DATA SCIENCE TECHNIQUES AND APPLICATIONS

Coursework 1: Dataset dimensionality analysis

Eirini Zygoura
Student ID: 13177951

Academic Declaration: “I have read and understood the sections of plagiarism in the College Policy on assessment offences and confirm that the work is my own, with the work of others clearly acknowledged. I give my permission to submit my report to the plagiarism testing database that the College is using and test it using plagiarism detection software, search engines or meta-searching software.”

Phase 1: Selection of the dataset

My first intention was to chose a dataset related to earth sciences or weather forecasting as I already have a MSc degree in Environmental Physics. Mining in Kaggle for datasets I decided to work with Breast Cancer Diagnosis, which is also a topic that I am interested in after being diagnosed almost 5 years ago.

The name of the chosen dataset is “Breast Cancer Wisconsin - benign or malignant” [1] and the goal is to predict whether a cancer cell is benign or malignant. The set consist of numerical features and is appropriate for the requirements of the coursework.

Phase 2: Description of the data

Origin of the data

The creator of the dataset is Dr. William H. Wolberg (physician), University of Wisconsin Hospitals (USA). The dataset Breast Cancer Wisconsin (Original) is taken from UCI Machine Learning Repository [2].

Samples of the dataset arrived periodically as Dr. Wolberg reported his clinical cases. As a result, the database reflects chronological grouping (8 groups from January 1989 to November 1991 with different number of instances). The dataset consist of 9 cytological characteristics of breast fine-needle aspirates (FNAs) that have been established to differ significantly between benign and malignant samples, as mentioned by Wolberg and Mangasarian [3]. It should be noted that no single characteristic alone or presently described pattern distinguishes between benign and malignant samples.

Kaggle challenges and discussion

There are no Kaggle challenges proposed for this dataset. A light guidance for beginners on processing the data and methods to evaluate their characteristics is provided. Also, there are posted recommendations dealing with classification methods that can be used to predict the type of the cancer cell, such as Logistic Regression and Random Forest Classification. In another version of the same data provided by Kaggle (version 1) [4], the use of Deep Neural Networks is presented.

Existing literature

For this dataset two papers are consider relevant: Wolberg and Mangasarian (1990) [3] and Zhang (1992) [5]. Up to now 40 papers cite this dataset, as mentioned in UCI Machine Learning Repository. Due to the automatic harvest of the papers that cite the dataset we cannot assume that all of these actually make use of the data, rather than mentioning them, but we can make an assumption that the majority of them are based on that dataset in order to investigate breast cancer.

Description of the data

Breast Cancer Wisconsin dataset has 683 observations. There are 10 input variables and 1 target variable. The first row of the dataset gives the variable names. The attributes on each column are:

1. Sample code number: id number
2. Clump Thickness
3. Uniformity of Cell Size
4. Uniformity of Cell Shape
5. Marginal Adhesion
6. Single Epithelial Cell Size
7. Bare Nuclei
8. Bland Chromatin
9. Normal Nucleoli
10. Mitoses
11. Class

The first column is an identifier of the sample and for further analysis it should be dropped from the data set since it does't contain any valuable information. The columns 2-10 (9 features) are the actual predictor variables that can be used to predict the class of the cancer cell. Each predictor has integer values scaled from 1 to 10, with 1 being the closest to benign and 10 the most anaplastic [3]. The class of each observation is listed in column 11 and takes the value 2 for benign cancer cells and value 4 for malignant.

Phase 3: Dimensional analysis

Size and shape of the data

Ignoring the first column (id number), it is a 10 dimensional dataset with 683 observations of 10 features.

Data quality

Breast cancer Wisconsin (version 3) is a high quality data with no missing values. The only disadvantage is that there are nearly double the number of observations with class 2 (benign) than there are with class 4 (malignant). Specifically:

- Class 2: 444 observations
- Class 4: 239 observations

Medically, it is useful to know the percentage of a fine-needle aspirate sample of being malignant (35% for this dataset), but in classification algorithms when the data is skewed in favour of one class, problems in training may be caused because the majority class can dominate and the classifier may distinguish in favour of that class.

