MS-E2112 – Multivariate Statistical Analysis

Project Report

Classification of Breast cancer based on Linear Discriminant Analysis

Yizhan Yang

Master of Science in **ICT Innovation**

Major in **Autonomous Systems**

Aalto University
School of Electrical Engineering

Date: 5/4/2019

1. Data Description

In this report, the Breast Cancer Coimbra Data Set was imported from UCI Machine Learning Repository.

The multivariate dataset contains 116 instances and 9 attributes (Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1. The 9 attributes indicating the presence or absence of breast cancer.

The attributes are anthropometric data and parameters which can be gathered in routine blood analysis.

2. Research Proposal

The main question is to create a classifier of the given data, and classify them into two different categories – Health Controls and patients. Linear Discriminant Analysis is the main tool to solve the research question.

3. Univariate and Bivariate Analysis

The real dataset was plotted by univariate scatter plots (see fig.1). It is clear that the group Insulin and HOMA are correlated.

As the amount of the data is much less, the data points do not indicate specific distribution.

Moreover, the correlation plot (see fig.2), which is completed by corrplot package, indicates that features Insulin and HOMA are highly correlated with correlation coefficients 0.93219777 respectively.

Age	57.301724
BMI	27.582111
Glucose	97.793103
Insulin	10.012086
HOMA	2.694988
Leptin	26.615080
Adiponectin	10.180874
Resistin	14.725966
MCP.1	534.647000

