Impact of new technologies for teaching statistics

Mathilde Mougeot

Université ParisX Modal'X
200 avenue de la République 92000 France
mathilde.mougeot@u-paris10.fr

Abstract:

This paper presents the way we take benefit of new technological improvements for teaching statistics. The underlying courses are given in « data mining » and in « statistics consulting » at a master level: Master ISIFAR[1] at Paris X University.

Introduction:

Nowadays, due particulary to new technologies, the data sources have deeply changed. Every compagny owns its datawarehouse to store current and strategic data. Every plant owns its Distributed Control System (DCS), where the information management system is used for long term data storage, data and alarm logging. Large data bases, available through internet or private networks, are then continuoulsy and automatically filled with data. One major task for statisticians is now to work on data samples to estimate, test, and produce decision rules. In order to create strategic indicators or to measure plant performances, current statisticians need to work on *large data bases*. The compilation and deployment of statistical techniques is then almost universally based on *computing systems* on large data bases.
In the areas of academic statistics, new technologies have also changed the approach of teaching to the whole range, from data collecting through processing, analysis and inference to decision.
In this paper, we describe how we take advantage of the new technologies in our statistics courses, at a Master level. Students attending this Master degree are looking for a job just after their academic year. An important aspect is to anticipate their needs for their futur job from a statistical point of view.

Simulations to illustrate statistical & theoretical properties

In the data mining course, various learning methods are introduced for prediction or classification. Historically, the Perceptron is one of the first learning method proposed by Rosenblatt for classification (1957). Despite the linear solutions provided by this method, the problem is interesting for studying the process of convergence. Starting from a random starting point, the algorithm converges progessively to an ending solution by minimizing a risk function after mutliple presentations of all the examples of the learning data set. Thanks to the power of current computers, it is nowadays possible to study dynamically the convergence of such an algorithm during a class.
Less than 30 seconds are needed to compute a solution with 500 examples using more than 100 presentations of the learning data set.
The following figure illustrate the process of convergence for a given example studied during the course. A set of 100 examples belonging to two disctinct clusters are previously artificially

1 MASTER ISIFAR: Ingéniérie et Informatique de la finance, de l'assurance et du risque

generated. For simplicity here, the two clusters are linearly separable. At the beginning of the algorithm, a separation line is chosen at random (figure 1.b). Here, in two dimensions, the separation is defined by two coefficients: slope and intercept. We can observe that this first boundary doesn't fit at all. Based on Rosenblatt's algorithm, the corresponding coefficients are iteratively updated as the data basis examples are progressively presented and the risk function minimized. At the end of the process, the computed coefficients correspond to the optimal solution, and we can observe that the two clusters are righlty separated (figure 1.d).

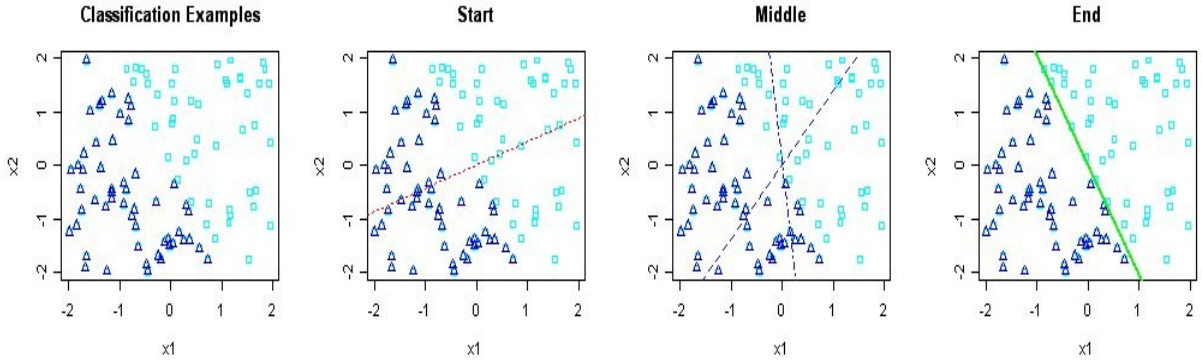

Figure 1: Illustration of convergence of the perceptron for classification. The set of data are here linearly separable(1/4). Initial condition are chosen randomly (2/4). After progressiv convergence (3/4), the algorithm finds the optimal solution (4).