To overcome the imbalance problem there are some techniques that convert the dataset into balanced, such as over-sampling (incrementing the size of the minority class) or under-sampling (reducing the size of the majority class). Over-sampling is prone to overfitting and in under-sampling critical information may be lost by decreasing the amount of the training data.

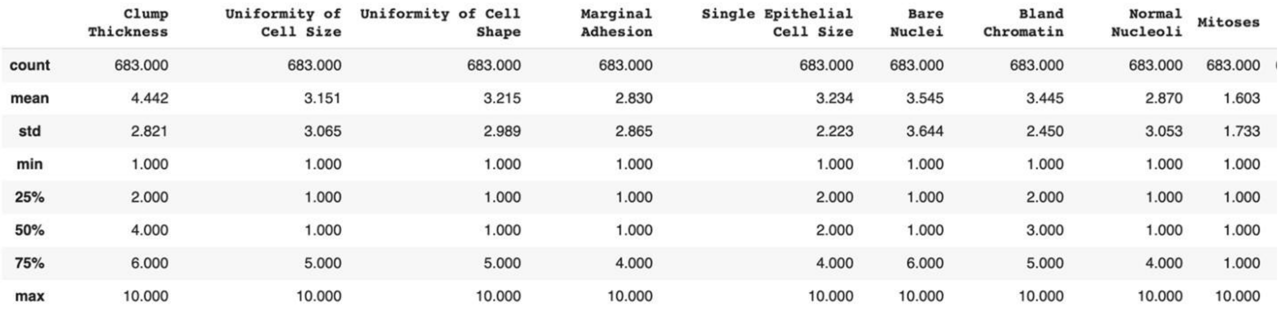
Also, feature selection is a function of intelligent subsampling and potentially helps reduce the imbalance problem. The scope is to identify the significant features for each class and take the union of the features to obtain the final feature set.

The right use of evaluation metrics is critical to ensure that a relatively small imbalance ratio does not affect the model performance.

In our dataset the imbalance is not of extremely high degree and there may not be a problem. For this part of the coursework I have decided to overlook the imbalance problem. Feature selection and correlations between features will be considered.

Feature characteristics

A brief description of the predictor features is provided in Figure 3, where the most important descriptive statistics are provided. Histograms in Figure 4 give a visualised description on how the data values are distributed.



	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
count	683.000	683.000	683.000	683.000	683.000	683.000	683.000	683.000	683.000
mean	4.442	3.151	3.215	2.830	3.234	3.545	3.445	2.870	1.603
std	2.821	3.065	2.989	2.865	2.223	3.644	2.450	3.053	1.733
min	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
25%	2.000	1.000	1.000	1.000	2.000	1.000	2.000	1.000	1.000
50%	4.000	1.000	1.000	1.000	2.000	1.000	3.000	1.000	1.000
75%	6.000	5.000	5.000	4.000	4.000	6.000	5.000	4.000	1.000
max	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000	10.000

