MACHINE LEARNING FROM DATA
Fall 2018

Report: Lab Session 0 – Exploratory data analysis

Names: Santagiustina Francesco, Simonetto Adriano
Group:

1. Instructions

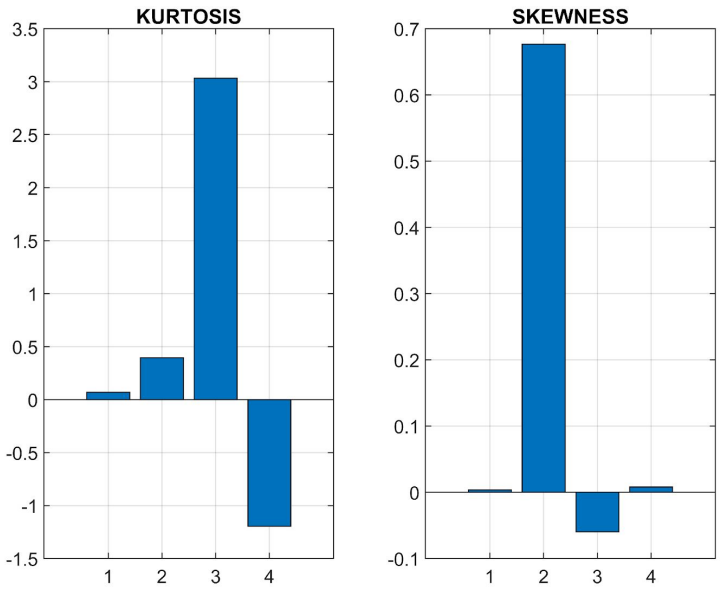
- Answer the questions
- Save the report, convert to pdf and upload the pdf file

2.

3. Questions

Q1. Briefly describe the conclusions of your analysis (you can insert plots)

As expected, observing the four different kind of plots, we can assume a Gaussian distribution only for the randomly generated Gaussian distribution.



In particular, from the analysis of kurtosis and skewness (shown above), it is possible to identify with sufficient confidence, the distribution that has originated the samples. The second one to begin with shows a particular skewness, which leads to believe to be the Rayleigh distribution as the other three are symmetric around their mean; the first one instead shows a kurtosis quite close to three, which suggests a Gaussian behaviour, while the one of the fourth is quite negative, a trait that between the four is only shown by the uniform. At last we can confidently say that the first one is actually the Laplacian as it is the only one left.

Q2. For each class and each feature, analyze histograms, cdfs and normal plots. Can we assume a Gaussian distribution for any of the features?

Amongst all the class-feature combinations the one that better follows a Gaussian behaviour is c2-f1. Its normal plot follows more or less the straight line of the Gaussian, its cdf is sufficiently close to the Gaussian one and the same can be said about its histogram. Other combinations with decent results are for example c1-f1 (whose normal plot however tends to slip a bit from the Gaussian one) and c3-f2 (its main issue is the histogram that presents a hole in the middle).

In the end however it is hard to confidently assume a Gaussian behaviour for any of the class-feature combinations as the amount of data appears to be too small for the tools to be accurate enough.

Q3. Analyze kurtosis and skewness values for each feature and class.

1. Sepal length:

Kurtosis is inferior to 3 for the three classes but at different values (about 2.65, 2.40 and 2.91 respectively) the last one being suitable to result from a gaussian.

Skewness of sepal length distribution is slightly positive and similar for all the classes (about 0.12, 0.10 and 0.11 respectively).

2. Sepal width:

No general consideration can be done about the kurtosis of this feature over the three classes as values vary widely (about 3.69, 2.55 and 3.51 respectively), we can see that the ones of class 1 and 3 are similar and denote more extreme deviations in sepal length for these classes.

For class one skewness related to sepal width (about 0.10) is very similar to the one related to sepal length while we see that class 2 and 3 skewness have both a moderate absolute value but different sign (respectively -0.35 and 0.35).

3. Petal length:

For petal length kurtosis is pretty high for the first class and close to the gaussian reference for classes 2 and 3 (about 3.81, 2.93 and 2.74 respectively).

Skewness is almost zero (about 0.07) for class one, while the same scheme as before repeat for skewness of class 2 and 3 petal length: quite high but opposite values (about -0.59 and 0.53).

