Department of Mathematics and Computer Science
University of Southern Denmark, Odense

April 26, 2011
Marco Chiarandini

DM825 - Introduction to Machine Learning

Obligatory Assignment 1, Spring 2011

Deadline: 19th May 2011 at noon.

Although it is acceptable that students discuss the assignment with one another, each student must write up his/her homework on an individual basis. Each student must indicate with whom (if anyone) they discussed the assignment.

Exercise 1 – Prediction of count outcomes

The software engineers of an online shopping portal wish to predict the number of clicks that a new product will receive in their online price search engine. New products are those that are not yet in the data base of the portal and for which there is no historical data available. The engineers would like to charge the vendors of the products on the basis of these predictions.
When a vendor inserts its product in the data base, the products are classified as belonging to an existing category or initiating a new category. In the following, we will assume that the category of a new product is known and it comprises products with similar names and characteristics.
The engineers collected historical data including the following variables:

- offer_id: a numerical identifier of the entry
- product_name: the name of the product (this field is at the moment problematic as it exhibits no formalism)
- category_id: a numerical identifier for the category.
- price: the price of the product in euro (-1 if not given)
- deliver_price: the price for delivery (-1 if not given)
- total_price: the sum of the previous two prices (-1 if not given)
- availability: -1 see merchant site, 0 not available, 1 available, 2 available soon, 3 limited availability
- merchant: an encrypted name of the vendor
- number_reviews: the number of reviewers who rated the vendor
- average_rating: the average rating of the reviews
- number_clicks: the response variable indicating the number of clicks.

The historical data are relative to one single day and the prediction is asked on the same unit of time.
A useful nonlinear regression model when the outcome is a count, with large-count outcomes being rare events, is the Poisson model. The Poisson probability distribution is given by

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, \ldots$$

Your tasks:

(a) The data set is confidential and it has been made available to you from the BlackBoard system. Retrieve the data set, load it in R via `load("clicks.rda")` and examine its content via `str(clicks)` and `head(clicks)`. You may explore possible relations between variables with the aid of some graphical representation like those in the example of the `splom` function of the package `lattice`. Use the obtained insights to preprocess the data and identify a subset of variables to use as predictors. Report the most significant insights eventually also with plots.[1]

(b) Write the Poisson distribution in the Exponential family form and indicate the canonical response and the link function.

(c) Fit a generalized linear regression model to the data. Interpret and assess its results. For the assessment you must choose an appropriate measure to gain a quantitative measure and distinguish training and test data.

(d) Which other predictive model studied in class can be applied to this task?

Exercise 2 – Credit risk assessment

The Basel II Capital Accord on Banking Supervision imposes that, in order for a bank to adequately compute its capital requirements, a reliable credit risk quantification technique should be applied.
Available are data of 76 small firms that are client of a bank. The sample is rather diversified and distributed across industries and the scenario is relevant in practical terms. Hence, also in this case, the data set is confidential and is available only from the BlackBoard system (package `credit-risk.tgz`, it includes an R script to load the data). There are annual data between 2001 and 2003, so that the sample period covers three years. For each firm there are 15 indices. The first 8 are financial ratios drawn from the balance sheet. The remaining 7 are credit-position ratios calculated by comparing the credit positions with respect to benchmarks.
The sample firms are split into two groups: the *in bonis* group (firms repaying their loan obligations at the end of the analysing period, for a total of 48 firms) and the *default* group (firms not repaying their loans at the end of the period, 28 firms).
A pre-processing phase is important to understand the type of data at hand, to reveal anomalies and to choose proper representations. Sometimes normalizations and transformations of data are convenient. A relevant problem with the data here is the presence of missing values. Under the hypothesis of missing data at random (MAR) three possible ways to handle missing values here are:

[1] Note that for plotting to a graphical device like pdf with the functions from the lattice package, you need to enclose the call of the function in print.

- *mean imputation*: assign the average of the available data for each variable.
- *ad-hoc refinement* of the mean imputation method: assign to the missing value the mean per firm over the available years. Only if, for a variable and a firm, all years are missing, assign the overall mean for that variable.
- *multiple imputation method*: first, create plausible values for missing observations that reflect uncertainty about a model for the input variables. Use these values to "fill-in" or impute the missing values. Repeat this process, resulting in the creation of a number of "completed" datasets, which allows the uncertainty regarding the imputation to be taken into account. Use the augmented data set that includes all the completed datasets in your analysis

For more ideas on how to handle missing data you are referred to [1].
For the assessment of your prediction models you will use a precomputed bootstrap. With the data set, you will find indication of which observations are to be used as *training data* and which as *test data*. Recall that the former data are to be used to learn the parameters of the models proposed and the latter data are for comparing and assessing their performance. Specifically, from the bulk of data available comprising all the 76 financial firms over a period of three years, a sample of 53 firms is indicated for the training data. In all, there are 50 such training data set indicated. For every training data the firms left out are considered for the test data. Note that you have only to consider the prediction on the third year.

Your tasks:

(a) Use logistic regression and neural networks to predict the group of the firms.

(b) Assess the classification performance of the prediction methods by computing the confusion matrix on the test data. Report the average confusion matrix over the 50 bootstraps.

(c) Generally speaking, banks are most interested in correctly classifying *default* firms, since mis-classifications of default would lead to grant the loan to an unsafe firm: it has been reported that trade experts deem misdefault as 2–20 times more serious than misbonis.

Use this information to create a loss matrix and reassess the prediction of the classifiers on the test set letting them to select for each sample the class that minimizes the corresponding loss function.

(d) Are there any other methods encountered during the course that you deem appropriate for this analysis and that you may be willing to try?

References

[1] Nicholas J. Horton and Ken P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1):79–90, 2007.