IHE-7510-01 - Data Mining

Project 1

Meenakshi Nagarajan
Nagarajan.12@wright.edu

#Dataset – Banknote Authentication
http://archive.ics.uci.edu/ml/datasets/banknote+authentication

#Initially the dataset has been loaded into a variable called ‘mydata’

```
> #Meenakshi Nagarajan
> #nagarajan.12@wright.edu
> library("dplyr")

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> #Load the data into 'mydata'
> mydata=read.csv("/Users/meenakshinagarajan/Desktop/Datamining/banknote_authentication.csv")
> head(mydata)
  Variance Skewness Curtosis  Entropy Class
1  3.62160   8.6661  -2.8073 -0.44699     0
2  4.54590   8.1674  -2.4586 -1.46210     0
3  3.86600  -2.6383   1.9242  0.10645     0
4  3.45660   9.5228  -4.0112 -3.59440     0
5  0.32924  -4.4552   4.5718 -0.98880     0
6  4.36840   9.6718  -3.9606 -3.16250     0
```

Fig. 1 *Read the data*

#Once the data has been loaded, the summary of data has been analyzed.

Background of data
The image of a bank note is pre-processed and the classification features were extracted. Later the note is classified as genuine or forged note based on the features extracted. In this project, the correct combination of features used to determine the authenticity of the banknote is being identified statistically.

The dataset used in this study is obtained from the UCI Machine learning repository. (**http://archive.ics.uci.edu/ml/datasets.html**)

1	**Dataset Characteristics**	Multivariate	4	**Number of instances**	1372
2	**Attribute characteristics**	Real	5	**Number of Attributes**	5
3	**Date Donated**	2013/ 04/ 16	6	**Missing values**	None

Attributes

Attribute Name	Datatype	Meaning
Variance	Numerical	Variance gives the amplitude distribution of the Wavelet coefficients around the center of histogram.
Skewness	Numerical	Skewness is the symmetry of the distribution of data around the center.
Kurtosis	Numerical	Kurtosis describes the deviation relative to the Gaussian distribution.
Entropy	Numerical	Entropy/ average information of an image.
Class	Binary	Authenticity (Output 0 means genuine and Output 1 means forged).

Insights

```
> summary(mydata)
    Variance          Skewness          Curtosis          Entropy      
 Min.   :-7.0421   Min.   :-13.773   Min.   :-5.2861   Min.   :-8.5482  
 1st Qu.:-1.7730   1st Qu.: -1.708   1st Qu.:-1.5750   1st Qu.:-2.4135  
 Median : 0.4962   Median :  2.320   Median : 0.6166   Median :-0.5867  
 Mean   : 0.4337   Mean   :  1.922   Mean   : 1.3976   Mean   :-1.1917  
 3rd Qu.: 2.8215   3rd Qu.:  6.815   3rd Qu.: 3.1793   3rd Qu.: 0.3948  
 Max.   : 6.8248   Max.   : 12.952   Max.   :17.9274   Max.   : 2.4495  
     Class       
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :0.0000  
 Mean   :0.4446  
 3rd Qu.:1.0000  
 Max.   :1.0000  
```

Fig. 2 *Summary of data*

#Histogram for all the attributes has been constructed to check the data distribution