Figure 3

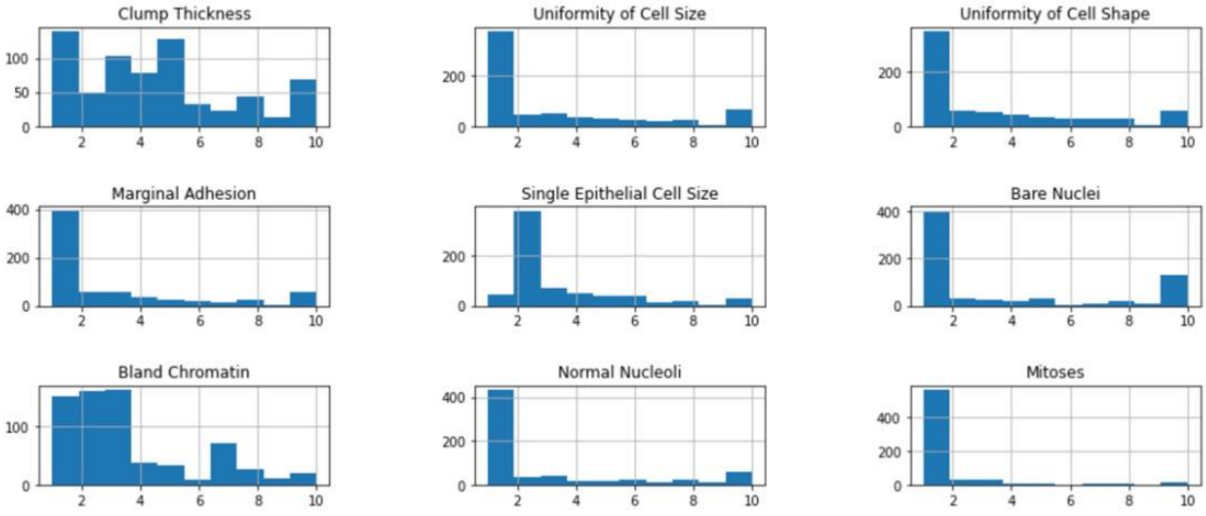
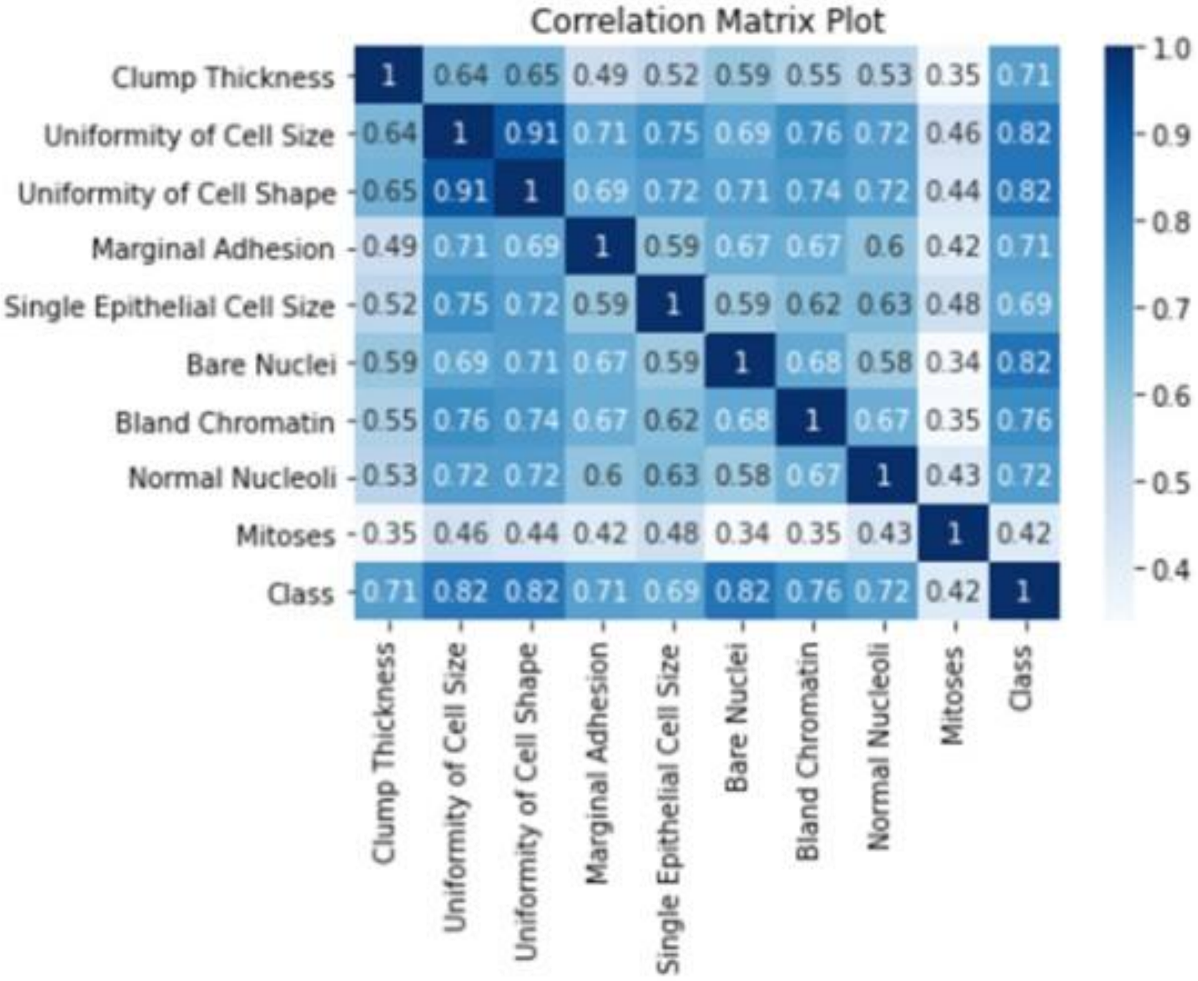


