cardinalities of the set of unprocessed clauses and the set of axioms)**
- **S11 : Maximum Clause Depth**



Lastly, we performed MDS using Euclidean and Correlation distances.

Results

- There are 10 principal components which impact heavily on how much time will a theorem take to be proven using first-order logic.
- The Maximum Clause Depth is the most important factor in deciding the said prediction.