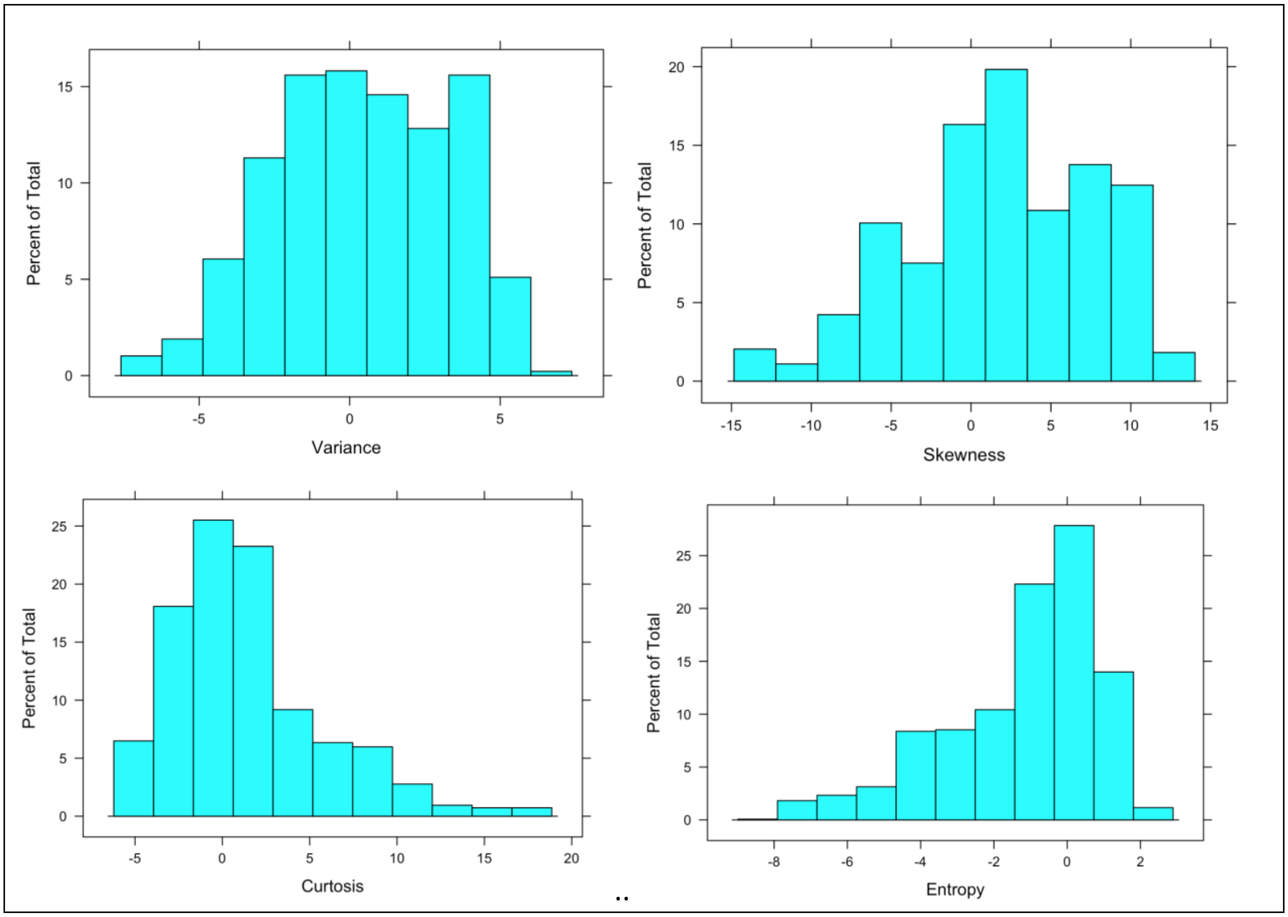

***Fig. 3** Histogram*

#From the distribution of data, some outlying observations could be seen for Kurtosis and Entropy.
#Here, it is not evident that the outliers are due to recording errors. Therefore, they are not removed for the data analysis

```
> #correlation of data
> out <- cor(mydata)
> round(out, 2)
          Variance Skewness Curtosis Entropy Class
Variance      1.00     0.26    -0.38    0.28 -0.72
Skewness      0.26     1.00    -0.79   -0.53 -0.44
Curtosis     -0.38    -0.79     1.00    0.32  0.16
Entropy       0.28    -0.53     0.32    1.00 -0.02
Class        -0.72    -0.44     0.16   -0.02  1.00
```

