APRENENTATGE AUTOMÀTIC (APA)

Practical work

FALL SEMESTER 2021/2022

Grau en Enginyeria Informàtica - UPC

CONTENTS

CHAPTER

ONE

GENERAL INFORMATION

This is the **guide** for the correct development of the practical work for the APA ("Machine Learning") course. The students must apply the different concepts and techniques lectured during the course to solve a real problem. The students must write a complete report describing the work carried out, the problems encountered and the solutions envisaged, as well as the final results and conclusions of the study.

All students enrolled in APA are required to complete a term project. The goal is to develop a **classification or regression model** to solve one of the problems that can be obtained from selected data repositories (detailed in 2.2). You can choose to explore any problem that motivates you, and even bring your own proposal.[1]

You are expected to write a complete **report** describing the work carried out, its motivation, the problems encountered and the solutions envisaged, and the final results and conclusions of the study. The main text is *strictly* limited to 20 pages (this includes graphics, tables and references; note the code must be submitted *separately*).

The main programming language used for the modelling part is `python`. Remember that there are *many* packages for `python` which probably contain useful routines you can use; just be sure to mention them in your final document.

Other software can be used as long as it serves a specific or secondary purpose.

Any additional information on the methods or on the problems should be **acknowledged and/or properly cited**.

[1] Upon explicit acknowledgement.

CHAPTER

TWO

DEVELOPMENT

2.1 Experiments procedure

It is strongly recommended that you follow these steps in the development of your practical work:

1. Choose the problem and make sure you can get the data
2. Read the available documentation on both problem and data; get and read some relevant previous work on the same (or very similar) problem and data, if any
3. Pre-process the data and choose the variables you are going to use
4. Perform a basic statistical description
5. Choose the resampling method to fit, select and test your models
6. Visualize the data
7. Perform a full modelling process, using linear/quadratic techniques
8. Perform a full modelling process, using non-linear techniques

2.2 Obtaining the problems

You need to decide a problem that fulfills all the next conditions:

- The dataset of this problem has numerical and categorical variables.
- The dataset of this problem is not synthetically generated.
- The dataset of this problem contains more than 10 variables.
- You need to have enough information about the problem to be able to understand and analyze your results. Just getting random data and feeding it to the machine is not valid.

- Datasets already pre-processed are not valid. You need a problem which data has any pre-processing work to do.
- The dataset of this problem contains more than 200 samples.
- The problem to solve is not one of the simple known problems like the iris, mnist or wine.
- The problem is complex enough so you don't get an almost perfect accuracy/R2 score with a linear/logistic regression.

You can choose any dataset that fills the conditions. It is recommended that you chose a problem from a field that you like and motivates you.

If you don't know where to look for problems you can use the next repositories:

1. The UCI machine learning repository:
   http://archive.ics.uci.edu/ml/
2. The School of Informatics (University of Edinburgh) repository:
   http://www.inf.ed.ac.uk/teaching/courses/irds/miniproject-datasets.html
3. The datasets that arose from the Delve project:
   http://www.cs.toronto.edu/~delve/data/datasets.html
4. Luis Torgo's regression data sets:
   http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html
5. Kaggle (You can only use problems that do not have a notebook with its analysis)
   https://www.kaggle.com/

You can also use Google's dataset search engine https://datasetsearch.research.google.com/ if you are interested on a specific topic.

Then choose *one* of the problems (most of them are real-world tasks and many are quite challenging, possibly going beyond the APA level, specially in the computational resources needed to solve them).[1] The problems are very diverse, not only regarding the area of work (biology, geophysics, medicine, etc) but because they show different data characteristics. For example, there are great differences in the number of variables and examples, number of classes, intrinsic difficulty, missing values, various errors, mixed nominal and/or continuous variables, etc. Some problems are synthetic (they have been generated by a program), and their characteristics are completely known. However, their study is interesting for a number of reasons, including meaningful (as well as significant) comparisons of different learning algorithms.

Some problems are then easier in some aspects and more difficult in others. Therefore, the selection of the particular problem does not have much importance for the grade. In particular, it is not at all advisable that you start to test problems to see how they "behave". It is recommended that you base the decision on the interest that it raises in you, although avoid very small datasets that are too easy (any method predicts with almost perfect accuracy).