Figure 4

A skewness on the lower values is apparent to all figures. It can be a reflection of the imbalance of the data towards class 2, as benign cells have lower values.

Correlation between features

For the whole set of data (2 classes contained) the correlations between the features have been calculated and are presented in Figure 1.



	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
Clump Thickness	1	0.64	0.65	0.49	0.52	0.59	0.55	0.53	0.35	0.71
Uniformity of Cell Size	0.64	1	0.91	0.71	0.75	0.69	0.76	0.72	0.46	0.82
Uniformity of Cell Shape	0.65	0.91	1	0.69	0.72	0.71	0.74	0.72	0.44	0.82
Marginal Adhesion	0.49	0.71	0.69	1	0.59	0.67	0.67	0.6	0.42	0.71
Single Epithelial Cell Size	0.52	0.75	0.72	0.59	1	0.59	0.62	0.63	0.48	0.69
Bare Nuclei	0.59	0.69	0.71	0.67	0.59	1	0.68	0.58	0.34	0.82
Bland Chromatin	0.55	0.76	0.74	0.67	0.62	0.68	1	0.67	0.35	0.76
Normal Nucleoli	0.53	0.72	0.72	0.6	0.63	0.58	0.67	1	0.43	0.72
Mitoses	0.35	0.46	0.44	0.42	0.48	0.34	0.35	0.43	1	0.42
Class	0.71	0.82	0.82	0.71	0.69	0.82	0.76	0.72	0.42	1

Figure 1

The features that are highly correlated with the target class (with correlations equal to 0.82) are:

- Uniformity of cell size,
- Uniformity of cell shape and
- Bare nuclei

Among them, Uniformity of cell size and Uniformity of cell shape are highly correlated with each other, a fact that can reduce the performance of some machine learning algorithms such as Logistic Regression.

It is worth mentioning that Mitoses shows very poor correlation, not only with the target class, but with all the other features as well. A non-linear relationship for Mitoses and the other features may exist.

Feature selection

In order to take account of the non-linear relationships that may exist, the predictor variables are ranked using Feature Importance obtained by Extra Trees Classifier. In the top 3 of the most important predictors are the 3 features with the highest correlation with the target class. The new information is the rank of importance of each feature. Feature Importance rankings are presented in Figure 2.