***Fig. 4** Correlation matrix*

#Correlation matrix has been constructed to see how well the variables are correlated with each other.
#Here, high negative correlation is observed between Skewness and a decent positive correlation is observed between kurtosis and entropy.
#using the library 'corrplot', a graphical display of correlation matrix is obtained, which can be used for quick interpretation

```
> library(corrplot)
> #correlation plot
> corrplot(out, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)
```

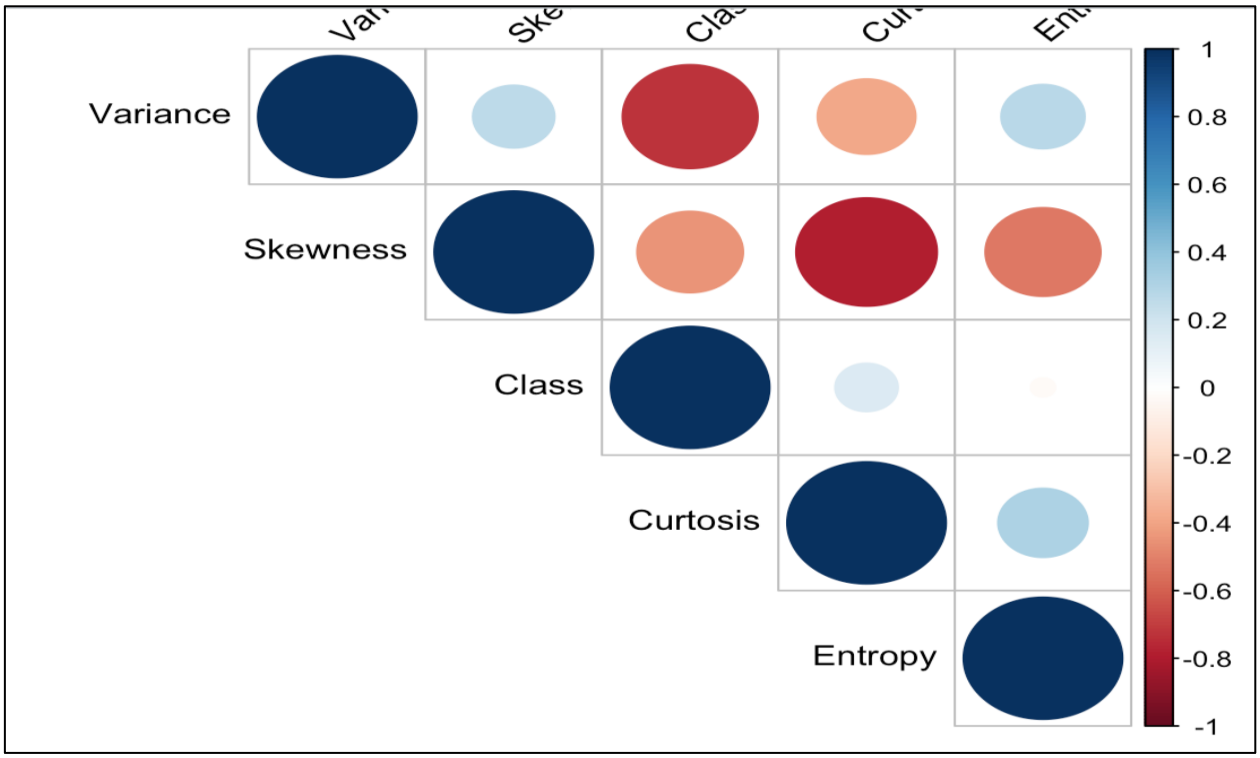

Fig. 5 *Correlation plot*

#From the plot, it is clear that there is zero correlation between the class attribute and Entropy. Therefore, the use of entropy in the classification, can be determined by classifying the dataset into training and testing data and computing the classification accuracy of the model with and without entropy.

```
> #Load the libraries to estimate the accuracy
> require(caret)
Loading required package: caret
Loading required package: lattice
Loading required package: ggplot2
>
> #Partition data into 80% (training) - 20%(testing)
> split=0.80
> trainIndex <- createDataPartition(mydata$Class, p=split, list=FALSE)
> mydata_train <- mydata[ trainIndex,]
> mydata_test <- mydata[-trainIndex,]
> #logistic regression with entropy
> model <- glm(Class ~.,family=binomial(link='logit'),data=mydata_train)
```