You must decide which **problem** you want to attack as soon as possible and communicate your choice (by e-mail) to `raquel.perez@bsc.es` no later than **November 8th, 2021**. Please indicate **all** the group member's names.

[1] In particular, avoid problems using images, unless the image size is very small.

Sending your problem for evaluation

In order to avoid having a too easy or too difficult problem for your project you should send a document with a basic analysis of your chosen dataset. This document should have:

- A small description of the dataset and its source. It should include which will be your target variable and if you will be solving a regression or classification problem. You should also comment specific problems of your data, like time related components and unbalance.
- A simple visualization of your variables comparing them with your target.
- A correlation matrix of your variables.

You can also add baseline results on a simple model. This is not mandatory but can help you find if your dataset is too easy.

2.3 On data pre-processing

Each problem requires a different approach concerning to data cleaning and preparation, and the selection of the particular information you are going to use can vary; this pre-process is very important because it can have a deep impact on future performance; it can easily take you a significant part of the time. It is then strongly advised that you analyse well the data before doing anything, in order to gauge the best way to pre-process it[2]. In particular, you shall pay attention to the following aspects (not necessarily in this order):

1. Treatment of lost values (missing values)
2. Treatment of anomalous values (outliers)
3. Treatment of incoherent or incorrect values
4. Coding of non-continuous or non-ordered variables (nominal or binary)
5. Possible elimination of irrelevant or redundant variables (feature selection)
6. Creation of new variables that can be useful (feature extraction)
7. Normalization of the variables (*e.g.* standardization)
8. Transformation of the variables (*e.g.* correction of serious skewness and/or kurtosis)

2.4 On model selection and estimation of performance

In accordance with the problem and the available data, you should design a set of experiments based on valid protocols to select models and to honestly estimate the generalization error (or any other measure of future performance) of the final proposed model or solution.

Some problems come with their own test data (data used for the estimation of true generalization error), some do not; in the latter case, you must obtain test data by splitting the full available data

[2] See Laboratory 0.

(once or several times, depending on the data size). For model selection, k-fold cross-validation will probably be necessary (the selection of the best value for k is your decision). It is methodologically prohibited to use as test data information that has already been used for the creation, adjustment or selection of the solution.

2.5 On model/examples interpretation

In order to assess the usefulness of a model it is also important to obtain some information about what attributes are considered more important in their decisions or what is important in the prediction of a selection of examples.

You should obtain/extract that information according to the type of model.

2.6 Project tasks delivery

The **final report** that you must deliver should include the following **sections**:

1. A brief but self-contained description of the work and its goals, and of the available data, and any additional information that you have gathered and used
2. A brief description of related previous work and results
3. The relevant data exploration process (pre-processing, feature extraction/selection, clustering and visualization)
4. The resampling protocol (training/test, cross-validation, etc) that you have used
5. The results obtained using **three linear/quadratic** methods (indicating the best set of parameters for each one):
   (a) If the task is **classification**, any of: logistic regression, multinomial regression (single-layer MLP), LDA, QDA, Naive Bayes, k-nearest-neighbours, linear SVM, quadratic SVM
   (b) If the task is **regression**, any of: linear regression, ridge regression, the LASSO, k-nearest-neighbours, linear SVM, quadratic SVM
6. The results obtained using **two general non-linear** methods (indicating the best set of parameters for each one); for both **classification** and **regression** tasks, any of: the MLP, the SVM with RBF kernel, a Random Forest
7. A description and justification of the final model chosen, and a honest estimation of its generalization error
8. The results of the interpretability analysis of the models:
   - What attributes are more important for the models and if there is a difference between lineal and non linear models
   - The explanation for a linear and a non linear model of your choosing of a selection of examples that seem relevant (e.g.: outliers, examples with extreme values, examples in the maximum/minimum values for the output value, examples classified incorrectly, examples where the more relevant attributes are not used, examples that seem prototypical...)

9. A final brief part containing:
    (a) A self-assessment of successes, failures and doubts
    (b) Scientific and personal conclusions
    (c) Possible extensions and known limitations
10. References to all your used sources: books, web pages, code, scientific papers, ...

