Pattern Recognition Coursework 2

Jakub Mateusz Szypicyn
CID: 00846006
EEE4
jms13@ic.ac.uk

Jacobus Jukka Hertzog
CID: 00828711
EEE4
jjh13@ic.ac.uk

Abstract

Line 1

Line 2

Line 3

1. Introduction

2. Distance Metrics

2.1. Distribution of Wine Data

The wine data (`wine.csv`) provided consisted of three classes. It cis made up of 13 features representing chemical and optical information. The data can be analysed *talis qualis*, however a more elaborate method is to investigate the covariance matrices. The raw data can supply us with information such as feature distribution (mean or spread), whereas covariance matrices will provide the dependency between features. This can be performed for all data altogether or for each class individually. Both methods will result in meaningful information as described in sections 2.1.1 and 2.1.2. Furthermore the data was then normalised and the above was repeated.[1]

2.1.1 All Classes

By investigating the data from all classes we will see that data varies altogether. We can investigate each feature and see what part of space it occupies. We can also determine how strongly all features are related to one another for all classes. Let us begin by examining the distributions of the raw data. The means of the feature distributions vary from around 0.36 for nonflavanoid phenols, through values of 1 to 3 and 10 to 100 to a single extreme at 737.1 for proline. The corresponding deviations are also quite spread out. The general trend is that the larger the mean the larger the standard deviation of the distribution. For example proline distribution has a standard deviation of 287 over 118 samples. The distributions of the features don't always seem to follow bell-shaped Gaussian distribution. However it should be noted that each distribution is in fact a sum of three independent distributions. Distribution of proline in Figure 1 resembles a gamma distribution.

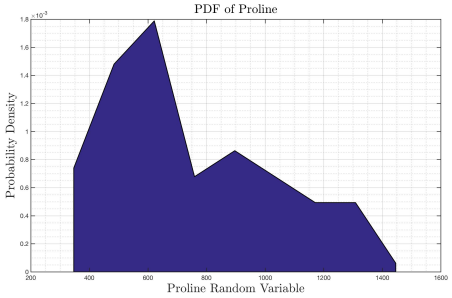

Figure 1. Distribution of Proline from all Classes

Table 1. Sample Means and Standard Dev. for all Training Data

RV	Alcohol	Ash	Hue	Proline
Mean	12.7	2.38	0.95	737
Std.	0.72	0.29	0.24	287

Looking at the covariance matrix of the raw training data we can see that proline covary with all other features quite strongly. This is however due to the fact that proline's distribution has a naturally high variance, thus corrupting the readings. The highest covariance for two different features excluding proline is that of magnesium (feature 5) and colour intensity (feature 10). However, these two features have second and third highest variance distributions. We have to therefore look at the covariance matrix of the normalised feature vectors.

Highest positive covariance is observed for features 6 and 7 (total phenols and flavanoids), whereas the most nega-

[1] Note that data was first separated as per instruction. Only training data is now being considered.

tive covariance is found for features 2 and 7 (malic acid and flavanoids). The plot of the above cases is shown in Figure 2.

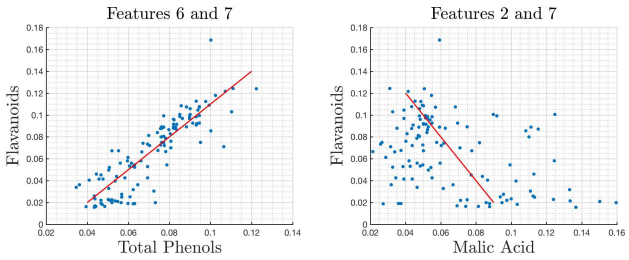

Figure 2. Covariance of Features 6 & 7 and 2 & 7 Visualised

2.1.2 Individual Classes

2.2. L1 and L2 Metrics

2.3. Chi-Squared and Correlation Metrics

2.4. Histogram Intersection Metric

2.5. Mahalanobis Distance Metric

3. K-means Clustering

4. Neural Network

5. Conclusion