Boheng Lin

Tufts University

Inconsistency of K-Means in Clustering High Dimensional Data Projected on Principal Components

Introduction

In unsupervised learning, K-Means is generally used for data with low dimensionalities as it often fails identifying the correct cluster when dimensionalities of the data increases. One method attempting to mitigate such error is to reduce the dimensionality using principal component. However, obtaining or measuring most of the dimensions for a multidimensional data may be difficult and K-means may inappropriately be applied to such data that has been force-reduced in dimension. In this project, a data matrix of dimension 569 * 32 is studied using K-Means to access its accuracy in identifying each data points. With 100 iterations of the K-Means over the data points projected on the first 2 principal component, any trend in accuracy will be analysed.

Data[1]

For this project, data of 569 breast cancer cells published by University of California Irvine machine learning repository is utilized with 32 features including a unique ID, a binary diagnosis of benign or malignant, and the mean, standard error and the largest or worst-case measure of the following 10 features

Radius	Texture	Perimeter	Area	Smoothness
Compactness	Concavity	Concavity Points	Symmetry	Fractal Dimension

Methodology

To reduce effects of multicollinearity and high dimensionality in applying K-Means, dimension reduction will be done using principal components analysis (PCA). Each PC is a direction vector in which the data points are most spread out in that level of dimension and is orthogonal to the one before and after. The proportion of the eigenvalues corresponding to each PC shows the extent to which each PC explains the variation in data. With explained ratio, top PCs will be selected to simulate a force-reduction on dimension and K-Means to be applied.

K-Means algorithm will generate two hypothesized centroids as the label representation of each label. The algorithm will first assign each data point to the centroid label closer to it. After assigning all the points, both centroids will be updated as the spatial mean of all the points of that label. Such assignment and update process is repeated 100 times to allow stabilisation of K-Means centroids. Labels classified using K-Means algorithm then to be compared to that in the original data and accuracy rate will be determined by the formula:

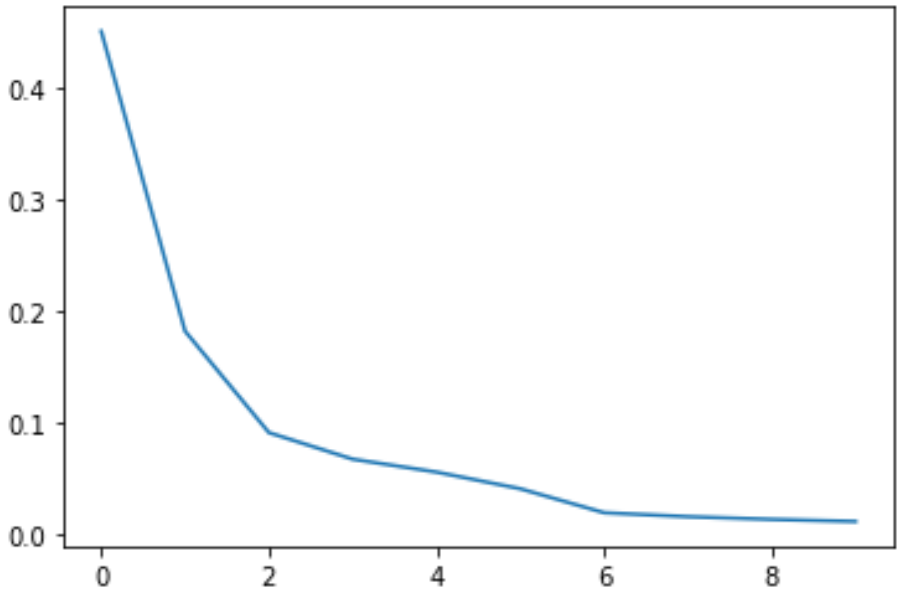
$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

[1] Data obtained from University of California Irvine Machine Learning Dataset Archive, via https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

Analysis

As part of the exploratory analysis and in attempts to have a preliminary understanding of how each of thirty variables perform against being 'benign' and being 'malignant', thirty plots of distributions of each variable against number of observations are constructed (Fig. 1). From the plot, it is clear that many of the features such as Mean Fractal Dimension, Standard Error in Texture, Standard Error in Smoothness, Standard Error in Concavity, Standard Error in Symmetry, Standard Error in Fractal Dimension and Worst Fractal Dimension have to large extent the same distribution, and the malignant distribution is almost entirely overlapped by benign distribution. In addition, along no feature is the two distribution completely distinct. This suggests that analysis done in any single variable will inherently produce misclassification errors.

