Technical University of Denmark

02450
Introduction to Machine Learning and Data Mining

Assignment 1

Work Distribution

Celia Sullca Cailloux (s132534) **50%**

Oseze Esther Iyore (s121922) **50%**

Technical University of Denmark

Febuary 27th, 2018

Contents

Data Description

Introduction

In this section a brief description of the data set used in this report, will be given. The name of the data set is "Breast Cancer Wisconsin (Diagnostic)" and is obtained from UCI Machine Learning Repository. As seen specifically in the two articles W. Nick Street, William H. Wolberg "Breast Cancer Diagnosis and Prognosis Via Linear Programming" and Kristin P. Bennett, Ayhan Demiriz, Richard Maclin "Exploiting unlabeled data in ensemble methods", the data is used to: "increase the accuracy and objectivity of breast cancer diagnosis and prognosis. The first application to breast cancer diagnosis utilizes characteristics of individual cells, obtained from a minimally invasive fine needle aspirate, to discriminate benign from malignant breast lumps. This allows an accurate diagnosis without the need for a surgical biopsy"[1].

The data contains 569 instances and 32 attributes, including the ID number and classification of the diagnosis (malignant M or benign B). The remaining 30 attributes are described by 10 characteristics of the cell nuclei in a breast mass. Each characteristic is represented by three attributes: 1) the mean, 2) the standard error and 3) the largest/worst value. It is stated with the data that there are no missing values in the data set. Table 1 shows a description and the data type of each attribute.

Primary Machine Learning Modeling Aim

When carrying out classification or clustering the relevant attribute to predict is the B/M diagnosis since this is the only discrete attribute. The aim would be to predict the diagnosis based on all the other attributes. In the case of classification this would be done by first training on part of the given data, with the correct diagnosis given, hence supervised machine learning. Finally, the model would be tested on the rest of the data. For the case of clustering the model is unsupervised which means that the training on a part of the data is done without the correct B/M diagnosis given. Thus, the main attributes for clustering are yet to be discovered. However, the overall tendency for malignant cell nuclei are that they are larger. From this point of view, the size-related attributes (mean radius, mean area and mean perimeter) are possibly suitable for clustering.

For the regression analysis the modelled attribute should be continuous. As mentioned before, the size of the cell nuclei are usually strong indicators of benign and malignant cases, thus the size-related attributes (mean radius, mean area and mean perimeter) could be predicted based on other the deformity-related attributes. When carrying out association mining the aim is to find a relationship between some of the attributes. In this case we expect certain associations between mean radius, mean perimeter, mean area and compactness, given that these are all related by mathematical formulas. In general, the comparing the size-related attributes with the deformity-related attributes is interesting, given that the size may reflect the state of the cell nucleus. When doing an anomaly detection the goal is to find and regard outliers in the data, and try to learn from these. For this purpose the same mathematically correlated attributes (mean radius, mean perimeter and mean area) are self-evident.

[1] W. Nick Street, William H. Wolberg "Breast Cancer Diagnosis and Prognosis Via Linear Programming"

Attribute Description

Table 1 summarizes the data type for the first ten attributes. As previously stated, the remaining twenty attributes are the standard deviation (STD) and largest/worst values of the same ten attributes listed in Table 1. It is seen that most of the attributes are ratios and continuous. The basic

Attribute	**Description**	**Type**	**Disc./Cont.**
Benign/Malignant	Diagnosis	Ordinal	Discrete
Mean radius (μm)	Mean of distances from center to points on the perimeter	Ratio	Continuous
Mean texture	Standard deviation of gray-scale values	Ratio	Continuous
Mean Perimeter (μm)	Perimeter of the cell nuclei	Ratio	Continuous
Mean Area (μm^2)	Surface area of breast mass	Ratio	Continuous
Mean Smoothness	Local variation in radius lengths	Interval	Continuous
Mean Compactness	Perimeter^2 / area - 1.0	Ratio	Continuous
Mean Concavity	Severity of concave portions of the contour	Ratio	Continuous
Mean Concave Points	Number of concave portions of the contour	Ratio	Continuous
Mean Symmetry	Length difference between the lines perpendicular to major axis	Ratio	Continuous
Mean Fractal Dimension	"Coastline approximation" - 1	Ratio	Continuous

Table 1: Attributes description and data type.

data statistic is performed to get an overview of all the attributes of the data set. For this part the data is separated into two parts based on the two diagnosis, benign or malignant. The following figures (Fig. 1 and 2) show the box plot of each of the diagnosis for each of the ten first attributes after the diagnosis.

