Plagiarism Detector

Date: 2020-04-07

PLAGIARISM SCAN REPORT

3% Plagiarised

97% Unique

841 Words

5709 Characters

Exclude Url : None

Content Checked For Plagiarism

Problem Statement Computers need to decide the gender of a person from his/her voice, the dataset for the same is available but contains many features and handling the same is not an easy task. There is a need for an efficient method to handle all the features without negatively affecting the ability of classifying the test dataset, also due the large number of features the response time of the system is slow which should be improved. Abstract Machine Learning is used widely today to provide solutions to many problems, increase profits, improve processes and at times for research and development. Machine Learning algorithms are particularly useful in this example as the presented data can be used for predicting the gender of the voice using supervised learning techniques. Due to the high number of attributes at times it becomes extremely tedious to handle the data and also the attributes may be correlated which may violate some assumptions made by classification algorithms for example GNB, which assumes conditional independence amongst the attributes. The project uses three classification algorithms and compares accuracy obtained from each of these and time required before and after the use Principal Component Analysis. Libraries used for the same include numpy, scikit learn, matplotlib amongst others. The project is built in modular fashion to help in dynamic and plug and play programming, the project is capable of analysing any other similar file. H/W & S/W Requirements The project has certain hardware and software requirements as follows: Hardware : 64 bit processor, 512 MB RAM, 4 GB HDD/SSD Software : Fedora 20 OS (any other Open Source OS), python3 with scikit learn, numpy and pandas libraries, text editor (gedit). Introduction In many applications including robotics, chat bots, virtual assistants, e-commerce websites amongst others it is a required for the machine to understand the gender of the voice to make better suggestions to the user, for example, consider an e-commerce website will try to recommend apparels for women/men depending on the gender. Even Though, the dataset for the same is available it has many attributes which are correlated, to reduce the dimensionality of the feature space and to make the attributes conditionally independent Principal Component Analysis is used. Objective The mini project aims to attain the following objectives: Analyse provided data present data in readable format for any user Reduce the dimensionality of the feature space. Generate conditionally independent attributes. Classify the data as per the labels and predict the outcome for similar data. Compare the performance of different classification algorithms with respect to accuracy and time required for training them. Scope The project deals with data consisting of physical parameters of voice generated by a set of candidates, but names and unique ids of the candidates have not been used in the project and no group member had access to them during any time of the project. The output of the project consists of these: Analysis Reports Feature space file after PCA Explained Covariance Ratio Graph Accuracy Comparison Graph for classification algorithms Time Comparison Graph for classification algorithms The project can be further expanded into a full fledge system by adding following functionality: Automated report mailing to users Creating database and combining data from different sources Developing an adequate front end for the system Project Development Outline For the successful implementation of the project we undertook the following activities: Brainstorming for project field selection Requirement gathering for the selected topic Requirement analysis of the collected requirements Finalizing requirements and problem statement. Brainstorming for data analysis techniques Incorporation of changes Implementation of individual modules Testing of individual modules Exception handling in individual modules Integration of modules Integration testing Integrated Exception handling Final testing Project Presentation Project Report Submission Functions The project uses following functions to achieve the required objectives: analysis(data_frame) : Generates analysis reports pca(data_frame, threshold) : Returns reduced feature vector space GNB(data_frame,target_frame) : Performs Classification using Gaussian Naive Bayes Algorithm KNN(data_frame,target_frame) : Performs Classification using K Nearest Neighbor Algorithm SVM(data_frame,target_frame) : Performs Classification using Support Vector Machine Algorithm Usage of Principal Component Analysis Principal Component analysis is used to reduce the dimensionality of the feature space while using the effects of all the attributes. Consider a dataset with m data instances and n features represented as matrix ZmXn. It is multiplied with its transpose ZTnXm . Z’nXn = ZTnXm X ZmXn. This matrix Z’ is represented as PnXn X Dnxn X PnXnT, where P is the matrix of eigenvectors and D is a diagonal matrix of eigenvalues. The eigenvalues in D are sorted in descending order and accordingly P is changed to P*nXn and then Z is pre multiplied to this matrix to obtain the reduced feature space. Z*mXn = ZmXn X P*nXn There are mainly three methods used for deciding the number of components to be retained in the feature space. They are : Deciding static number of attributes to be retained Deciding a threshold on explained variance percentage Deciding the number of attributes using cumulative explained variance graph

3% Plagiarised

Feb 13, 2020 - The mini project aims to attain the following objectives: ? Clean data by using various data cleaning techniques. ? Pre-process the data so ...

https://www.scribd.com/document/446851892/DA-Project-Report-1