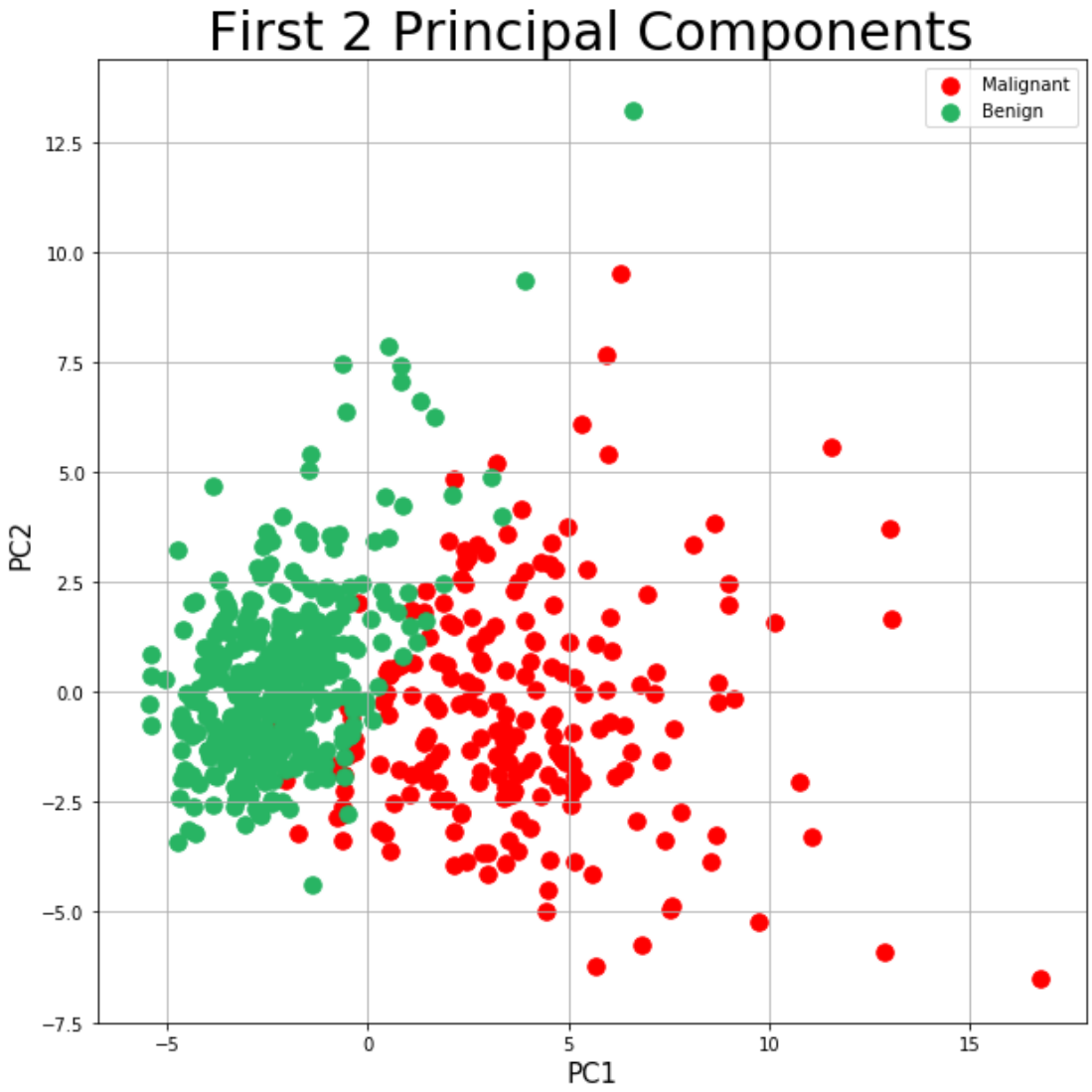
In order to avoid inheriting such errors, one of the most commonly used method is to analyze along principal components rather than variable dimensions. To account for at least 95% of the variance in data, 10 principal components have been computed. From the diminishing marginal variance explained plotted below, despite the over 95% explained variance by top 10 PCs, only the top two accounts for the most significant variance explanation.