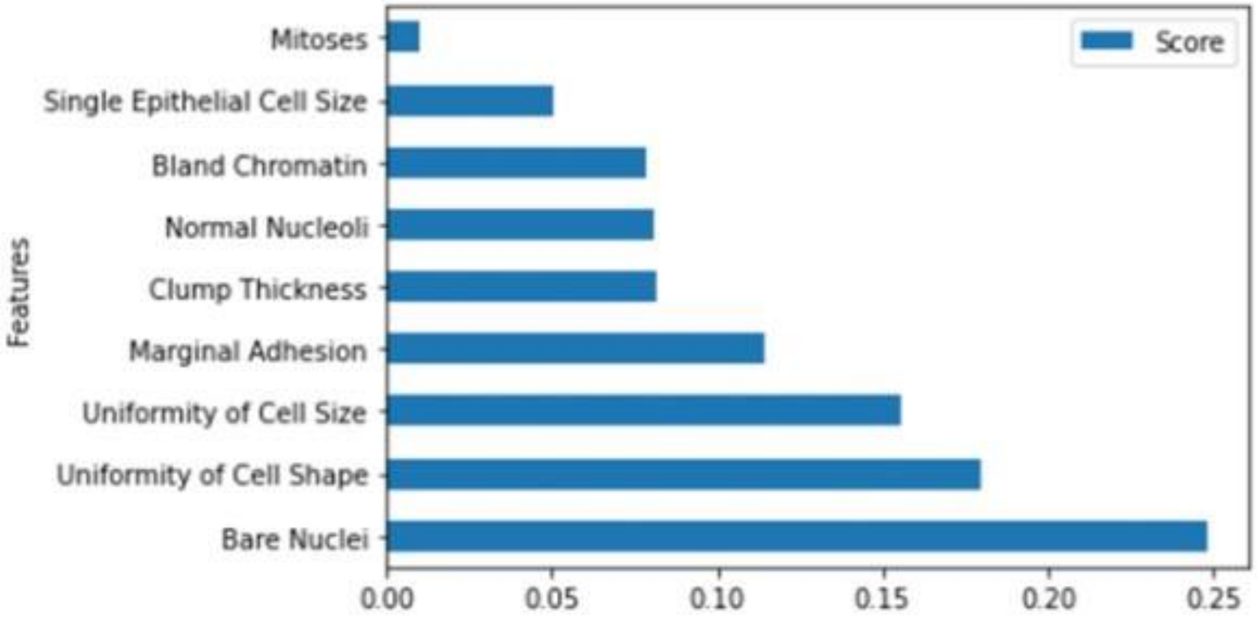


Figure 2

Marginal Adhesion ranked 4th, followed by Clump Thickness, Normal Nucleoli and Bland Chromatin, having the same score. Due to the closeness of the scores perhaps it should be better to consider more than the 3 variables that have the highest correlation with the target for classification modelling. Personally, I would choose to continue with the 4 higher scored features (Bare Nuclei, Uniformity of Cell Shape, Uniformity of Cell Size and Marginal Adhesion) as a start and then compare the model results choosing the 7 higher score features to see if the extra features add noise to the model due to the correlation between them, or if they improve the model predictions.

Mitoses has a very low ranking indicating that there is no specific relationship between this variable and the type of the cancer cell.

Conclusions

A. Proposed dimensions for further analysis are (in descending relevance order):
   - Bare Nuclei,
   - Uniformity of Cell Shape,
   - Uniformity of Cell Size and
   - Marginal Adhesion

B. Balance techniques can be applied to examine wether they improve the performance of a classification model.

C. More feature selection techniques may be considered and the number of selected features should also be examined.

REFERENCES

1. Kaggle Breast Cancer Wisconsin dataset, available online at: *https://www.kaggle.com/ninjacoding/breast-cancer-wisconsin-benign-or-malignant*

2. UCI web page on Breast Cancer Wisconsin (Original) Data Set, available online at: *https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)*

3. Wolberg, W.H., Mangasarian, O.L. (1990), *"Multisurface method of pattern separation for medical diagnosis applied to breast cytology"*, Proceedings of the National Academy of Sciences of the United States of America, 87 (23), pp. 9193-9196. doi:10.1073 pnas.87.23.9193

4. Kaggle Breast Cancer Wisconsin dataset (version 1), available online at: *https://www.kaggle.com/salihacur/breastcancerwisconsin*

5. Zhang, J. (1992), “Selecting typical instances in instance-based learning.”, Proceedings of the Ninth International Machine Learning Conference (pp. 470-479). Aberdeen, Scotland: Morgan Kaufmann.