The attributes are standardized to allow comparison and also for the coming PCA analysis where standardized values are needed. For the benign diagnosis the attribute *mean concavity* stands out as it has the most distance outlier. For the malignant diagnosis this is true for mean texture and fractal dimension. In general for both diagnosis there are a lot of outliers. For benign (B) diagnosis the attributes mean texture, mean concave point and mean fractal dimension have the most outliers. For malignant (M) diagnosis meant compactness and mean concave point have the most. This could be an indication that these attributes are not the most defining attributes when trying to diagnose.

Table 2 shows the mean, standard deviation and range for the ten first attributes after for each diagnosis (M/B). In general, the tendency is that the mean values of the attributes increase for the malignant cell nuclei in comparison to the benign cell nuclei. This is the case both he attributes describing the size and deformity of the cell nuclei. However, the standard deviation also increases for the attributes belonging to malignant cell nuclei. Thus the range of the attributes is more or

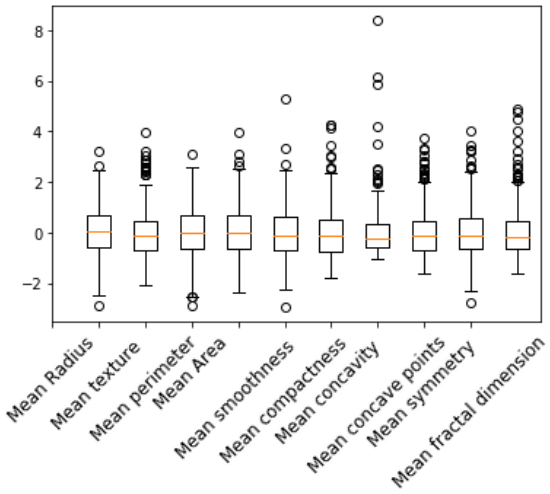

Figure 1: Box plot for the ten first attributes for the B diagnosis. The data is standardized.

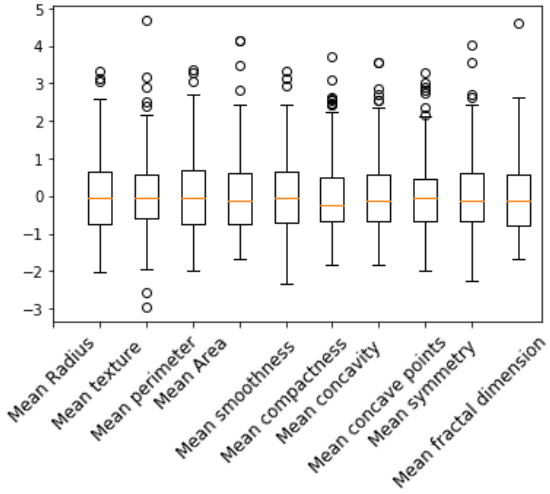

Figure 2: Box plot for the ten first attributes for the M diagnosis. The data is standardized.

less the same for both malignant and benign cell nuclei apart from the attributes describing the size (mean radius, mean perimeter and mean area). The latter ones have both lower mean values, smaller standard deviation and range that shifted towards lower values in the case of benign diagnosis compared to malignant diagnosis.

Attribute	**Mean (M)**	**STD (M)**	**Range (M)**	**Mean (B)**	**STD (B)**	**Range (B)**
Mean radius	17.5	3.20	10.95 - 28.11	12.1	1.78	6.98 - 17.9
Mean texture	21.6	3.77	10.38 - 39.28	17.9	3.99	9.7 - 33.8
Mean perimeter	115.0	21.8	71.9 - 188.5	78.1	11.8	43.8 - 114.6
Mean area	978.0	367.0	361.6 - 2501.0	463.0	134.0	143.5 - 992.1
Mean smoothness	0.10	0.01	0.074 - 0.14	0.093	0.013	0.053 - 0.16
Mean compactness	0.15	0.0539	0.046 - 0.35	0.08	0.03	0.019 - 0.22
Mean concavity	0.161	0.0748	0.02 - 0.43	0.046	0.04	0.0 - 0.41
Mean concave points	0.088	0.0343	0.02 - 0.20	0.026	0.02	0.0 - 0.085
Mean symmetry	0.19	0.03	0.13 - 0.3	0.17	0.02	0.106 - 0.27
Fractal Dimension	0.06	0.008	0.05 - 0.097	0.06	0.01	0.052 - 0.096

Table 2: Mean, standard deviation and range for the ten different attributes for the two diagnosis.

