Data Mining I: First Assignment

Deadline: November 23rd 2018

September 24th, 2018

The main goal of this work is to apply feature importance methods to rank variables that are more predictive of cardiac pathologies. Before trying to achieve this goal, it may be necessary to preprocess the data. For example, data cleaning, data transformation, data normalization, as well as removal of irrelevant features (such as the ID), will be needed to help improving data quality.

In data cleaning, you need to look for **inconsistencies**, **noise** and **missing values**. For example, you may look for different variable values that convey the same meaning (M or Masculine, where both mean the Masculine gender) or variables that contain spurious values (for example, impossible values for ages, dates etc), you may need to remove duplicate entries/objects, remove variables that contain too many missing values (for example, more than 50% missing values, if missing values have no meaning), or remove variables that have the same value for all objects.

Data transformation involves applying functions that will convert variables or objects from one space to another. Methods commonly used are smoothing (binning, clustering or regression, for example), aggregation, generalization (for example, to transform low level feature values into higher level feature values - age represented in higher level intervals), normalization (for example, $x_{new} = \frac{x-x_{min}}{x_{max}-x_{min}}$) and standardization $x_{new} = \frac{x-\mu}{\tau}$ to make data fall into a small specified range, or feature construction (for example, creation of new attributes such as mass body index from height and weight). Be careful with outliers. If your data has too many outliers, normalization may push all your "normal" values to a very small interval. In that case, it is recommended to use standardization.

After preprocessing, it is usual to proceed with univariate analysis (summaries, histograms, boxplots, mean, standard deviation, range percentiles, interquartile range etc to measure data spread, centrality etc), bivariate analysis (correlations, for example Pearson, Spearman, Kendall etc), and mutivariate analysis (regression, clustering, factoring-PCA, mutual information etc). These are useful to study relevant or redundant variables and help to perform a pre-selection of features. Note that strong correlations among variables mean that they are dependent in some way. When performing feature selection, usually (but not always, because it depends on the domain), if two or more variables are strongly correlated, it may mean that they equally contribute to the analysis and some of them may be discarded. For example, if height and weight are strongly correlated with IMC, we should use only IMC. Methods of **multiple mutual information** perform feature selection by choosing only relevant and non-redundant variables.

Besides performing statistical analysis, we can also apply machine learning algorithms to extract **knowledge** from data and generate predictive models. Examples of such algorithms are Support Vector Machines, Decision Trees, Random Forests, Bayesian Networks, Ensemble methods, Neural Networks etc.

There are several different libraries available out there that can perform all those tasks. Be aware that algorithms offer a wide range of parameter choices. You need to be careful with that choice, and need to understand what algorithm is implemented and what the results mean.

In order to assess the performance (quality-wise or quantity-wise) of each method on the same dataset, we need to choose some evaluation metric. This can be the error rate, the rate of correctly classified instances (CCI), sensitivity (Recall, True Positive Rate), specificity, precision, Receiver Operating Characteristic (ROC) curves, Area Under the (ROC) Curve, Precision-Recall curves etc. The metric to be used to assess performance will usually depend on the domain in hand. For example, if data is skewed (unbalanced number of objects per class), error rate may not be the adequate choice (why?).

Data Dictionary The data to be used for this work was collected in the Real Hospital Português (RHP), Brazil, anonymized and shipped to Portugal with the approval of the RHP Ethics Commitee. The Ethics Committee of the University of Porto, Portugal, also approved the use of this data for academic studies. Entries consist of children between 0 and 19 years old, with or without a cardiac pathology (variable NORMAL X ANORMAL).

Table 1 gives a brief description of the variables.

ID	anonymized patient ID
Peso	patient weight
Altura	patient height
IMC	body mass index
Atendimento	date of visit
DN	birth date
IDADE	age
Convenio	health care insurance
PULSOS	pulses
PA SISTOLICA	systolic blood pressure
PA DIASTOLICA	diastolic blood pressure
PPA	result SBP/DBP
NORMAL x ANORMAL	absence or presence of pathology
B2	type of the second heart sound
SOPRO	murmur type
FC	cardiac frequency
HDA1	histoy of disease 1
HDA2	histoy of disease 2 (other history)
SEXO	patient gender
MOTIVO1	first reason for being forwarded to the cardiology clinic
MOTIVO2	second reason for being forwarded to the cardiology clinic

Table 1: Data Dictionary

Variable PPA is calculated from PA SISTOLICA and PA DIASTOLICA, according to cardiac clinical tables that relate the patient gender with age and blood pressures (systolic and diastolic).

In order to be able to improve the quality of your analysis, it is required that you search for intervals normally used in pediatric cardiology, for example, for variables IDADE and IMC.

What to deliver? A report containing:

1. Raw data
    - describe variables according to their types: interval-scaled, binary, nominal, ordinal, ratio-scaled. Be aware that there are specific methods suitable to each type of variable.
    - Preliminary analysis (summaries, histograms, boxplots, spread measures, density). These are interesting to be applied to the raw data to "uncover" inconsistencies, outliers, duplicates etc.
    - List of main changes needed to be performed with the raw data.
2. Preprocessed data
3. Basic description (summaries, histograms, boxplots, spread measures, density).
4. Analysis
    - Bivariate analysis (correlations, regression)
    - Multivariate analysis (multiple variable regression, mutual information, cluster analysis - once more, be aware that some methods used to calculate similarity (dissimilarity) depend on the type of the variable. In the context of this dataset, when performing custer analysis we are interested to know if there are groups of similar patients)
    - Predictive Models (use of supervised machine learning)
        - Decision Tree learning
        - Support Vector Machines
        - Others (if time allows)
    - Comparison
5. Discussion and Main Conclusions

This work is to be performed by groups of at most two people.