# Reconhecimento de Padrões
# Pattern Recognition

2015/2016

*Project Assignment*

## Default of credit card clients

## 1 Background

Banks loose money when clients are not able to pay credit cards debts. The prediction that a given client will be unable to pay its obligations in the future is a challenging task that can be faced as a binary classification problem. The classification task is: given a set of indicators (features) infer if the client is able or not to succeed in paying its credit card.

## 2 Objective

Your task is to develop classifiers to predict if a given client will be able to pay (or not) its credit card in the next month based on a dataset collected in Taiwan.

## 3 Practical Assignment

### 3.1 Dataset Description

The dataset contain data, collected in October of 2005, from 30,000 clients from an important bank in Taiwan, and is available at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#. Among the total 30,000 clients, 5529 clients (22.12%) are the cardholders with default payment[1].

The 23 indicators (features) that aim to discriminate clients as being able or not to pay their credit card in the next month are summarized in Table 1.

### 3.2 Feature Selection and Reduction

Some of the supplied features may be useless, redundant or highly correlated with others. In this phase, you should consider the use of feature selection and dimensionality reduction techniques, and see how they affect the performance of the pattern recognition algorithms. Analyze the distribution of the values of your features and compute the correlation between them. Make sure you know your features! Do not forget to present your findings in the final report.

### 3.3 Experimental Analysis

You should be able to design experiences in order to run the pattern recognition algorithms in the given data and evaluate their results. Keep in mind that this is an unbalanced binary data set. Try to design the

Table 1: Features

| ID | Name | Possible Values |
|---|---|---|
| X1 | Amount of the given credit | Includes both the individual consumer credit and his/her family (supplementary) credit in dollars. |
| X2 | Gender | 1 = male;<br>2 = female |
| X3 | Education | 1 = graduate school;<br>2 = university;<br>3 = high school;<br>4 = others. |
| X4 | Marital status | 1 = married;<br>2 = single;<br>3 = others. |
| X5 | Age | Age in years. |
| X6 - X11 | History of past payment<br>from April to September, 2005:<br>X6 = the repayment status in September, 2005<br>X7 = the repayment status in August, 2005<br>...<br>X11 = the repayment status in April, 2005. | -1 = pay duly;<br>1 = payment delay for one month;<br>2 = payment delay for two months;<br>8 = payment delay for eight months;<br>...<br>9 = payment delay for nine months and above. |
| X12-X17 | Amount of bill statement:<br>X12 = in September, 2005;<br>X13 = in August, 2005;<br>. . .<br>X17 = in April, 2005. | Amount in dollars. |
| X18-X23 | Amount of previous payment.<br>X18 = amount paid in September, 2005;<br>X19 = amount paid in August, 2005;<br>...<br>X23 = amount paid in April, 2005 | Amount in dollars. |

classifier taking into account these issues. Justify your assumptions and decisions.
Define the performance metrics to evaluate your method (e.g. (AER) Average Error Rate, F-measure, ROC Curves, etc.). To run the experiments multiple times and to be able to present average results and standard deviations (of the metrics used) you should use cross-validation.
Do not forget that manually inspecting the predictions of your algorithm can give you precious insights of where it is failing and why, and what you can do to improve it (e.g. what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm deal better with those cases?). Go back and forward to the Pre-processing, Feature reduction and Feature Selection phases until you are satisfied with the results. It is a good idea to keep track of evolution of the performance of your algorithm during this process. Try to show these trends in your final report, to be able to fundament all the issues involved (choosing parameters, model fit, etc.)

### 3.4 Pattern Recognition Methods

You can write your own code or use the functions and methods available in Matlab and in the Statistical Pattern Recognition STPRTool used in the classes (since you are already familiarized with it). The methods used in your work should be described as well as discussion of the parameters used. Try out different pattern recognition algorithms. You should try to understand how they perform differently in your data.

### 3.5 Results and Discussion

Present and discuss final results obtained in your Project assignment. This problem was already studied by other authors. Compare your results with the results from other sources.

### 3.6 Code & Graphical User Intreface (GUI)

You should deliver your software code in MATLAB, or any other programming language you used during the project.

For your project you should write code for a graphical user interface in MATLAB. To aid you, MATLAB has for that purpose the built-in tool "guide" which can be called from the console. The GUI should improve the interaction of the user with the code by providing options for data-loading, feature selection/dimensionality reduction, classification, post-processing, validation and visualization.
Remember to comment your code. Write also a help section to your code that tells the purpose of the function, usage, and explanation of parameters. In MATLAB, comments following the first line of a function will show when help command is used with the name of the function.

## 4 Documentation

Write documentation (in Portuguese or in English) about your project. The documentation should include a cover page where course name, project title, date, names and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that reader would be able to implement the same kind of functions for feature extraction and classification just based on your documentation and some basic background in pattern recognition. Always justify your choices, even when their are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data to your documentation. At the end of your documentation you should have a list of all references used.

### 4.1 Requirements

Practical assignment is meant to be done in groups of two persons. If someone wants to work alone, this is also possible. Larger groups are not allowed.

### 4.2 Project Submission & Deadlines

1. **Project First Milestone (Deadline: 22th April 2016!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Minimum Distance classifier;
   - Matlab Code + short report.
2. **Project Final Goal (Deadline: 27th May 2016!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Several classifiers;
   - Final Report
   - Matlab code + Matlab GUI.
3. **Presentation and Discussion (3rd and 6th of June 2016!)**

### Acknowledgments

Credits to the UCI Machine Learning Repository[2].

## References

[1] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[2] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.