PRACTICA 1: AUDITORIA DE DATOS (Parte I)

Una de las partes más costosa en aprendizaje automático es preparar los datos para ser procesados posteriormente. Lo primero que tenemos que hacer es entender el problema al que nos vamos a enfrentar, invirtiendo tiempo en estudiar/visualizar la base de datos que nos facilita el cliente/colaborador.

En esta práctica se pide realizar una auditoría de los datos de la base suministrada. Esto nos permitirá conocer los datos con los que se va a trabajar.

El cliente (en este caso tus profesores) nos marca unos hitos para realizar el estudio de la base de datos (ten en cuenta que el cliente puede pedir puntos que son irrealizables en su base de datos):

- Descripción de las variables y valores estadísticos (mínimo, máximo, media, desviación, mediana, etc.). Estudia qué valores estadísticos son los convenientes según el tipo de variable y procede en consecuencia.
- Describe y realiza modificaciones en la base datos si lo consideras necesario. Por ejemplo, qué harías con valores nominales, si los hubiera.
- Estudia si es necesario normalizar los datos y cómo lo harías. Procede a modificar la base de datos (normalizar) si lo consideras necesario.
- Detección de valores extremos (outliers) y descripción de qué harías en cada caso.
- Detección de valores perdidos (missing values) y descripción de cómo actuarías para solventar el problema.
- Buscar correlaciones entre:
  - las variables predictoras, lo que permitirá ver si hay variables redundantes.
  - variables predictoras y la clase (target).
- Detecta, si hubiera, falsos predictores.
- Estudia si fuera conveniente segmentar alguna de las variables.
- Estudia si fuera conveniente crear nuevas variables sintéticas basada en las variables originales.

El cliente solicita un documento (audit), de un máximo de 10 páginas, que recoja las conclusiones de los puntos anteriores así como otras deducciones inferidas del estudio de la base de datos y que aportan conocimiento al problema.

PRACTICE 1: DATA AUDIT (Part I)

One of the most expensive parts of machine learning is preparing the data for further processing. The first issue we have to do is to understand the problem we are going to face, investing time to study/visualize the database provided by the client/collaborator.

In this practice we are asked to carry out an audit of the data of the supplied database. This will allow us to know the data we are going to work with.

The client (in this case your teachers) sets us some milestones to carry out the study of the database (bear in mind that the client can ask for points that are unachievable in their database):

(1) Description of the variables and statistical values (minimum, maximum, mean, deviation, median, etc.). It studies which statistical values are suitable according to the type of variable and proceeds accordingly.
(2) Describe and modify the database if necessary. For example, what would you do with nominal values, if any.
(3) Study if it is necessary to normalize the data and how you would do it. Proceed to modify the database (normalize) if you consider it necessary.
(4) Detection of extreme values (outliers) and description of what you would do in each case.
(5) Detection of missing values and description of how you would act to solve the problem.
(6) Search for correlations between:
    * the predictor variables, which will allow to see if there are redundant variables.
    * Predictor variables and the class (target).
(7) Detection of false predictors, if any.
(8) Study if it is convenient to segment some of the variables.
(9) Study if it is convenient to create new synthetic variables based on the original variables.

The client requests a document (audit), of a maximum of 10 pages, that gathers the conclusions of the previous points as well as other inferred deductions of the study of the database and that contribute knowledge to the problem.

PRACTICE 1: Dimensionality Reduction (Part II)

Let's get our hands now on a different real dataset from the UCI Machine Learning Repository.
We are going to test some of the methods explained on the Wine Dataset, an admittedly easy dataset. It consists of 178 samples with 13 constituents drawn from three types of wine.

1. Mutual information. We are going to investigate the use of the mutual information criterion to evaluate a set of candidate features and to select an informative subset to be used as input data for a typical classifier.

First of all, you should select a suitable subset of variables and plot those with higher mutual information. Are you able to distinguish the three types of wine?

2. Chi-Square. Repeat the selection of variables with Chi-Square method. Do you get the same results as with the previous one?

3. Principal Components Analysis (PCA). Now we are going to work with PCA as a method for dimensionality reduction. The Principal Component Analysis (PCA) was independently proposed by Karl Pearson (1901) and Harold Hotelling (1933) to turn a set of possibly correlated variables into a smaller set of uncorrelated variables. The idea is that a high-dimensional dataset is often described by correlated variables and therefore only a few meaningful dimensions account for most of the information. The PCA method finds those directions in the original dataset that account for the greatest variance in data, also called the principal components.

a) PCA without normalization:

- Calculate the eigenvalues and plot them. How many components do you need to explain 90% of the total variance?

- Plot the two first components, are the resulting clusters clearly separated?

b) PCA with normalization: Repeat the two previous steps but in this case scaling the input to zero mean and unit variance N(0,1), it is also called z-scores. What do you see now? In our dataset, why does PCA without normalization perform poor?

4. Linear Discriminant Analysis (LDA). What we aim for is a projection that maintains the maximum discriminative power of a given dataset, so a method should make use of class labels (if they are known a priori). The Linear Discriminant Analysis, invented by R. A. Fisher (1936), does so by maximizing the between-class scatter, while minimizing the within-class scatter at the same time.
a) LDA without normalization: Calculate the two components (C-1) and plot them, are the resulting clusters clearly separated?
b) LDA with normalization: Repeat the previous step, what do you see now? Which is the difference with the previous one?

5. Logistic Regression and Model Evaluation

From the wine dataset, we are going to study the importance (or not) of reducing dimensionality.

We are going to apply logistic regression as a predictive model and see the influence of increassing the number of predictor variables. In the wine data set the dependent variable is not binary, therefore we need to perform a study two to two classes: 1 vs. 2, 1 vs. 3 and 2 vs. 3.

To do this, you must create two datasets : training and test, with a 70:30 ratio.

We will apply a forward method increasing the number of variables one by one and we will observe how the prediction outcomes vary.

A summary table similar to the one shown here will be constructed:

LOGISTIC REGRES-SION	Accuracy	Precision	Sensitivity	Specificity	AUC-ROC
Class 1 vs. 2					
Full Data - No reduction					
MI – 1 variable					
MI – 2 variables					
…..					
Chi2 – 1 variable					
Chi2 – 2 variables					
…..					
PCA (1 component)					
PCA (2 components)					
PCA (3 components)					
….					
LDA (1 component)					
	Accuracy	Precision	Sensitivity	Specificity	AUC-ROC
Class 1 vs. 3					
Full Data - No reduction					
MI – 1 variable					
MI – 2 variables					
…..					
Chi2 – 1 variable					
Chi2 – 2 variables					
…..					
PCA (1 component)					
PCA (2 components)					
PCA (3 components)					
….					
LDA (1 component)					
	Accuracy	Precision	Sensitivity	Specificity	AUC-ROC
Class 2 vs. 3					
Full Data - No reduction					

MI – 1 variable					
MI – 2 variables					
…..					
Chi2 – 1 variable					
Chi2 – 2 variables					
…..					
PCA (1 component)					
PCA (2 components)					
PCA (3 components)					
….					
LDA (1 component)					