The following table displays both the variance explained and cumulative variance explained at each principal component level. The top two principal components are deemed as most significant each account for more than 15% of variance explained.

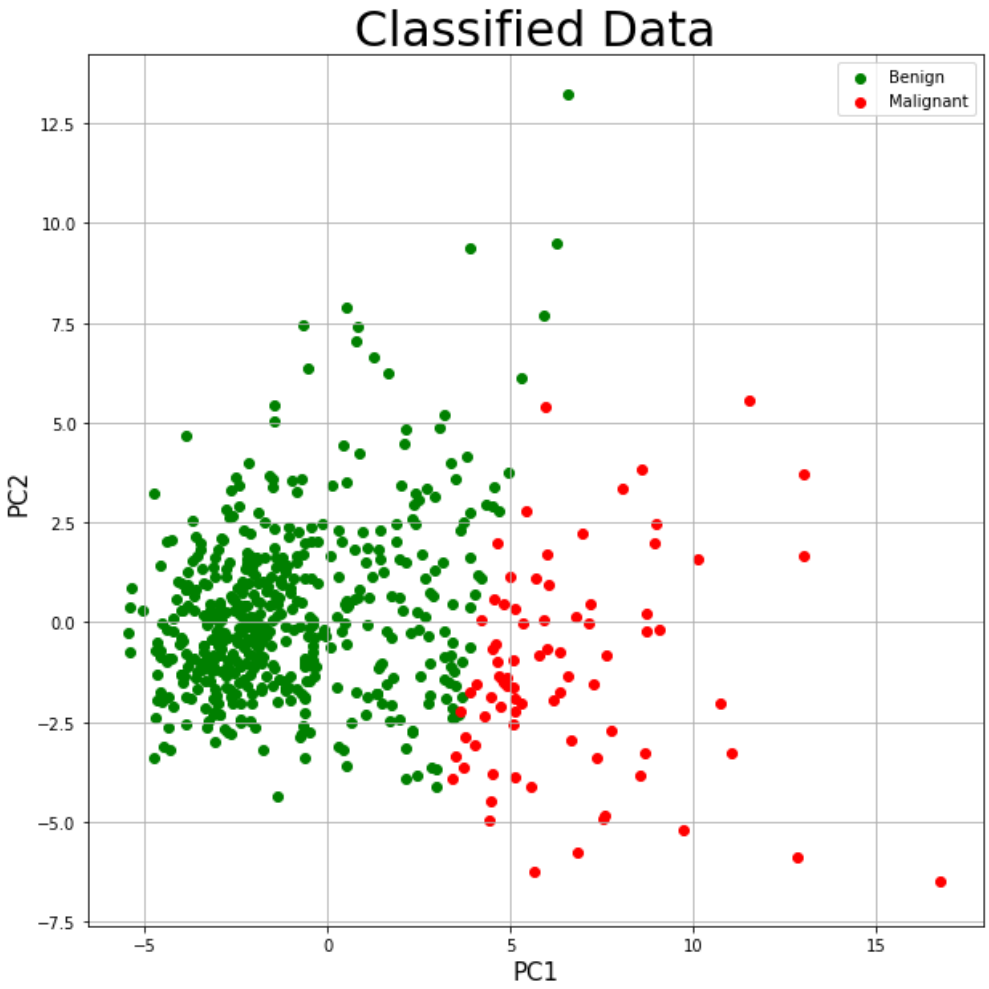
	Variance Explained (3 s.f.)	Cumulative Variance Explained
PC1	45.1%	45.10%
PC2	18.2%	63.30%
PC3	9.16%	72.46%
PC4	6.78%	79.24%
PC5	5.63%	84.87%
PC6	4.14%	89.01%
PC7	1.99%	91.00%
PC8	1.63%	92.63%
PC9	1.40%	94.03%
PC10	1.21%	**95.24%**

Each data is projected on the two principal components and standardized such that the mean of the data is set at the origin (0,0). Visually, the plot demonstrated some degrees of difference between malignant and benign cells or data as well as some degrees of similarities.

