Deloitte. Digital

uc3m | Universidad Carlos III de Madrid

Big Analytics:
de la información al conocimiento

Programa

Introducción
Arquitectura/Bases de datos relacionados /Infraestructuras (AMAZON)
R
Introducción al análisis masivo de datos: descriptivos y visualización de Big Data
Hackathon + Series Temporales
Aprendizaje supervisado de datos
Máquinas de vectores soportes (SVM) y algoritmos genéticos
Técnicas de regresión
Técnicas no supervisadas de análisis de datos
Introducción a Python
Spark Core +(parte 1)
Spark Core + SparkSQL (parte2) + Introducción a Spark MLlib
SparkStreaming
Deep learning
Técnicas de clasificación 2: Combinación de clasificadores
Modelos probabilísticos y gráficos
Hackathon + Presentación Final

Programa Big Analytics: de la información al conocimiento

Introducción
Introducción general al Big data y la necesidad del Analytics

Arquitectura/Bases de datos relacionados /Infraestructuras (AMAZON)
On premise vs Cloud. Soluciones Big Data en el Cloud. Diseño de sistemas inteligentes. Tipos de problemas que nos encontramos en data science y como abordarlos. Intro al aprendizaje automático. Caso práctico 1

Bases de datos. Introducción. Tratamiento, transformación y limpieza de datos. Caso práctico 2 y 3.

Bases de datos II. Obtención de datos, inferencia de datos y modelado de datos. Diseño de las necesidades del sistema. Casos prácticos 4 y 5.

R
1. Introducción a R.
2. Introducción a los paquetes de R que se utilizarán en otras sesiones.
3. Ejemplos prácticos.

Introducción al análisis masivo de datos: descriptivos y visualización de Big Data

En la primera clase, el objetivo es entender los datos que nos de Deloitte y enseñar que muchas veces la parte más tediosa y la que, en ocasiones, lleva más tiempo es preparar y entender los datos. Para ello, empezaré presentando las técnicas más básicas: histogramas, scatter plots, … y mostraré varios ejemplos que he trabajado en el pasado con los que estas técnicas tan rudimentarias fueron capaces de mejorar el detector facial de Viola-Jones o segmentar imágenes dermatológicas.

En la segunda clase, se avanzará al exploratory pursuit y al independent component analysis. Y se verá el análisis discriminante de Fisher. Se enseñara rápidamente que es el análisis factorial para explicar el concepto de dimensiones/rasgos latentes. Tras ello, concluiremos la segunda clase

entendiendo la visualización no para explorar sino para mostrar nuestros resultados.

Hackathon + Series Temporales

- Presentación del hackathon en el que van a poder aplicar las distintas técnicas que se vayan presentando a lo largo del curso
- Introducción a las series temporales con ejercicios teóricos/prácticos: definición de series temporales, descomposición de series temporales, series estacionarias y técnicas de modelización
- Primera prueba del Hackathon y subida de resultados a la plataforma. Formación de equipos
- En series temporales seguimos con una colección de ejercicios teóricos/prácticos: modelización de series temporales y visualización. Ejercicio final.
-

Aprendizaje supervisado de datos

Introducción al machine learning y tipos de problemas: supervisado vs no supervisado vs semi-supervisado, regresión vs clasificación… Algoritmos supervisados sencillos: métodos lineales (discriminante lineal), cuadráticos y no paramétricos (vecinos próximos). Aspectos importantes en el proceso de clasificación: Selección de características y reducción de la dimensión.elección del clasificador, problema de sobreajuste, validación.

Máquinas de vectores soportes (SVM) y algoritmos genéticos

Breve introducción a la optimización. Introducción a las máquinas de vectores soporte (SVM): motivación, optimización, kernel trick, ajuste de parámetros. Introducción a los algoritmos genéticos: cómo buscar en el espacio de soluciones, heurísticas, motivación de los algoritmos genéticos, metodología, tipos.

En todas las sesiones se motivarán los contenidos con ejemplos ilustrativos y reales en la medida de lo posible. Se harán prácticas de los distintos temas con R y se usará como hilo conductor el problema general de todo el curso.

Técnicas de regresión
Introducción de las principales técnicas de regresión: Lineal, Splines, Quantiles, Lasso y regressión logística.

Técnicas avanzadas de regresión, diseño de experimentos para mitigar problemas de causalidad: Diferencias en diferencias, Variables instrumentales y regresion por discontinuidad.

Técnicas no supervisadas de análisis de datos
Cluster Analysis: k-means (color quantization, pattern recognition examples), k-medoids (face recognition examples). Association Rules: The apriori algorithm (examples on Association rules sequences. The cspade algorithm (examples on tag recommendation, market basket, etc) market basket analysis), Association rules sequences. The cspade algorithm (examples on tag recommendation, market basket, etc)

Técnicas no supervisadas de análisis de datos
Cluster Analysis/ Hierarchical clustering : Agnes - Diana, Types of linkages . Examples on movie suggestion engines, cell phone towers placement, etc.

Introducción a Python
Conceptos básicos e introducción a la programación en Python, cubriendo las librerías más empleadas en el tratamiento de datos y en el desarrollo de modelos de machine learning (numpy, pandas, scikit-learn, etc.). En las prácticas se utilizarán Jupyter Notebooks para documentar el código y facilitar la ejecución interactiva durante la sesión

Spark Core +(parte 1)
Introducción a las funcionalidades básicas de Spark. Partiendo de la definición y manejo de RDDs hasta la manipulación de DataFrames y DataSets, pasando por las transformaciones y acciones más comunes en el procesamiento de datos distribuidos sobre Spark. Para ello se pueden utilizar distintas APIs y durante el curso se utilizará PySpark (de ahí la introducción de la sesión anterior), empleando además distintos formatos y fuentes de datos en el origen. Siguiendo un enfoque práctico, se aplicarán estos conceptos a ejemplos con datos reales de manera interactiva.

Spark Core + SparkSQL (parte2) + Introducción a Spark MLlib
Continuación de la sesión anterior, incorporando la librería de modelado MLlib de Spark. Se explicará cómo construir los algoritmos descritos en la primera sesión, esta vez en formato distribuido. A su vez se hará un repaso de todo aquello necesario en la construcción de features y de un pipeline completo de machine learning con PySpark.

SparkStreaming
En esta sesión se incorporará una componente de real time al sistema desarrollado durante las sesiones anteriores, haciendo uso de los modelos generados en "batch" para completar una arquitectura lambda. Para ello se hará una introducción al manejo de streams de datos y sus bloques de procesamiento mediante colas de Kafka. Utilizando como unidad básica los DStreams y modelos entrenados en batch veremos cómo utilizar SparkStreaming para hacer predicciones en real time.

Deep learning
Introducción a las redes neuronales básicas y a las redes profundas (deep learning) utilizando "Tensor Flow".
Técnicas de clasificación 2: Combinación de clasificadores

Técnicas de clasificación 2: Combinación de clasificadores

Modelos probabilísticos y gráficos
Introducción a los modelos gráficos probabilísticos. Caracterización. Redes Bayesianas. Tablas de probabilidad condicionada. Inferencia
Ejemplos Aprendiendo modelos gráficos probabilísticos e inferencia con ellos

Hackathon + Presentación Final
Trabajo por equipos & presentación de resultados