CS-7641 Machine Learning:

Unsupervised Learning and Dimensionality Reduction

Junle Lu
junle.lu@gatech.edu
Georgia Institute of Technology
Nov 4, 2018

Abstract – This paper investigates unsupervised learning algorithms for clustering and dimensionality reduction on two datasets. For clustering, the implemented algorithms are k-means clustering and expectation maximization. For dimensionality reduction, the implemented algorithms are principal component analysis, independent component analysis, randomized projection, and truncated singular-value decomposition. All the algorithms are implemented in scikit-learn python library. Using the dimensionality reduction algorithms, the reduced data or new space are applied with the clustering algorithms and the neural network classifier(MLP).

1. Introduction

In the last paper, the performance of a neural network was analyzed with two datasets, but only one is chosen to be used for the experiments: gender recognition by voice.

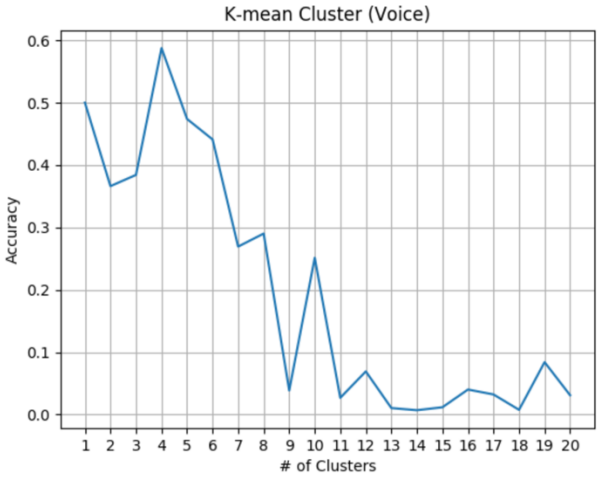
Gender Recognition by Voice: this database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz. It has 3168 instance and 20 features.

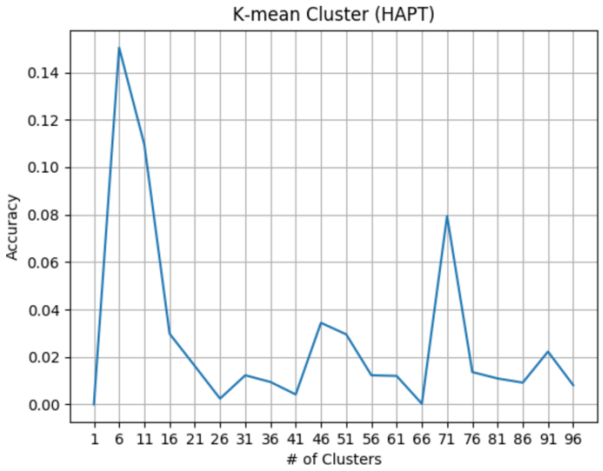
Human activity recognition (HAR): The experiments were carried out with a group of 30 volunteers within an age bracket of 19-48 years. They performed a protocol of activities composed of six basic activities: three static postures (standing, sitting, lying) and three dynamic activities (walking, walking downstairs and walking upstairs). The experiment also included postural transitions that occurred between the static postures. These are: stand-to-sit, sit-to-stand, sit-to-lie, lie-to-sit, stand-to-lie, and lie-to-stand. All the participants were wearing a smartphone (Samsung Galaxy S II) on the waist during the experiment execution. The majority of the smartphones are embedded with tri-axial accelerometer and gyroscope sensors which are crucial for the HAR system. In most smartphones, these sensors are primarily used for automatic screen rotation and indirect user control by detecting the phone's orientation as user inputs. The processing power of the smartphones is more than enough to implement HAR system. In particular, this dataset was generated based on the sensor readings from a Samsung Galaxy II smartphone.

2 Clustering Algorithms

2.1 k-means Clustering

K-mean clustering simply group all the similar data points into different clusters. The voice dataset only has two labels, which means two clusters are expected. We ran the data with different k values and compare the prediction accuracy.

