Actuarial Applications Seminar

Seminar on Statistics I

Data Science Applications with Python

The objective of this course is to prepare students with the basic practical tools to perform a Data Science analysis using Python software. In particular, the Keras, Scikit-learn and TensorFlow libraries will be used for the development of the projects.

Professor: Dr. Arrigo Coen Coria

The syllabus is as follows which is divided into 5 modules:

1. **Python for Data Science Overview**
   - 1.1. Arithmetic and variables
   - 1.2. Conditionals and flow control
   - 1.3. Functions
   - 1.4. Scikit-learn y TensorFlow
   - 1.5. Jupyter Notebook
   - 1.6. Git and GitHub tools
   - 1.7. The CRISP-DM model to carry out a project

2. **Classification algorithms**
   - 2.1. Definition and general concepts
   - 2.2. Classification algorithms
   - 2.3. Classification algorithm generalizations
   - 2.4. Precision and error measurements
   - 2.5. ROC Curve
   - 2.6. Error analysis
   - 2.7. Multiple classifiers
   - 2.8. Application with Scikit-Learn

3. **Linear, nonlinear regression and its generalizations**
   - 3.1. Definition and general concepts
   - 3.2. Linear regression
   - 3.3. Gradient Descent Algorithm
   - 3.4. Polynomial regression
   - 3.5. Learning curves
   - 3.6. Regularized linear models
   - 3.7. Logistic regression

3.8. Application with Scikit-Learn

4. Decision trees

4.1. Definition and general concepts
4.2. Generation of decision trees for regression and classification
4.3. CART algorithm
4.4. Tree mogging and cultivation strategies
4.5. Tree hyperparameters
4.6. Random forests
4.7. ADA algorithm
4.8. Application with Scikit-Learn

5. Neural networks

5.1. Perceptron
5.2. Activation functions
5.3. Network depth and Backpropagation algorithm
5.4. Network training and tuning
5.5. Network acceleration algorithms
5.6. Application with TensorFlow

Evaluation

The course will be evaluated as follows:

- 75% Projects per module: For each of the five modules there will be a corresponding task.
- 25% Final Project: The final project is the analysis of a data set using the appropriate Data Science techniques for its analysis. This project consists of a written work and an oral presentation of it.

Calendar

- Module 1: 29 Ene – 7 Feb
- Module 2: 10 Feb – 6 Mar
- Module 3: 9 Mar - 27 Mar
- Module 4: 30 Mar - 24 Apr
- Module 5: 27 Apr – 21 May

Bibliography:

Géron, A. (n.d.). *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*.

Harrington, P. (2012). *Machine learning in action*. Manning Publications Co.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Elements of Statistical Learning 2nd ed. *Elements*, *27*(2), 745. https://doi.org/10.1007/978-0-387-84858-7

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. MIT Press.

Müller, A. C., & Guido, S. (n.d.). *Introduction to machine learning with Python : a guide for data scientists*.

Shalev-Shwartz, S., & Ben-David, S. (n.d.). *Understanding machine learning : from theory to algorithms*.

VanderPlas, J. (2016). *Python Data Science Handbook:ESSENTIAL TOOLS FOR WORKING WITH DATA*. *O'Reilly*. Retrieved from http://shop.oreilly.com/product/0636920034919.do%0Ahttps://jakevdp.github.io/PythonDataScienceHandbook/05.01-what-is-machine-learning.html

BibliografíaCook, D. and Swayne, D.F. (2007). Interactive and Dynamic Graphics for Data Analysis With R and GGobi

Efron, B., Hastie, T. (2016). Computer Age Statistical Inference. Algorithms, Evidence and Data Science. Cambridge University Press.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd ed., Springer. TEXTO a seguir en el curso y disponible en Springer a trav́es de la UNAM

Hastie, T., Tibshirani, R., Wainwright, M. (2015). Statistical Learning with Sparsity. The lasso and generalizations. Chapman and Hall.

Højsgaard, S., Edwards, D., Lauritzen, S.L. (2012). Graphical Models with R. Springer. Disponible en Springer a trav́es de la UNAM

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to Statistical Learning. With applications in R, Springer. TEXT to follow in the course laboratory and available in Springer through the UNAM

Kuhn, M, Johnson, K. (2013). Applied Predictive Modelling. Available in Springer through unam

Ripley, B.D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

Scutari, M and Denis , J-B. (2015). Bayesian networks . With examples in R. Chapman and Hall.

Venables, W.N. and Ripley, B.D. (2002). Modern Applied Statistics with S. Springer– Verlag.