Visualization

In this section the data will be visualized graphically. Firstly the histogram of the ten first attributes after the diagnosis are regarded. These are shown in figure 3 and show that the attributes generally seem to be normally distributed with the exception of the mean concavity and mean concave points that seem to have an exponentially decaying character. In figure 4 five of the attributes are plotted

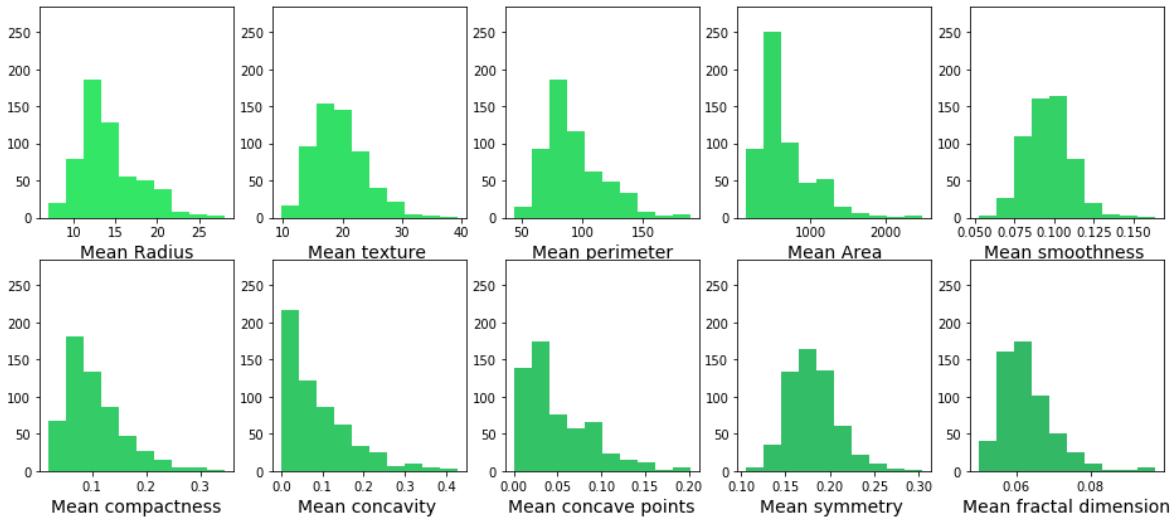

Figure 3: Histogram of the first ten attributes (listed in Table 2)

against each other, of which four are size-related attributes. The diagnosis indicated with color. The figure shows the correlation between the selected attributes. The strongest correlation observed are between the size-related attributes. All the scatter plots show a positive linear correlation reflecting that the larger the deformity and the size of the cell nuclei, the more probable the diagnosis of a malignant cell nuclei. In the previous section it was discussed to use all the parameters to predict the diagnosis class. This figure shows a clear separation of the two diagnosis (the color separation) which indicates that all of these attributes could in principle be used to predict diagnosis. It was also discussed to use mean radius to model for the regression. The scatter plots shows that radius is strongly correlated to two other attributes, so this confirms that this analysis could be of interest to conduct. Finally it was discussed to use the mathematically correlated attributes for association mining. The strong correlating between mean perimeter and mean area confirm that this analysis would make sense to perform.

The main goal of principle component analysis (PCA) is to reduce the amount of dimensions in a multidimensional data set without loss of any signification information. The idea is thus to minimize abundance and maximize signal. When the PCA is performed the data may be reduced to a dimension that allows it to be visualized clearly. It is therefore a very useful tool when regarding data attributes. The principle components are the new basis vector created by the analysis, the point in the direction where the most variance is explained in order. So the first principle component explains the most variance and so on. When performing a principle component analysis the components used must account for 90 % of the variance. In figure 5 the amount of variance explained is shown as a function of the number of principle components used. Outputting the numbers it shows that the first component explains 98 % of the variance and the two first component together explain