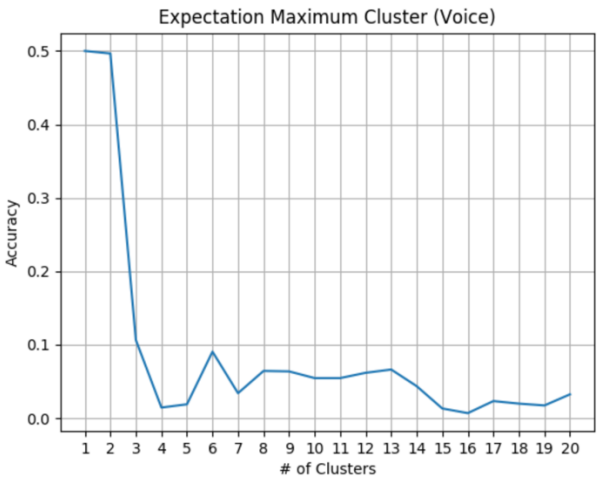


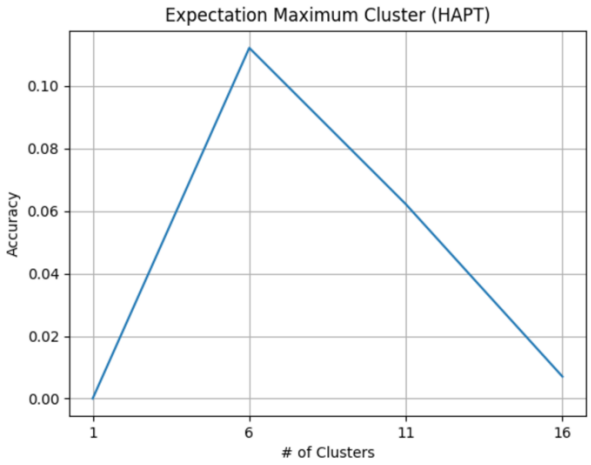


For both datasets, they are performing as expected. For the voice dataset, the accuracy goes down significantly after 5 clusters. It has only two actual labels. For HAPT dataset, the accuracy goes down significantly after 6 clusters, which is expected number of activities to classify. The accuracies are not so great here.

2.2 Expectation Maximization (EM)

The expectation maximization used is a Gaussian mixture model. It is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians.



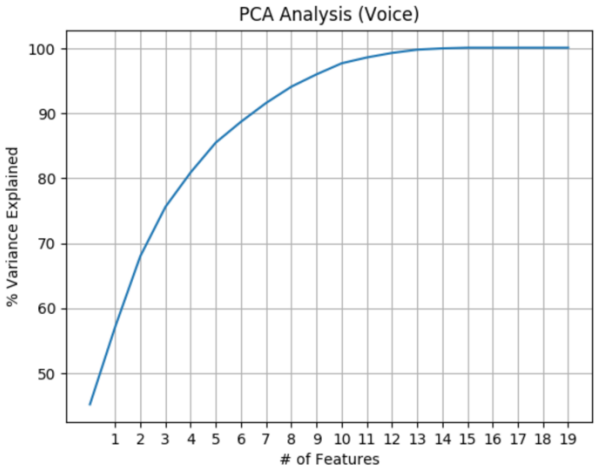


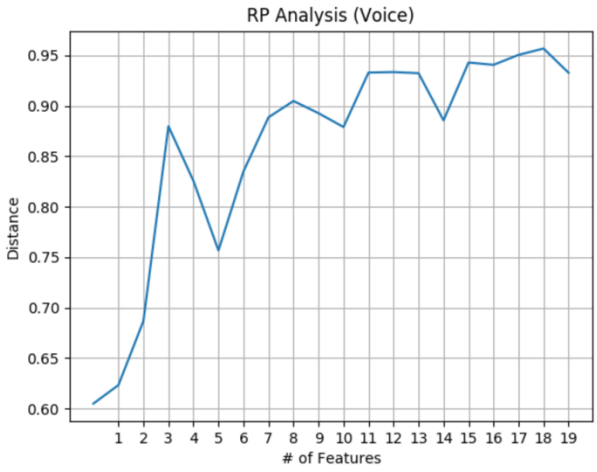
From the graphs below, this model strongly agree with the number of labels that the datasets are designed: 2 and 6.

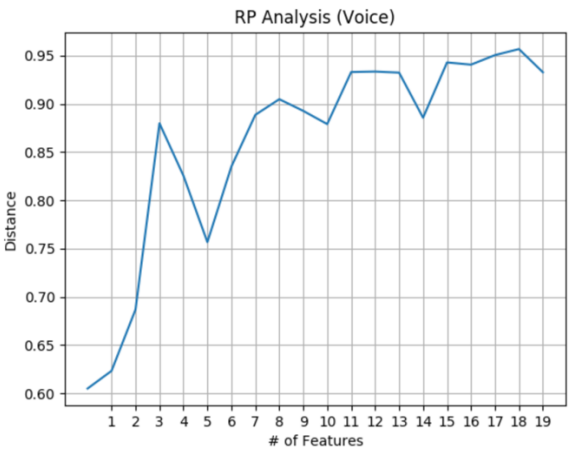
3 Dimensionality Reduction

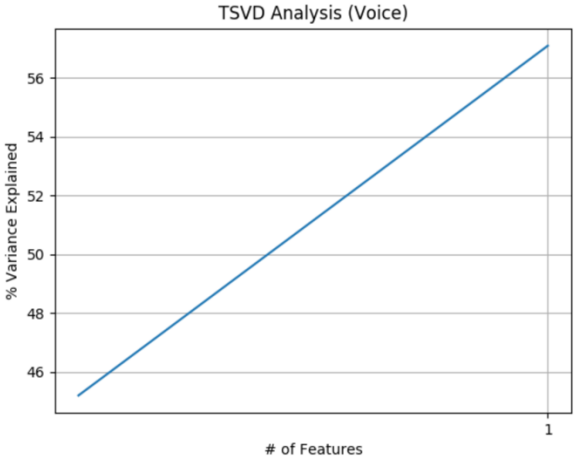
3.1 Principal Component Analysis

PCA is used to decompose a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance. In scikit-learn, PCA is implemented as a transformer object that learns components in its fit method, and can be used on new data to project it on these components. The percentage of the variance with number of features are computed to find the optimal number of feature to extract. The voices dataset only has 10 features in total, and the HAPT data has 561 features. Independent component analysis separates a multivariate signal into additive subcomponents that are maximally independent. It is implemented in scikit-learn using the Fast ICA algorithm. The random projections reduces the dimensionality by projecting the original input space using a sparse random matrix. Sparse random matrices are an alternative to dense Gaussian random projection matrix that guarantees similar embedding quality while being much more memory efficient and allowing faster computation of the projected data. This transformer performs linear dimensionality reduction by means of truncated singular value decomposition (SVD). Contrary to PCA, this estimator does not center the data before computing the singular value decomposition.

