Abstract

In recent times, many disciplines (like biology, chemistry or finance) have seen an explosion of data. The statistical methods face challenging tasks when dealing with such high-dimensional, multi-variate data. However, much of the data is highly redundant and can be efficiently brought down to a much smaller number of variables without a significant loss of information. The mathematical procedures making this reduction possible are called Dimensionality Reduction Techniques. Each and every technique reduces the dimensions of the data based on different criteria. This project has been done in three parts. In the first part, a simulation-based comparative study of variable selection was done in a linear-regression setting using a penalized-regression method - Least Absolute Selection and Shrinkage Operator (LASSO) versus univariate regression followed by the False Discovery Rate (FDR). Sensitivity, Specificity and Receiver Operating Characteristic (ROC) curves were used for comparison of these methods. In the second part, one of the Dimension Reduction Technique the Principal Component Analysis (PCA) was used to compare codon usage bias of HIV-1 viral genomes and genes to its human host using whole genome sequences. In the third part, Single Nucleotide Polymorphism (SNP) selection was done using Empirical Bayes strategy in Genome-wide Association Studies (GWAS).