Delivery: You will be required to submit a written report (a pdf file) and the **full code** (in separate files, and only electronically) and a brief text file with instructions on how to execute your code. The report should *not* include explanations of those methods seen in class, unless relevant for your needs. All deliveries are exclusively by means of the "Racó" in a **single compressed file**; you do not have to print anything. The deadline is **January 10th, 2022**.

CHAPTER

THREE

EVALUATION

3.1 Task evaluation

The grade will be partly based on the **clarity** of your report, so please make sure your final report is well organized and clearly written. There should be an introductory part explaining the basics of your work, and a conclusions section, basically stating what you know compared to what you knew before the work started; also any gaps, possible extensions or limitations in your development should be noted and explained.

Your work will also be evaluated based on **technical quality**. This means that the techniques you use should be reasonable, the stated results should be accurate, and the technical results should be correct and complete.

In summary, these are the conditions for a high score (in this order):

1. The (good) use of techniques and methods presented in class
2. The care and rigor for obtaining the results (validation protocol, statistical significance)
3. The quality of the obtained results (generalization error, simplicity)
4. The quality of the written report (conciseness, completeness, clarity).

3.1.1 Evaluation criteria

The next items will be used as a guide for evaluating the task. Every item will be given a mark between 0 and 10 based on this table. Note that there is no item about model accuracy, that is because having a greater accuracy or R2 will not affect the grade. Which will affect the grade is to justify properly every step and use adequate methodologies so every result you give is real.

1.	Code Execution
8-10	The code has a valid link or script for downloading the data, has a requirements.txt file with all its dependencies, has a README.md with valid instructions and can be executed without error following these instructions.
5-7	The code has a valid link for downloading the data, has a README.md with valid instructions and can be executed without error following these instructions.
0-4	There is no documentation about how to get the data or execute the code and the code returns an error when executed.
2.	**Data pre-processing**
8-10	Data pre-processing is adequate to the dataset, handles correctly its particularities and is coherent with the models used later. Every decision about it is properly explained in the document. Pre-processing is adequately applied to the different data partitions.
5-7	Data pre-processing is enough for fixing most the dataset "problems" like missing values. Decisions about pre-processing are not explained.
0-4	There is no pre-processing performed to the data or the pre-processing made makes no sense. Pre-processing is applied wrong to the different partitions of the data.
3.	**Resampling protocol**
8-10	Resampling protocol is adequate to the problem and is used right. Resampling protocol chosen is properly explained in the report.
5-7	Resampling protocol has any minor mistake that does not affect validity of the results or it is not properly explained.
0-4	Resampling protocol makes no sense at all and implies that all the results are not valid. Resampling protocol chosen is not properly explained in the report and has mistakes.
4.	**Hyperparameter selection**
8-10	Hyperparameter selection uses an adequate methodology, it is explained properly and every decision regarding it is justified.
5-7	Hyperparameter selection uses an adequate methodology but it is not explained well enough or justified.
0-4	Hyperparameters selection is not explained or ignored completely.

5.	**Final model selection**
8-10	Model selection uses an adequate methodology, it is explained properly and every decision regarding it is justified. An adequate methodology is used to estimate its generalization error.
5-7	Model selection has any minor mistake or is not justified enough, but it does not affect the validity of the results.
0-4	Model selection is not explained or ignored completely. Generalization error is not estimated or it is estimated wrong.
6.	**Adaptation to your problem context**
8-10	All decisions taken are adequate to your problem context and properly justified.
5-7	Most decisions taken are adequate to your problem context and properly justified.
0-4	No decisions taken are adequate to your problem context and properly justified.
7.	**Decision justification**
8-10	All design decisions taken during the project are justified and explained.
5-7	Most design decisions taken during the project are justified and explained.
0-4	No design decisions (or almost no design decisions) taken during the project are justified and explained.
8.	**Final error generalization**
5-10	Final error generalization is properly estimated using an adequate methodology.
0-4	Final error generalization is estimated wrongly or using the wrong data partition.
9.	**Interpretability analysis**
7-10	The analysis compares the weights for the different types of models and analyzes the explanation for a selection of relevant examples
3-6	The analysis compares the weights for the different types of models
0-2	The analysis is limited to showing the weights of the models.