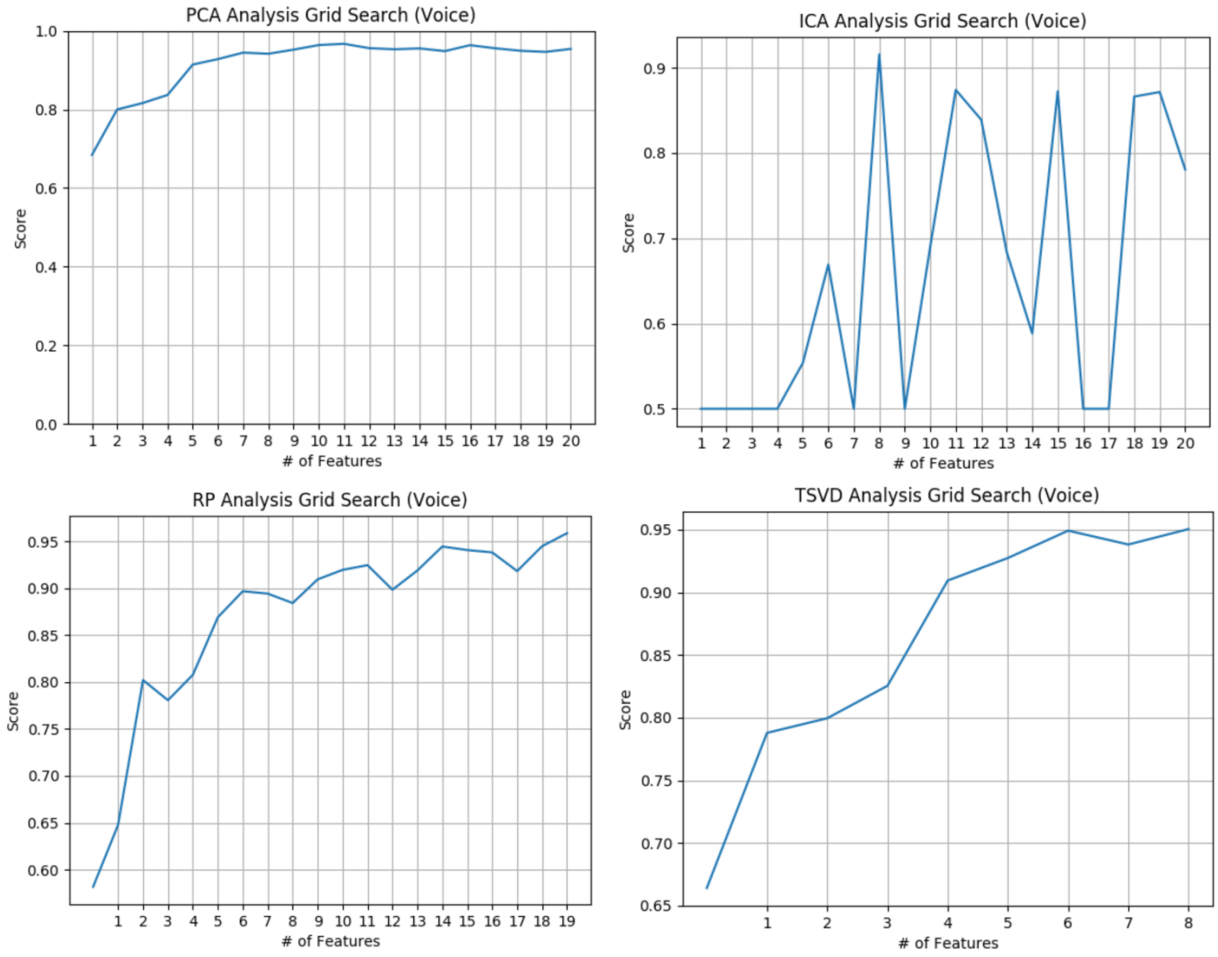

PCA Analysis (Voice)
% Variance Explained
# of Features


RP Analysis (Voice)
Distance
# of Features


RP Analysis (Voice)
Distance
# of Features


TSVD Analysis (Voice)
% Variance Explained
# of Features

4. Neural Network with Dimension Reduction



The grid search is applied to all four dimensionality reduction algorithms. From the graphs, the best number of features are between 4 to 6

5. Conclusion

The dimension can be used to reduce the number features used for the classification algorithm without significant lose in the accuracy.