Analysis of Tumor Types

Rosebella Capio

University of Idaho

November 27, 2018

Outline

Introduction

Background on Tumors

A tumor, also known as neoplasm, is an abnormal mass of tissue that can be solid or fluid-like. That is, a tumor is a kind of lump or swelling that does not necessarily pose a health threat. A tumor is NOT the same as CANCER although some can develop into one.

General Types of Tumors

- **Benign**: These are not cancerous, do not spread, remain in its current form and do not return after being removed
- **Premalignant**: These are not yet cancerous but appear to be developing the properties of cancer
- **Malignant**: They are cancerous, grow, spread and get worse

Introduction Cont'd

Tumors Types Considered

The tumors considered in this work are;

- **BRCA:** Breast Invasive Carcinoma
- **KIRC:** Kidney Renal Clear Cell Carcinoma
- **COAD:** Colon Adenocarcinoma
- **LAUD:** Lung Adenocarcinoma

Introduction Cont'd

Data Source

- The dataset was obtained from the UCI Repository
- It consists of random extraction of gene extractions of patients having different tumor types

Introduction Cont'd

Objectives

- Perform Classification Analysis
- Perform Clustering Analysis

Outline

Analysis

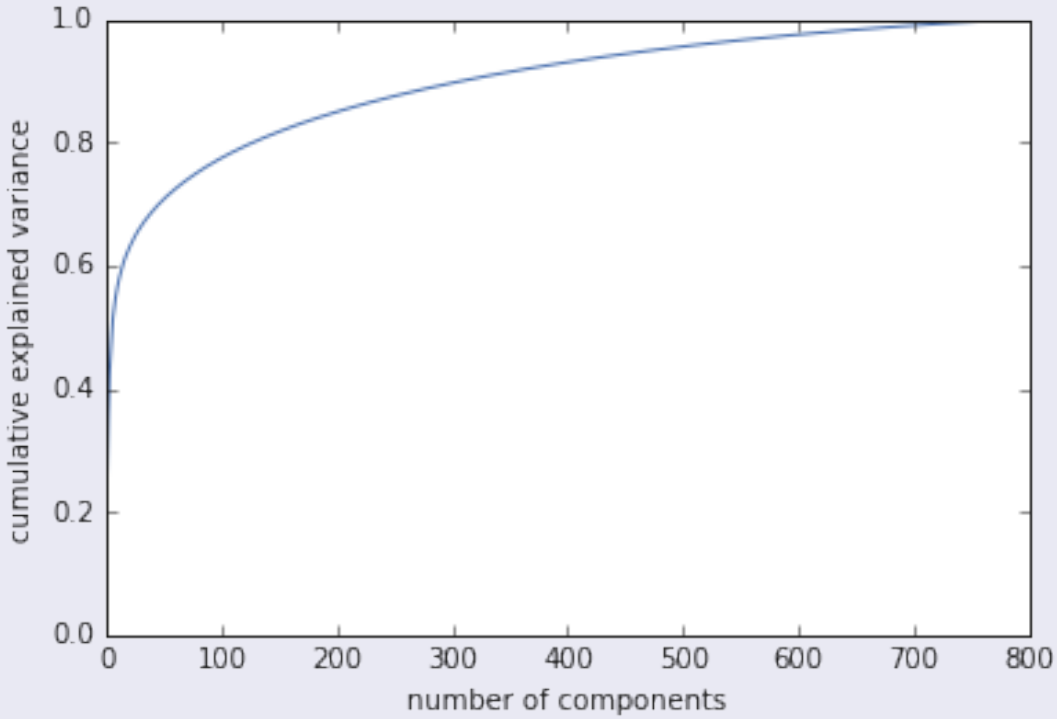
Exploratory Data Analysis

- It has 801 observations and $20,532$ features
- The feature names are gene1 to gene$20,532$ and the observations are patients with response variable labelled as different tumor types: BRA, KIRC, COAD, LUAD, PRAD.
- The features are numerical
- There are no missing values in this data set

Analysis

Dimensionality Reduction

PCA was used in reducing the dimension of the data. The graph below shows that about 700 components can explain all the variation in the data.



