PR0MiDAT

IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

Machine Learning con Python 2

Aprendizaje no supervisado

(Métodos Exploratorios)

(506) 2268.8823 - (506) 8708.9091 info@promidat.com facebook.com/oldemarrodriguez www.promidat.com

Tutor: El curso será impartido por Dr. Oldemar Rodríguez graduado de la Universidad de París IX y con un postdoctorado en Minería de Datos de la Universidad de Stanford.

Duración: Cuatro semanas.

Descripción:

En este curso se presentarán los principales conceptos y métodos en Minería de Datos. El énfasis principal del curso será examinar dichos métodos desde un punto geométrico y de sus aplicaciones concretas. Se le dará especial importancia al uso de los conceptos de minería de datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas y paquetes especializados en Machine Learning en Python.

Objetivos:

En este curso el estudiante será capaz de:

1. Entender la necesidad de la utilización de modelos, algoritmos, software especial para el descubrimiento de conocimiento en grandes volúmenes de datos.
2. Conocerá la metodología del ciclo de desarrollo usado para el descubrimiento del conocimiento en grandes bases datos (KDD – “Knowledge Discovery in Databases”).
3. Entender las diferencias entre: estadística, análisis de datos, recuperación de la información, ML – “Machine Learning”, minería de datos y Ciencia de Datos.
4. Conocer los principales modelos, técnicas y algoritmos utilizados para descubrir el conocimiento en grandes volúmenes de datos.
5. Efectuar Análisis Exploratorio de Datos con Python.
6. Utilizar Python en métodos no supervisados de análisis de datos, como son Análisis en Componenentes Principales, Clusterización Jerárquica, K-Medias.

Metodología:

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas.

- Una vídeo conferencia semanal, las cuales quedarán grabadas en Webex, para que los alumnos la puedan acceder en cualquier momento.
- Trabajos prácticos semanales.
- Foros para plantear dudas al tutor y compañeros.
- Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Machine Learning, por ejemplo, que involucren segmetación de carteras de clientes.

Contenido:

1. Conceptos de la Machine Learning

a. Definiciones básicas en Machine Learning y Ciencia de Datos.
b. Instalando Anaconda, Spyder, Scikit-Learn, NumPy, SciPy, IPython, Jupiter, Matplotlib, Pandas y Sympy.

2. Análisis Exploratorio de Datos

a. Tipos de variables.
b. Estadísticas básicas y matriz de correlaciones.
c. Tablas de datos y datos atípicos.
d. Uso de NumPy.
e. Manipulación de Datos con Pandas.
f. Visualización de Datos con Matplotlib.

3. Métodos de condensación de la información

a. Análisis en Componentes Principales – ACP (PCA, Karhunen-Loeve o K-L Method)
    - Plano principal

- Círculo de correlaciones
- Dualidad y sobre-posición de gráficos
- Análisis Factorial de Correspondencias Múltiples

b. Aplicaciones en casos reales con los paquetes de Python Scikit-Learn y Prince.

4. Clustering Jerárquico Aglomerativo

a. ¿Qué es “cluster analysis”?
b. Clustering Jerárquica Aglomerativa.
    - Distancias y matrices de distancias.
    - Agregaciones.
    - Jerarquías binarias.
    - Jerarquías Binarias sobre las Componentes Principales.
c. Aplicaciones en casos reales con Scikit-Learn.

5. Método de k-medias (k-means)

a. Inercia total, inercia inter-clases e inercia intra-clases.
b. Teorema de Fisher.
c. Problema combinatorio.
d. Método de Forgy.
e. Método de las nubes dinámicas.
f. Aplicaciones en casos reales con R y el paquete Scikit-Learn.

Bibliografía:

1. Berry M. and Linoff G. “Data Mining Techniques”. John Wiley & Sonsa, 1997.
2. Bry X. “Analyses factorielles simples”, Ed. Economica, Paris, 1995.
3. Hastie, Tibshirani and Friedman. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer-Verlag, 2009.
4. Giudici Paolo. “Applied Data Mining: Statistical Methods for Business and Industry”. Wiley, 2005.
5. Dunhan M. “Data mining: Introductory and Advanced Topics”. Prentice Hall, 2002.
6. Han J. and Kamber M. “Data Mining: concepts and techniques”, Morgan Kaufman Publishers 2001.

7. Jambu M. “Introduccion au Data Mining: Analyse Intelligente des données”. Eyrolles, Paris, 1999.
8. Mirkin Boris. “Clustering for Data Mining, a data recovery approach”. Chapman & Hall. Boca Raton FL, 2005.
9. Rodríguez O, P.J.F. Groenen S. Winsberg and E. Diday. "I-Scal: Symbolic Multidimensional Scaling of Interval Dissimilarities". COMPUTATIONAL STATISTICS & DATA ANALYSIS the Official Journal of the International Association for Statistical Computing, London, 2006.
10. Andreas C. Müller and Sarah Guido. Introduction to Machine Learning with Python: A Guide for Data Scientists. O’Reilly, 1st Edition, 2017.
11. Dusty Phillips. Python 3 Object-oriented Programming, Second Edition. Packt Publishing Ltd, 2015.
12. Eric Matthes. Python Crash Course A Hands-On, Project-Based introduction to Programming. No Starch Press, Inc. 2016.
13. Jake VanderPlas. Python Data Science. O’Reilly, 2017.
14. John Paul Mueller (Author) and Luca Massaron. Python for Data Science For Dummies (For Dummies (Computer/Tech)) 1st Edition, 2015.
15. Steven F. Lott. Mastering Object-oriented Python. Packt Publishing Ltd, 2014.
16. Python Software Foundation. 2017. Python 3.6.2 documentation. python.org.
17. Anaconda 2017. Download Anaconda Distribution Python 3.6 version. Anaconda Inc.
18. Anaconda 2017. Anaconda Documentation. Anaconda Inc.
19. Richard, C. (2013). Learning R: A Step-by-Step Function Guide to Data Analysis. Sebastopol: OReilly.
20. Matloff, N. (2013). The art of R programming: A tour of statistical software design. San Francisco: No Starch Press.