3.2 Generic Competence Evaluation

In addition, as you probably know, there is a **generic competence** (or *skill*) associated to this course: "*Effective communication*"[1], which is worth an additional 10% of the final grade. In order to help you deal with this, the **rubric** with which the competence will be evaluated is available to everyone prior to delivery.

[1] Expressió oral i escrita.

Each indicator has four possible answers and four notes: A, B, C, D. The teacher decides first on which of the four cases each document is found for each indicator and then puts a numerical note according to the table:

[0-5)	D
[5-6.5)	C
[6.5-8.5)	B
[8.5-10]	A

For example, if an indicator has a B, then a note is placed in the range [6.5 - 8.5) as a function of their degree of performance. The grades are then promised and a competence grade (CG) is obtained. This mark counts 10.0% of the final mark of the subject. The conversion of CG to A, B, C, D (applying the table in reverse) gives the final note of the competence.

3.2.1 Evaluation criteria

1.	**General presentation**
A	The document has a cover with relevant information, uses appropriate margins, does not cut tables or figures and number the pages. It has a numbered index.
B	The document has a cover with the relevant information, numbers the pages and has an index.
C	The document numbers the pages, but has an upgradeable cover, not very suitable margins, and does not have one index (for medium or large size documents.)
D	The document has no cover, no proper margins, no number of pages, and no index by document medium or large size.
2.	**Introduction**
A	Poses the main topic and anticipates the structure of the work. The purpose, general exposition of the subject and the objectives are very clear, and allow a complete and quick idea of the work, including the results.
B	The purpose, general presentation of the topic and objectives are reasonably clear, but do not allow it to be done a complete and quick idea of the work.
C	The purpose, topic and objectives are presented, but in an unobjective and clear way.
D	No Introduction, or what is there is completely ineffective.
3.	**Organization and general structure**
A	Ideas are presented in a logical order, always with coherence and fluency, and enjoyed reading.
B	Ideas are presented in a logical order, usually with coherence and fluency.
C	Ideas are presented in a certain order, but many details are not in a logical or expected order, and distract the reader.
D	Ideas are not presented in a logical order, lack of coherence and the order of the paragraphs does not reinforce the content but obscures it.

4.	Grammatical correctness
A	There are no spelling or syntactic errors or typographical errors. The language used is rich.
B	There are no spelling or syntactic errors, but some typographical errors. The language used is reasonably varied.
C	There are some spelling or syntactic errors, or typographical errors. The language used is fair.
D	There are many spelling or syntactic errors, or typographical errors. The language used is poor.
5.	**Language style**
A	The document is written using precise and didactic language.
B	The document is written using acceptably accurate and understandable language, even if it is not completely didactic.
C	The document uses a language that is not very precise or grammatically dubious (ambiguities, contradictions...).
D	The document uses language full of inconsistencies and ambiguities and often grammatically incorrect.
6.	**Visual elements (tables, graphics...)**
A	Authors use visual elements where appropriate, are well done (not loaded or poor) and help a lot in understanding the work.
B	Authors use visual elements where appropriate, are acceptably performed, and often help understanding work.
C	Authors use visual elements, but not always where appropriate, they are not always acceptable performed and do not always help in understanding the work.
D	Authors do not use visual elements when it is convenient, or they are not well done or difficult understanding of work.
7.	**Conclusions**
A	The document ends with a clear summary describing what can be deduced from the work done, what work it could have been done differently, what would be the work to be done if you wanted to keep working because you have learned by doing the work.
B	The document ends with a clear summary describing what can be deduced from the work done and what has been done learned by doing the work.
C	The document ends with a summary that describes in part or in doubt what can be deduced from the work done.
D	The document does not include any conclusions or these are manifestly incomplete or incorrect.

8.	References (books, web pages, newspapers, courses, notes...)
A	The document correctly identifies and cites the information used and gives additional references.
B	The document identifies and cites the information used.
C	The document identifies and cites the information used, but not correctly.
D	The document does not cite the information used.