Classification Models

Logistic Regression

Logistic regression performed poorly on the training data set with accuracy of .29 with very low precision values for each class as seen in the output below;

```
Accuracy of Logistic regression classifier on training set: 1.00
Accuracy of Logistic regression classifier on test set: 0.29
```

Logistic Regression

	precision	recall	f1-score	support
0	0.31	0.30	0.31	1007
1	0.28	0.27	0.28	1002
2	0.32	0.30	0.31	998
3	0.27	0.30	0.29	1005
4	0.26	0.25	0.25	988

Classification Models

K Nearest Neighbor

The accuracy score from using KNN on the data was .99379 implying that the KNN classifies the data very well. Precision for each class was almost 1 or 1. The output can be seen below;

```
Accuracy Score: 0.99379
```

K Nearest Neighbor

	precision	recall	f1-score	support
BRCA	0.98	1.00	0.99	55
COAD	1.00	1.00	1.00	17
KIRC	1.00	1.00	1.00	25
LUAD	1.00	0.97	0.98	32
PRAD	1.00	1.00	1.00	32

Classification Models

Random Forest

Random Forest classification also did a poor job at classifying the tumor types and had the following values for precision;

	precision	recall	f1-score	support
0	0.258893	0.548117	0.351678	239
1	0.328947	0.187970	0.239234	266
2	0.377309	0.600840	0.463533	238
3	0.128492	0.090551	0.106236	254
4	0.323529	0.043478	0.076655	253

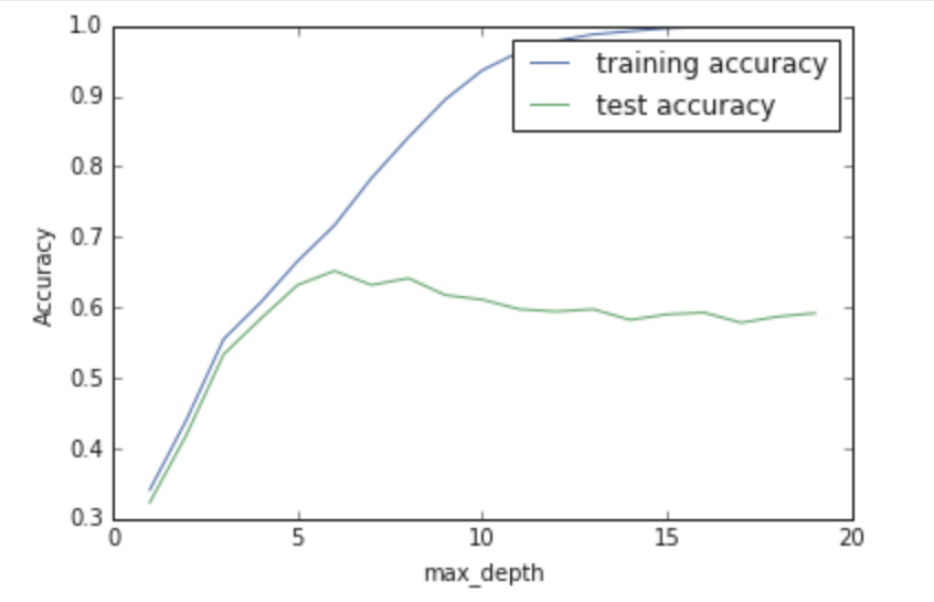
Classification Models

Decision Trees

Decision Trees gave a training set accuracy of 1 and a little over 0.5 for the test set; The plot of the graph is shown;

```
Accuracy on training set: 1.000
Accuracy on test set: 0.583
```

Decision Trees



Classification Models

Naive Bayes

Naive Bayes produced a test accuracy score of .25 implying that it is a poor classifier for this data set;

```
Accuracy of Naive Bayes classifier on training set: 1.00
Accuracy of Naive Bayes classifier on test set: 0.25
```

Classification Models

Neural Networks

Neural Networks also produced a test accuracy score of .27 but 1 for the training set implying that it is a poor classifier for this data set;

```
Accuracy of NN classifier on training set: 1.00
Accuracy of NN classifier on test set: 0.27
```

Classification Models

Support Vector Machine

Support Vector Machine produced similar scores as those produced by Naive Naive Bayes and Neural Networks hence a poor classifier for this data set;

```
Accuracy of SVC classifier on training set: 1.00
Accuracy of SVC classifier on test set: 0.26
```

