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. Generalidades de Scikit-learn y TensorFlow
   - 1.7. El modelo CRISP-DM para realizar un proyecto

2. **Algoritmos de clasificación**
   - 2.1. Definición y conceptos generales
   - 2.2. Algoritmos de clasificación
   - 2.3. Generalizaciones de algoritmos de clasificación
   - 2.4. Medidas de precisión y error
   - 2.5. Curva ROC
   - 2.6. Análisis del error
   - 2.7. Múltiples clasificadores
   - 2.8. Aplicación con Scikit-Learn

3. **Algoritmos de Regresión**
   - 3.1. Definición y conceptos generales
   - 3.2. Regresión lineal
   - 3.3. Algoritmo de Descenso por gradiente
   - 3.4. Regresión polinómica
   - 3.5. Curvas de aprendizaje
   - 3.6. Modelos lineales regularizados
   - 3.7. Regresión logística

3.8. Aplicación con Scikit-Learn

4. Arboles de decisión

4.1. Definición y conceptos generales
4.2. Generación de arboles de decisión para regresión y clasificación
4.3. Algoritmo CART
4.4. Estrategias de podando y cultivo de árboles
4.5. Hiperparámetros de arboles
4.6. Bosques aleatorios
4.7. Algoritmo ADA
4.8. Aplicación con Scikit-Learn

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
5.6. Aplicación con TensorFlow

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: 20 sep – 8 oct (3 semanas)
- Módulo 2: 11 oct – 29 oct (3 semanas)
- Módulo 3: 1 nov – 19 nov (3 semanas)
- Módulo 4: 22 nov – 10 dic (3 semanas)
- Módulo 5: 13 dic – 28 ene (3 semanas)

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

Bibliografía

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 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. TEXTO a seguir en el laboratorio del curso y disponible en Springer a trav́es de la UNAM

Kuhn, M, Johnson, K. (2013). Applied Predictive Modelling. Disponible en Springer a trav́es 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.