Tutorial 3

Statistical models

École Nationale des Ponts et Chausées
Département Ingénieurie Mathématique et Informatique – Master II

Loïc BRIN • Benoit ROGER

Exercise 1: Term structure of default probability deduced from a Transition Matrix.

This exercise is based on the following S&P transition matrix:

	AAA	AA	A	BBB	BB	B	CCC	D
AAA	90,8%	8,3%	0,7%	0,1%	0,1%	0,0%	0,0%	0,0%
AA	0,1%	91,2%	7,9%	0,6%	0,1%	0,1%	0,0%	0,0%
A	0,9%	2,4%	90,0%	5,4%	0,7%	0,3%	0,1%	0,1%
BBB	0,0%	0,3%	5,9%	86,9%	5,3%	1,2%	0,1%	0,2%
BB	0,0%	0,1%	0,7%	7,7%	80,5%	8,8%	1,0%	1,2%
B	0,0%	0,1%	0,2%	0,5%	6,5%	82,7%	4,1%	5,9%
CCC	0,2%	0,0%	0,2%	1,3%	2,3%	12,9%	60,6%	22,5%
D	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	100,0%

available here `http://defaultrisk.free.fr/data/TD2_1.csv`.

1. Load the database and define a function that for a given number of year (n) and a given credit rating (l), returns a n-vector with the PD for each n years, deduced from the S&P transition matrix.

2. Plot the PD for all the ratings, for the next 15 years, deduced from the transition matrix.

3. What do you observe?

Exercise 2: Load and explore the "German Credit" database.

1. Load the "German Credit" database available here `http://defaultrisk.free.fr/data/TD2_2.csv`. Name it `Credit` and display the first ten rows.

2. Display the size of the dataframe `Credit`.

3. Which columns are numerical? Which are categorical?

4. The variable `Default` is the variable we want to predict. Is the dataset balanced?

5. Split the data set into a training set and a test set (70-30%) keeping the same proportion of "Default" and "No Default" in both sets.

Exercise 3: Predict Default using a logistic regression.

1. Fit a logistic regression on the training set to predict the binary variable `Default` using the whole set of

predictors. Fit another logistic regression using only `Age` and `Status`.

2. What are the significant predictors? As a matter of simplicity use only the predictors `Age` and `Status` ?

3. Fit a lasso logistic regression and optimize the regularization parameter using cross-validation. AUC will be used to select the regularization/penalty parameter.

4. Compare the AUC of the two models (before and after regularization) on the test set and conclude.

Exercise 4: Predict Default using tree-based methods.

1. Fit a tree to classify the variable `Default`.

2. Display the confusion matrix of the predictions on the test set.

3. Try to improve your classifier with bagging (optional: modify the cutoff to improve your detection of Default). Why modifying the cutoff is of particular interest when trying to predict default?

4. Use a random forest algorithm to improve your classifier. What are the most important variables?