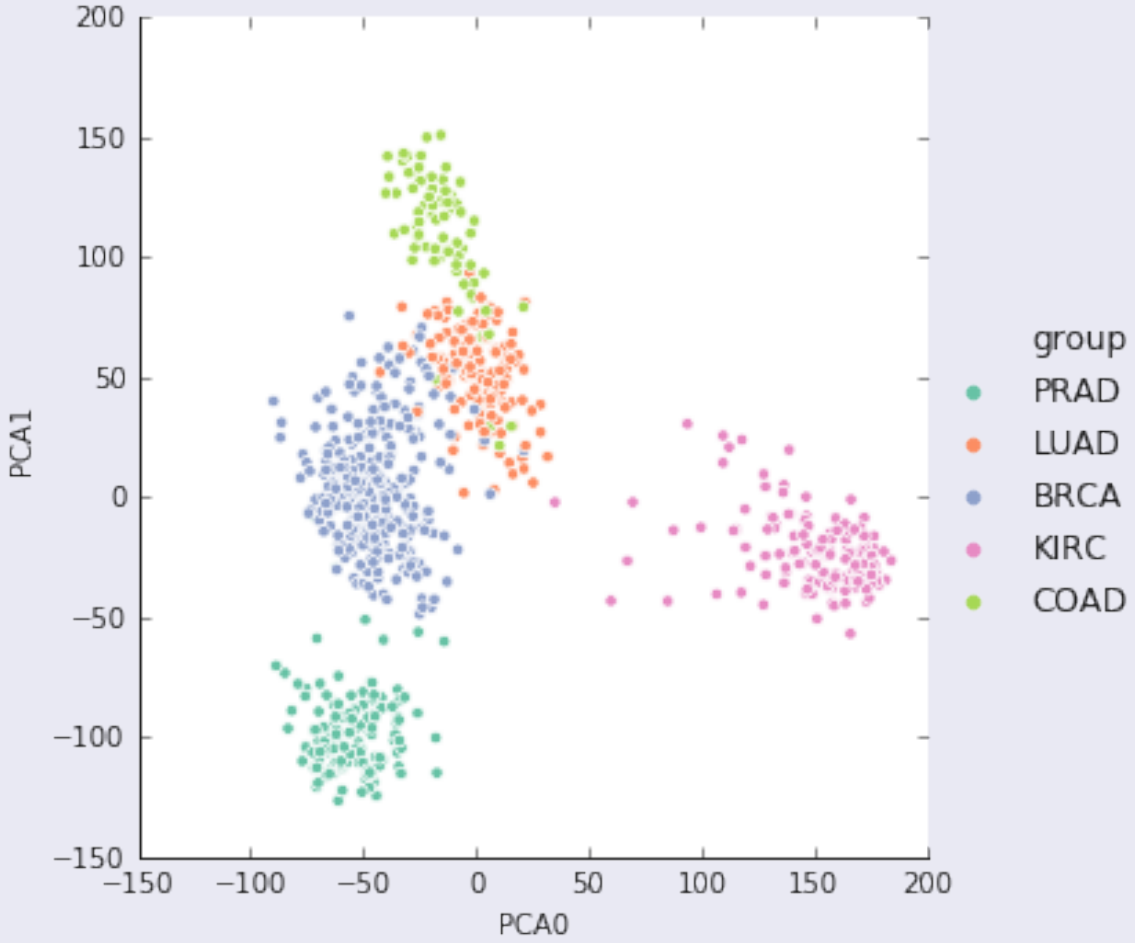
Comparison of Test Accuracy Scores for Classification Models

- Logistic Regression: **0.29**
- KNN: **0.99379**
- Decision Trees: **0.583**
- Naive Bayes: **0.25**
- Neural Networks: **0.27**
- SVM: **0.26**

Outline

Plot of tumor clusters from PCA

PCA tumor clusters



Crosstab of tumor types against PCA clusters

Crosstab

group	BRCA	COAD	KIRC	LUAD	PRAD	All
cluster						
0	0	1	1	138	0	140
1	0	0	0	0	136	136
2	0	0	145	0	0	145
3	50	0	0	3	0	53
4	250	0	0	0	0	250
5	0	77	0	0	0	77
All	300	78	146	141	136	801

Clustering

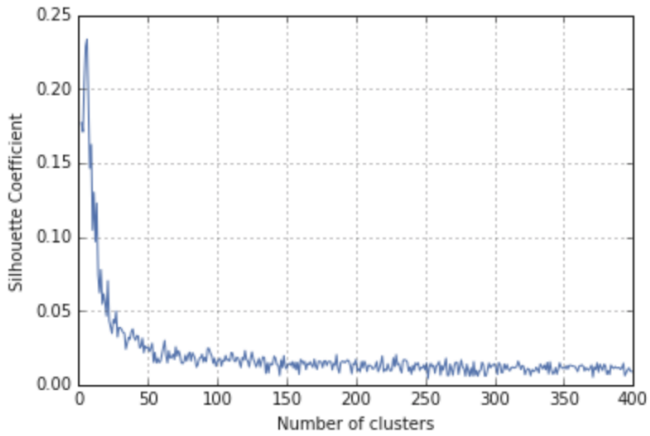
K-Means

K-Means clustering was tested on this classification problem and produced a low sihouette score and produced an optimal number of clusters for PCA and non-PCA data to be 6;

```
0.2287493724969654
```

K Means

```
The optimal number of clusters is 6
```



Conclusion

Conclusion

- Of all the classification models looked at, K Nearest Neighbor provided the highest accuracy with a score of .99 followed by Decision Trees with .583. The rest did not perform so well as classifiers.
- Association could not be performed on this data set but will be addressed in final report

THANK YOU