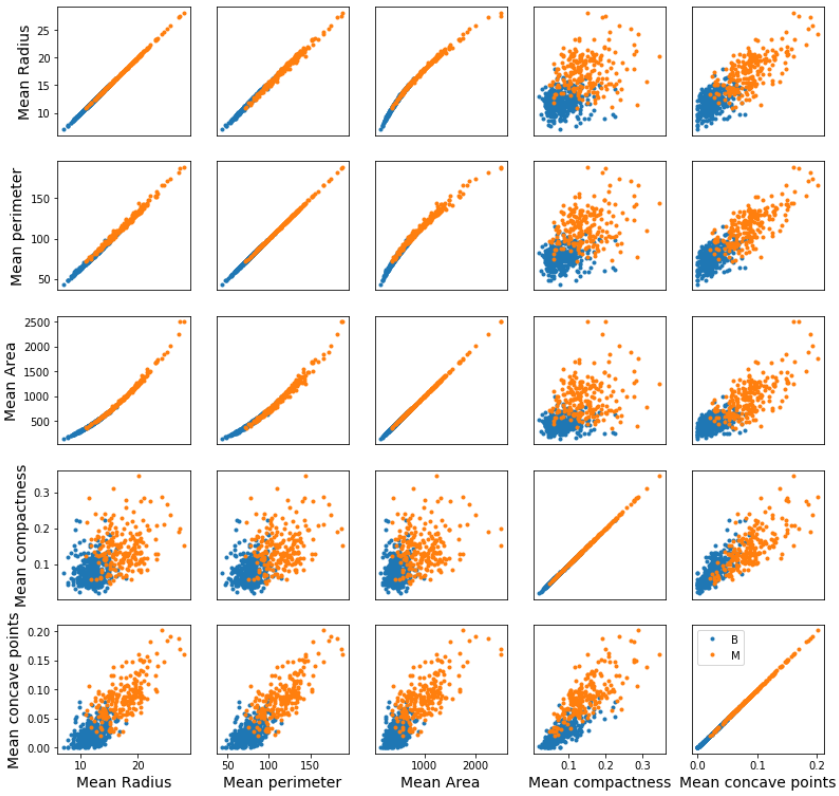

Figure 4: Scatter plot of selected attributes plotted against each other.

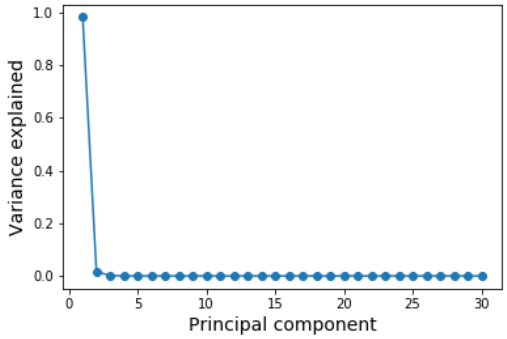

Figure 5: The amount of variation explained as a function of the number of PC included.

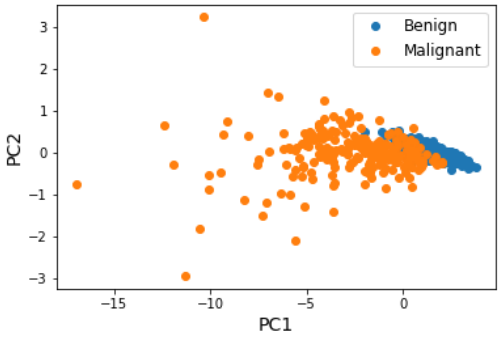

Figure 6: The data projected onto the two principal components axis

99.8 % of the variance. This means these two axes are sufficient to describe the data without much loss of information. Figure 6 shows the data projected on the two principle components. The data belonging to the benign diagnosis are clear on one place on the figure very consistently where the data for the malignant diagnosis is slightly more spread out and single observations lie far away from the rest.

Brief Discussion

Through this report the data for breast cancer diagnostics has been visualized and analyzed. It was briefly discussed that for supervised learning, the the class attribute (benign/malignant) is the relevant attribute to compare with for classification. However, for regression and unsupervised learning, the size-related attributes (mean radius, mean perimeter, mean area and compactness) may enable the desired regression or clustering of benign and malignant cell nuclei. The summary statistics showed a tendency that higher mean values of all the attributes occur for malignant cell nuclei. In addition, the range of the size-related attributes shifted positively with malignant cell nuclei. The visualization of the attributes showed that most attributes where normally distributed. The strongest correlations between the attributes where between the size-related attributes and in general a positive correlation between the attributes. By principal component analysis the variance explained for the two first principal components account for 98 % of the variance of the data, why the data "Breast Cancer Wisconsin" can be fairly represented by the first two principle components.