Fig.1 Means

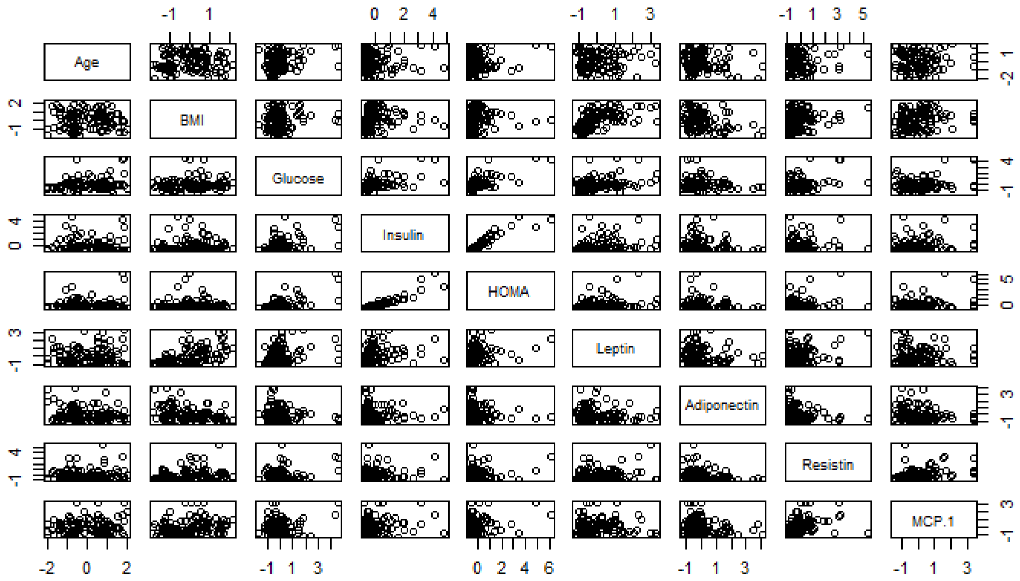

Fig.2 Scatter Plot

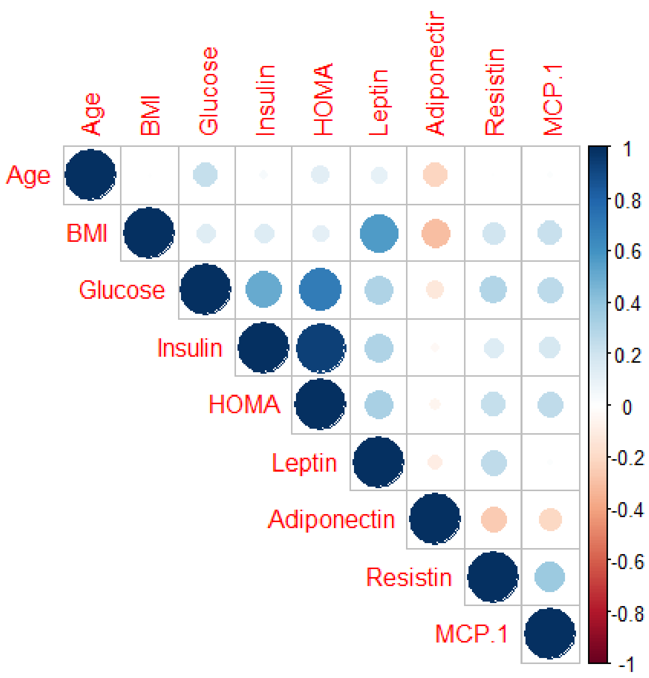

Fig.3 Correlation Plot

In addition, to inspect the data deeper, histogram plots are required (see fig.3). The interval of all the data is [0,1.5 x 10^3]. We can find that group Age and BMI has normal distribution. And attributes Glucose, Insulin, HOMA, Resistin have outliers points.