4. Petal width:

Kurtosis values (about 4.30, 2.51 and 2.33 respectively) show us that petal length distribution is very outlier-prone for class 1 while classes 2 and 3 show frequent modestly sized deviations.

For petal width skewness is very low for classes 1 and 2 (about -0.03 and -0.13) but very large and positive for class 1 (about 1.16).

Q4. Analyze boxplots by feature. Are there 'significant' differences between the classes?

The classes do present significant differences. With the exception of feature 2 where the median values and quartiles of the three classes appear to be similar, it is possible to observe a clear trend in the other three boxplots. In fact class 1 appears to have a median value much lower than that of classes 2 and 3 up to the point that for features 3 ad 4 even the 95% confidence interval results disjoint from that of classes 2 and 3. This is a sign of significant difference between class 1 and the other two. Moreover we can spot differences also between classes 2 and 3 as median and quartiles values of class 3 appear to be higher than those of class 2 for all of the features.

Q5. Analyze the scatter plot. Are features related in any way? What can you say about the separability of the classes?

There appears to be a strong relation between some of the features (e.g. the scatter plot of features 3 and 4 presents a clear tendency along the main diagonal, impling a relation of direct proportionality between the two). As for the classes' separability we can say that class 1 seems to be easily distinguishable from the other two as in almost every scatter plot it is clearly divided from the others. On the other hand, even if we can still see some differences between the behaviour of class 2 and 3 it appears way harder to separate the two as in almost every plot their respective points appear to be mixed together. In particular, it is possible to choose features 3 and 4 in order to obtain a good separation between the three classes using linear classifiers.

Q6. Edit the script ML_Lab0_irisdataset.m. Choose one feature (among the four available) and compute the feature mean and confidence intervals at confidence levels 95%, 99% and 99.9% for the three classes.

Hint: use Matlab functions `tinv` and `var`

Feature #3	Mean	CI at 95%	CI at 99%	CI at 99,9%
Class 1	1.4640	[1.4147, 1.5133]	[1.3982, 1.5298]	[1.3781, 1.5499]
Class 2	4.2600	[4.1265, 4.3935]	[4.0819, 4.4381]	[4.0274, 4.4926]
Class 3	5.5520	[5.3952, 5.7088]	[5.3428, 5.7612]	[5.2788, 5.8252]

Q7. Copy the code used to answer Q6.

```
i_feat = 3;
alphas=[0.05, 0.01, 0.001];
```

```
for j = 1:length(alphas)
        alfa=alphas(j);
        for i_class = 1:N_class
        index=find(Labels==i_class);
        df=length(index)-1;
        M_mean=mean(X(index,i_feat));
        S_deviation=sqrt(var(X(index,i_feat)));
        P=1-(alfa/2);
        t_alfa_2=tinv(P,df);

        Confidence_I=[M_mean-(t_alfa_2*S_deviation/sqrt(df+1));M_mean+(t_al
        fa_2*S_deviation/sqrt(df+1))]
        end
end
```

Q8. Choose one feature K (among the four available). Edit the script ML_Lab0_irisdataset.m to conduct the following hypothesis tests, using a chi-squared test

- Null hypothesis H_0 : Feature K from class 1 comes from a Gaussian distribution at the significance level 0.001
- Null hypothesis H_0 : Feature K from class 2 comes from a Gaussian distribution at the significance level 0.001
- Null hypothesis H_0 : Feature K from class 3 comes from a Gaussian distribution at the significance level 0.001

Complete the following table with the decisions (acceptance/rejection) for the null hypothesis H_0 (feature Gaussianity), p-value and degrees of freedom for α = 0,001.

Explain the meaning of the p-value and interpret the results accordingly.

Feature # 3	Acceptance / rejection of H_0	*p*-value	Degrees of freedom
class 1	H_0 accepted	0.2645	3
class 2	H_0 accepted	0.4050	3
class 3	H_0 accepted	0.0381	4

Given a hypothesis H0 supposed true, the p-value is the probability of obtaining a result equal or more extreme than the observation.With this in mind we cannot reject any H0 as all the p-values are much higher than the set value of alpha.

Q9. Copy the code used to answer Q8.

```
i_feat = 3;
P = zeros(1,3);
H = zeros(1,3);
 for i_class=1:N_class
      index=find(Labels==i_class);
      V=X(index,i_feat);
      [H(i_class),P(i_class),STATS] = chi2gof(V,'ALPHA',0.01,'nbins' ,10)
 end
```