Fig.6 *Logistic regression model*

#The dataset is classified into 80 and 20%. 80% data has been used for training and 20% of data was used for testing.
#Since the datase has two possible outcomes, 0 or 1, logistic regression is appropriate.
#To perform logistic regression, glm() function is used.

```
> summary(model)

Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = mydata_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50383   0.00000   0.00000   0.00037   2.51400

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.8221     1.5513   4.398 1.09e-05 ***
Variance     -7.6291     1.8181  -4.196 2.71e-05 ***
Skewness     -3.8843     0.9241  -4.203 2.63e-05 ***
Curtosis     -4.9388     1.1799  -4.186 2.84e-05 ***
Entropy      -0.4719     0.3580  -1.318    0.188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1506.244  on 1097  degrees of freedom
Residual deviance:   40.075  on 1093  degrees of freedom
AIC: 50.075

Number of Fisher Scoring iterations: 12
```

Fig.7 *Model summary*

#From the model, we could see that Entropy has higher p value >0.05.
#It suggests that there is no strong association between entropy and class attribute.

```
> results <- predict(model,newdata=mydata_test,type='response')
> results <- ifelse(results > 0.5,1,0)
> #accuracy calculation
> error <- mean(results != mydata_test$Class)
> print(paste('Accuracy',1-error))
[1] "Accuracy 0.989051094890511"
```

Fig.8 *Predictive ability of model*

#From the accuracy obtained, it could be seen that the model is good for predicting the authenticity of bank notes.
#Again, a model is constructed with Entropy removed and the predictive ability of that model will be assessed.

```
> #Removing entropy
> mydata[4]<-NULL
> head(mydata)
  Variance Skewness Curtosis Class
1  3.62160   8.6661  -2.8073     0
2  4.54590   8.1674  -2.4586     0
3  3.86600  -2.6383   1.9242     0
4  3.45660   9.5228  -4.0112     0
5  0.32924  -4.4552   4.5718     0
6  4.36840   9.6718  -3.9606     0
>
> #logistic regression without entropy
> trainIndex <- createDataPartition(mydata$Class, p=split, list=FALSE)
> mynewdata_train <- mydata[ trainIndex,]
> mynewdata_test <- mydata[-trainIndex,]
>
> model <- glm(Class ~.,family=binomial(link='logit'),data=mynewdata_train)
```

Fig.9 *New model with Entropy removed*

```
> summary(model)

Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = mynewdata_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73794  -0.00001   0.00000   0.00156   2.32091

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.4494     1.4124   4.566 4.96e-06 ***
Variance     -6.1081     1.4710  -4.152 3.29e-05 ***
Skewness     -3.2003     0.7165  -4.467 7.94e-06 ***
Curtosis     -4.0327     0.9265  -4.352 1.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1507.660  on 1097  degrees of freedom
Residual deviance:   39.596  on 1094  degrees of freedom
AIC: 47.596

Number of Fisher Scoring iterations: 12
```

Fig.10 *Summary of new model*

```
> results <- predict(model,newdata=mynewdata_test,type='response')
> results <- ifelse(results > 0.5,1,0)
> error <- mean(results != mynewdata_test$Class)
> print(paste('Accuracy',1-error))
[1] "Accuracy 0.981751824817518"
```

Fig.11 *Predictive ability of new model*

#We see that there is not much difference in the accuracy obtained for the new model compared to the old model.
#Therefore, it could be concluded that, Variance, Skewness, Kurtosis together can effectively classify between genuine and forged notes.

Reference

http://archive.ics.uci.edu/ml/datasets.html

IHE-7510-01 - Data Mining- Lecture