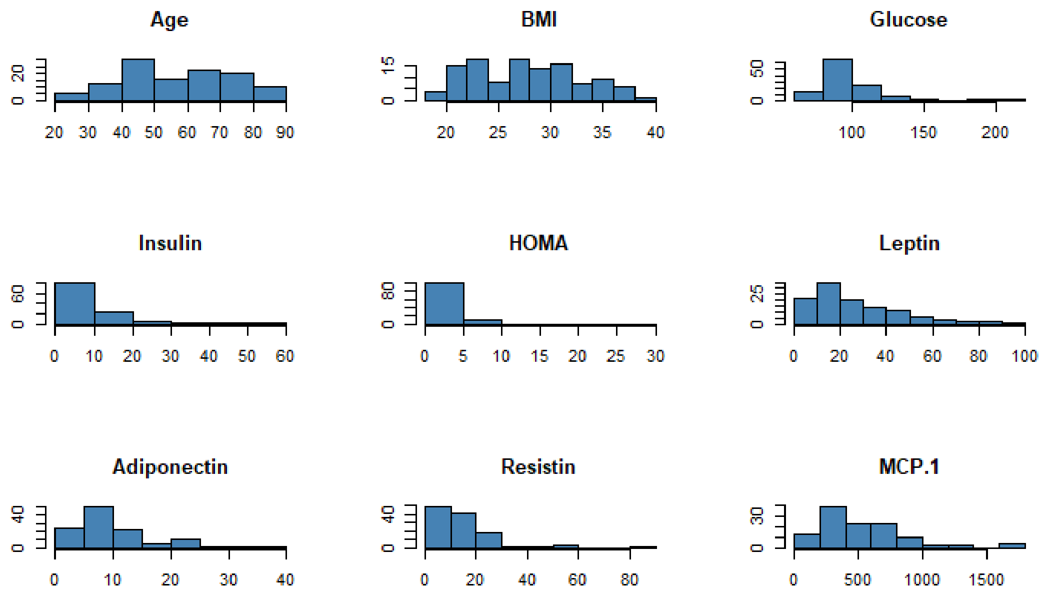

Fig.4 Histogram

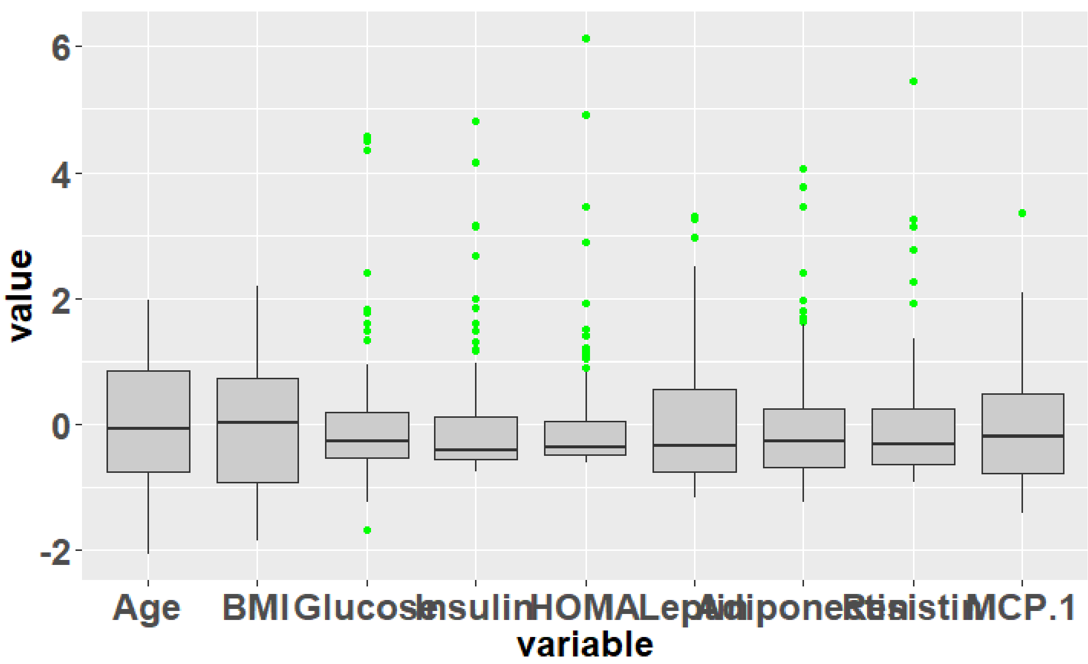

Fig.5 Outliers

Based on the boxplot (see fig.4), the range of the data is varying. Outliers points seems to be too much. The large range of the data leads to the sensitive boxplot.

4. Linear Discriminant Analysis

Linear discriminant Analysis can classify the health people and the patients.

After dividing the dataset into 80% training and 20% test data is 75%. The result and error are shown (see Fig.5).

```
    truth
est  1  2
  1 39 18
  2 13 46
> 1-sum(diag(tab))/nrow(cancer)
[1] 0.2672414
>
```

Fig.6 Result and Error

The result and error are shown above (see Fig.5). The error of cross validation is 26.7%, which means the accuracy is 73.3%.

5. K-Means

K-Means algorithms can easily classify create two clusters of breast cancer, before apply K-means, we should apply PCA first.

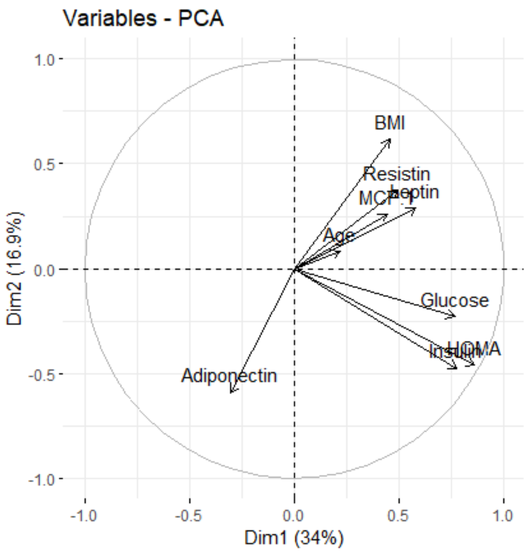

Fig.7 PCA

Then, we apply K-Means with K equals 2.