During the course, the students can visualize the convergence process step by step. Experience shows that the convergenc sequence helps the students to undertand afterwards more complex optimization methods.
R software is used for these simulations. The first script, containing all the instructions in the R language, is given to the students. The students have to modify progressively the script to solve more complex solutions, in particular for non linear cluster solutions using neural networks techniques.

Data mining on large data base:

At Paris X University, a specific course at a 2nd year Master level has been developed, called « Statistics Consulting Courses ». The approach is to consider that the data are the heart of all statistical education and that students should be introduced to statistics through data-centered courses. The purpose of this course is to apply and merge statistical methods to solve real industrial problems based on real industrial data. Students work by group of two, in a small class limited to 20 students in order to promote exchanges and discussions. One part of the program is to create a statistical method to be able to monitor an equipment using appropriate variables originated by sensor data. A decision rule should be able to trigger alarm when the equipment is malfunctioning and no more efficient. We use an extract of a industrial data basis. The data basis contains more than 260 000 recordings corresponding to hourly values for 10 sensors over 3 years. Students can not work « manually » on such a large data basis and should develop automatic procedures and requests to analyse first the data and then build the methodology.

Missing values.. The first task is to analyse each variable independently. The reflex of each student is to compute basic statistics on each variable as mean, median, sd.... (table 1). The students are very surprised by the first displayed results: most of statistical indicators are providing « NA[2] »

2 NA: Not A number

values. The students discover that the corresponding data basis contains missing values which prohibit any computations as is mostly the case in industrial data bases.

	E	***X_1***	***X_2***	***X_3***	***X_4***	***X_5***
Raw data: 8760 observations: Mean, sd	NAN	NAN	NAN	NAN	NAN	NAN
Missing values pre-treatments: 8333 observations Mean, sd	16 000	20 017	-58	4,60	17,59	11,98

Table 1: On real data bases, data pre-treatments (elimination or replacement) are compulsory in order to compute any statistics. E is the target variable which should be predict using the other explanatory variables $X_1..X_5$.

Each group of students work independently and is free to choose its own method to handle the missing values, that is to say: elimination or replacement. In this particulary case, after elimination of the missing values, 95% of the original data are kept.

Extreme or abnormal values..: The distribution of each variable can then be displayed on each computer. The first analysis exhibits extreme values in the distributions (figure 2.a). The first question is to determine if the extreme values are belonging or not to the underlying distributions. Extreme values can be linked to extreme phenomena or to abnormal phenomena, as sensors failures for examples. The students know the impact of extreme values on computing indicators. Each group of students have to make their own decision in order to eliminate or keep these extreme values (figure 2.b).

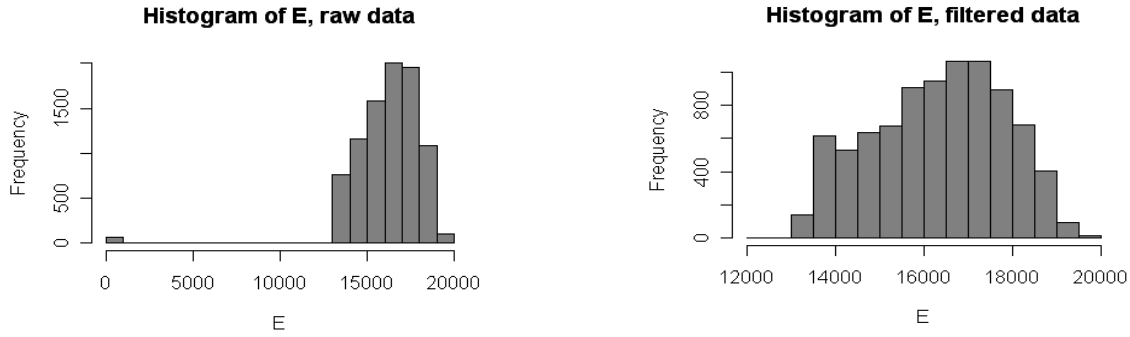

Figure 2: extreme or abnormal values should be eventually replaced to provide robust models. On the left figure, a class of extreme values can be observed far away from the main part of the distribution. Extreme values can be eliminated by chosing an appropriate threshold (right figure) to provide a continuous histogram.