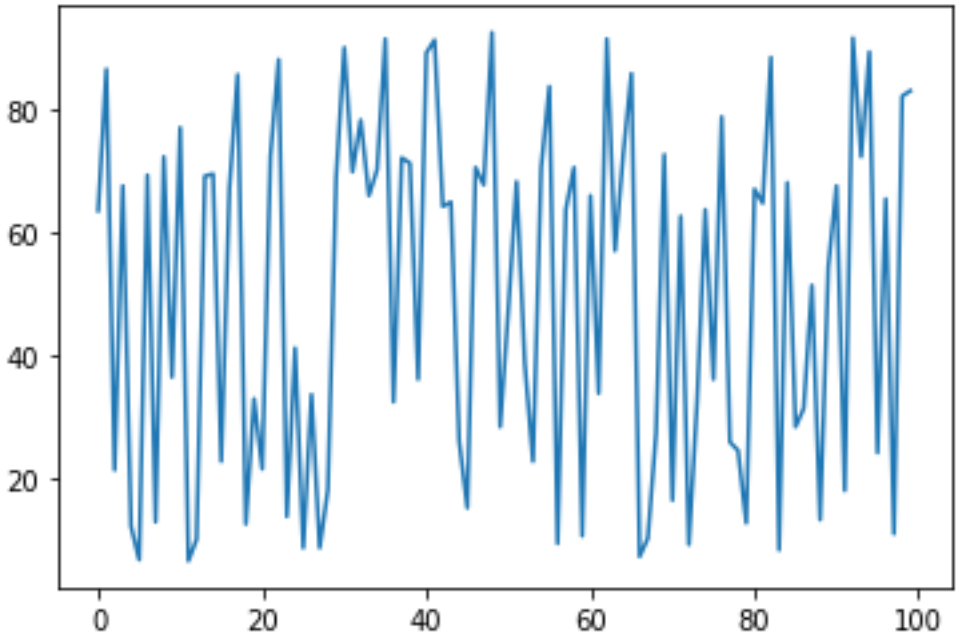


For both labels, more data points are found at relatively lower left coroner with few anomalies spreading out rightwards and upwards while the malignant data has relatively higher value in PC1. The sole visual difference on the location of data of different labels is to a large extent without absolute reference even though there exists a visual line between data from each label.

With the labels removed from the projection, K-Means are applied to reidentify each data point and assign relevant label – malignant or benign – to it. The preliminary round of K-Means classification produces the following visual results with an accuracy rate of 76.8%. This suggests that about a quarter of the cells have been misclassified either false positive or false negative.



With a repetition of 100 times deploying the same K-Means algorithm on the same data set, the accuracy rate shows no visual trend or change over the number of repetitions.



Repetition	Min Acc.	Med Acc.	Mean Acc.	Max Acc	SE. Acc
100	6.50%	63.53%	50.10%	92.44%	2.83%

While the top two principal components account for the most degree of variation in the data, reducing the dimension of the data from thirty dimensions to two is likely to introduce more and larger random errors than doing analysis using variables themselves. Therefore, to consider whether KMeans algorithm will perform better when the reduction in dimension is less, clustering using the top five principal components and the top ten principal components, each explains about 85% and 95% variation in the data respectively, are conducted over 200 iterations of the KMeans algorithm. The results plots are in Fig.2

Conclusion

Accuracy rates in identifying benign and malignant cells fail to converge to a reasonably useful measure of more than 50% in all cases. The large fluctuation and instability of the accuracy rate and un-trended feature of the accuracy rate by deploying K-Means multiple times suggests the inappropriateness of K-Means in dimension reduced data set even as the data points are projected on most important principal components. Even as the projection consists of all principal components that explain 95% of variation in the data, KMeans algorithm fail to improve on its convergence on accuracy rate. Force reduction of dimensionality on data will make K-Means classification highly unreliable, inconsistent and of minimum effectiveness in classifying unknown labels. In addition, data passed in K-Mean with unknowingly multiple omitted dimensions is likely to similar problems.

Appendix

Fig 1. Summary plot of benign and malignant cell distribution along 30 different variables.

Fig.2 Accuracy Rate (%) of the K-Means Algorithm Using 5 & 10 PCs