

Seminario de Aplicaciones Actuariales
Seminario de Estadística I
Aplicaciones de Ciencia de Datos con Python

El objetivo de este curso es preparar a los alumnos con las herramientas prácticas básicas para realizar un análisis de Ciencia de Datos utilizando el software Python. En particular, se utilizarán las bibliotecas Keras, Scikit-learn y TensorFlow para el desarrollo de los proyectos.

Profesor: Dr. Arrigo Coen Coria

El temario es el siguiente el cual está dividido en 5 módulos:

1. Aspectos generales de Python para Ciencia de Datos

- 1.1. Jupyter Notebook
- 1.2. Las herramientas Git y GitHub
- 1.3. Aritmética y variables
- 1.4. Condicionales y control de flujo
- 1.5. Funciones
- 1.6. El modelo CRISP-DM para realizar un proyecto

2. Algoritmos de regresión

- 2.1. Definición y conceptos generales
- 2.2. Regresión lineal simple
- 2.3. Regresión lineal múltiple
- 2.4. Modelos lineales regularizados (Ridge & LASSO Regression)
- 2.5. Regresión polinómica
- 2.6. Algoritmo de Descenso por gradiente
- 2.7. Curvas de aprendizaje

3. Algoritmos de clasificación

- 3.1. Definición y conceptos generales
- 3.2. Medidas de precisión y error
- 3.3. K-Nearest Neighbors (KNN)
- 3.4. Support Vector Machines (SVM)
- 3.5. Regresión logística
- 3.6. Curva ROC
- 3.7. Análisis del error
- 3.8. Generalizaciones de algoritmos de clasificación
- 3.9. Múltiples clasificadores

4. Árboles de decisión

- 4.1. Definición y conceptos generales
- 4.2. Generación de árboles de decisión para regresión y clasificación
- 4.3. Algoritmo CART
- 4.4. Estrategias de podado y cultivo de árboles
- 4.5. Hiperparámetros de árboles

4.6. Bosques aleatorios

4.7. Algoritmo ADA

5. Redes Neuronales

5.1. Perceptron

5.2. Funciones de activación

5.3. Profundidad de redes y algoritmo Backpropagation

5.4. Entrenamiento y ajuste de redes

5.5. Algoritmos de aceleramiento de redes

Evaluación

El curso será evaluado de la siguiente manera:

- 75% Proyectos por módulo: Para cada uno de los cinco módulos habrá una tarea correspondiente.
- 25% Proyecto final: El proyecto final es el análisis de un conjunto de datos utilizando las técnicas de Ciencia de Datos adecuadas para su análisis. Este proyecto consta de un trabajo escrito y una exposición oral del mismo.

Calendario

- Módulo 1: 30 ago – 3 sept
- Módulo 2: 6 sept – 24 sept
- Módulo 3: 27 sept – 15 oct
- Módulo 4: 18 oct – 12 nov
- Módulo 5: 15 nov – 26 nov

Bibliografía

- 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>
- Cook, 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 través 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 través de la UNAM
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to Statistical Learning. With applications in R, Springer. TEXTO a seguir en el laboratorio del curso y disponible en Springer a través de la UNAM
- Kuhn, M, Johnson, K. (2013). Applied Predictive Modelling. Disponible en Springer a través de la 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.