Nowadays, the use of statistical software are very convenient. Students can easily produce graphics, histograms in order to visualize the results of their pre-treatments. At the end of each step, we select groups of students, who choose a different approach. The first conclusions of each group of students are presented to the whole class, by plugging a video projector on each computer. The main graphical results of each group can be easily displayed on a screen to analyse the impact of each choice. All the students are then invited to participate to the discussion.
After this first step, each student remembers that the *quality of the data base* should imperatively be analysed, that any raw data base should be pretreated before any statistical treatments.

Modeling..:
The previous explanatory analysis leads to reliable data which can be then used to build an empirical model. The equipment target variable is explained by other explanatory variables. Using regression, neural network, and regression tree methods, various models are built and then tested

(figure 3). In order to build a robust model, the initial pretreated data basis is split in two. A first learning data basis, is devoted to estimate the coefficients of each underlying model; the second test data basis is used to estimate the model error independantly of previous estimation. The modeling step is followed by a selection step where the best model is retained from all the the different produced models. All these treatments can be executed during the class using, for example, SAS Entreprise MinerTM to produce the different analyses and the corresponding diagram (figure 3).

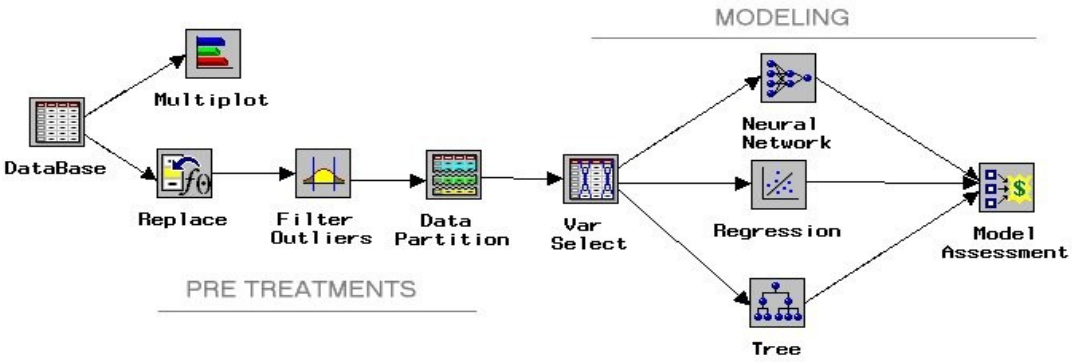

Figure 3: Diagram of all the treatments using SAS Entreprise Miner^{TM}. Each box corresponds to a specific methodology step and the methodology is given by the set of boxes.

Today, this course can be carried out due to the new statistical softwares that are very easy to use and can provide rapid results even on large data bases. Students can test different statistical approaches and immedialty use the software to produce and analyse the corresponding results. The different chosen approaches of the groups of students can be easily displayed on screen. The advantages or disadvantages of each approach can be discussed by the whole class.

References:

- Carter L. & Mougeot M. (1998) Use of Excel in a first course in Statistics for Mathematical Students. ICOTS-5, The first International Conference on Teaching Statistics, Singapore, June 1998.
- Carter L. & Mougeot M. (1994) Simulations to illustrate results in theoretical statistics. Proceedings of four International Conference on teachings Statistics Marrakech, july 1994.
- Hastie T, Tibshirani R & Friedman J (2001) The elements of statistical learning. Springer.
- Rosenblatt F. Principle of neurodynamics, Spartan Books, 1962
- Tukey J.W. (1977) Exploratory data analysis, EDA 1977