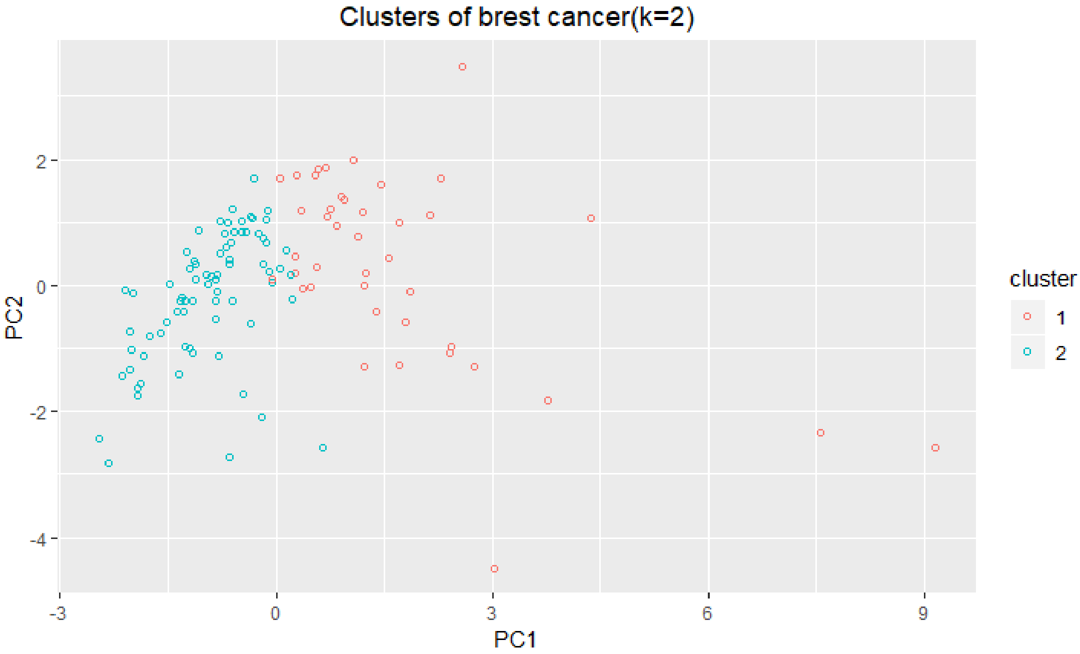

Fig.8 K-Means

6. Conclusions

As a result, the classification task was hardly done with linear discriminant analysis, since the outliers and inseperatable of the real world data. As the accuracy reached 73.3%, the classifier was good. One possible reason for the reduction of the accuracy could be some data cannot be separated, as it is indicated by scatter pair plot and plot of K-Means Algorithm.

7. Appendix

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Age	1.000000000	0.008529857	0.2301056	0.03249535	0.12703259	0.10262605	-0.21981289	0.002741708	0.01346168
BMI	0.008529857	1.000000000	0.1388452	0.14529526	0.11448013	0.56959261	-0.30273476	0.195350206	0.22403821
Glucose	0.230105617	0.138845189	1.0000000	0.50465307	0.69621182	0.30507994	-0.12212131	0.291327462	0.26487927
Insulin	0.032495353	0.145295260	0.5046531	1.00000000	0.93219777	0.30146162	-0.03129608	0.146730986	0.17435580
HOMA	0.127032593	0.114480131	0.6962118	0.93219777	1.00000000	0.32720986	-0.05633712	0.231101229	0.25952919
Leptin	0.102626049	0.569592606	0.3050799	0.30146162	0.32720986	1.00000000	-0.09538874	0.256233522	0.01400862
Adiponectin	-0.219812891	-0.302734758	-0.1221213	-0.03129608	-0.05633712	-0.09538874	1.00000000	-0.252363303	-0.20069450
Resistin	0.002741708	0.195350206	0.2913275	0.14673099	0.23110123	0.25623352	-0.25236330	1.000000000	0.36647421
MCP.1	0.013461678	0.224038215	0.2648793	0.17435580	0.25952919	0.01400862	-0.20069450	0.366474210	1.00000000

Fig.9 Correlation

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
Age	259.6212144	0.6899654	83.51514	5.271382	7.454703	31.72129	-24.237703	0.5473753	75.03014
BMI	0.6899654	25.2017631	15.70053	7.343449	2.093103	54.85333	-10.400302	12.1513148	389.04875
Glucose	83.5151424	15.7005265	507.38291	114.444261	57.115555	131.82711	-18.824679	81.3098725	2063.87004
Insulin	5.2713825	7.3434492	114.44426	101.359945	34.181124	58.22217	-2.156212	18.3041243	607.20600
HOMA	7.4547030	2.0931033	57.11556	34.181124	13.264479	22.86097	-1.404132	10.4289668	326.96236
Leptin	31.7212930	54.8533291	131.82711	58.222171	22.860971	367.99877	-12.522427	60.9050169	92.95764
Adiponectin	-24.2377028	-10.4003022	-18.82468	-2.156212	-1.404132	-12.52243	46.831322	-21.3987474	-475.08370
Resistin	0.5473753	12.1513148	81.30987	18.304124	10.428967	60.90502	-21.398747	153.5281003	1570.73824
MCP.1	75.0301391	389.0487525	2063.87004	607.205999	326.962361	92.95764	-475.083702	1570.7382390	119655.57060

Fig.10 Covariance