Guía docente
200645 - PBDE - Programación y Bases de Datos Estadísticas

Última modificación: 01/06/2023

Unidad responsable: Facultad de Matemáticas y Estadística
Unidad que imparte: 723 - CS - Departamento de Ciencias de la Computación.
707 - ESAII - Departamento de Ingeniería de Sistemas, Automática e Informática Industrial.

Titulación: MÁSTER UNIVERSITARIO EN ESTADÍSTICA E INVESTIGACIÓN OPERATIVA (Plan 2013). (Asignatura optativa).

Curso: 2023 **Créditos ECTS:** 5.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: ALEXANDRE PERERA LLUNA

Otros: Primer quadrimestre:
ALEXANDRE PERERA LLUNA - A
ORIOL RICART VILARRUBIAS - A

CAPACIDADES PREVIAS

Asignatura no obligatoria.
El estudiante ya ha desarrollado diversas capacidades estadisticas y/o de investigación operativa anteriormente.
Se requiere un nivel B2 (Cambridge First Certificate, TOEFL PBT >550) de inglés.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
3. CE-1. Capacidad para diseñar y gestionar la recogida de información, así como la codificación, manipulación, almacenamiento y tratamiento de esta información.
4. CE-4. Capacidad de utilizar los diferentes procedimientos de inferencia para responder preguntas, identificando las propiedades de los diferentes métodos de estimación y sus ventajas e inconvenientes, adaptados a una situación concreta y con un contexto específico.
5. CE-5. Capacidad para formular y resolver problemas reales de toma de decisiones en los diferentes ámbitos de aplicación sabiendo elegir el método estadístico y el algoritmo de optimización más adecuado en cada ocasión.
6. CE-6. Capacidad para utilizar el software más adecuado para realizar los cálculos necesarios en la resolución de un problema.
7. CE-7. Capacidad para comprender artículos de estadística e investigación operativa de nivel avanzado. Conocer los procedimientos de investigación tanto para la producción de nuevos conocimientos como para su transmisión.
8. CE-8. Capacidad de discutir la validez, el alcance y la relevancia de estas soluciones y saber presentar y defender sus conclusiones.

Transversales:
2. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisición, la estructuración, el análisis y la visualización de datos e información en el ámbito de especialidad, y valorar de forma crítica los resultados de dicha gestión.

10. TERCERA LENGUA: Conocer una tercera lengua, preferentemente el inglés, con un nivel adecuado oral y escrito y en consonancia con las necesidades que tendrán los titulados y tituladas.

11. TRABAJO EN EQUIPO: Ser capaz de trabajar como miembro de un equipo interdisciplinar, ya sea como un miembro más o realizando tareas de dirección, con la finalidad de contribuir a desarrollar proyectos con pragmatismo y sentido de la responsabilidad, asumiendo compromisos teniendo en cuenta los recursos disponibles.

METODOLOGÍAS DOCENTES

El curso está dividido en 2 módulos que se imparten de forma sucesiva. Cada módulo consta aproximadamente de la mitad de las sesiones. Todas las clases son teórico-prácticas y en ellas el profesorado presenta y discute los conceptos básicos de cada módulo. El material de soporte que se utilizará será publicado con anterioridad en Atenea (guía docente, contenidos, transparencias del curso, ejemplos, programación de actividades de evaluación, bibliografía,...).

El estudiante deberá dedicar las horas de aprendizaje autónomo al estudio de los temas del curso, ampliación bibliogràfica y seguimiento de las prácticas de laboratorio.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

En este curso presentan y discuten herramientas y técnicas para preparar a los estudiantes a la ciencia de los datos. Los principales conceptos introducidos en clase abarcarán herramientas y métodos para el almacenamiento y análisis de datos, incluyendo bases de datos relacionales, noSQL y distribuidas, computación científica, "machine learning" aplicado y "deep learning" con Python. También se estudiaran Scala y Spark. El curso consta de dos módulos principales.

MÓDULO 1:
El primer módulo cubrirá un curso intensivo de python científico para el análisis de datos. Este curso incluirá cuatro puntos:
* Introducción al lenguaje Python como una herramienta. ipython, ipython notebook (jupyter), tipos básicos, mutabilidad e inmutabilidad y programación orientada a objetos.
* Breve introducción a Python numérico y matplotlib para visualización gráfica.
* Introducción a los kits científicos para el análisis de datos con mchinelearning. Análisis de componentes principales, clustering y análisis supervisado con datos multivariados.
* Introducción al Deep Learning con Python.

MÓDULO 2:
Presentamos el lenguaje Scala y la arquitectura Spark.
* Scala como un lenguaje funcional y las colecciones de Scala.
* Spark y RDD (Resilient Distributed Data Sets).
* Spark y SQL.
* Introducción a MLlib.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas aprendizaje autónomo	80,0	64.00
Horas grupo grande	30,0	24.00
Horas grupo pequeño	15,0	12.00

Dedicación total: 125 h

CONTENIDOS

Introducción a Python

Descripción:
a. 'Por qué Python?
b. Historia de Python
c. Instalación de Python
d. Recursos de Python

Dedicación: 1h
Grupo grande/Teoría: 1h

Trabajar con Python

Descripción:
a. Flujo de trabajo
b. Ipython vs CLI
c. Editores de texto
d. IDEs
e. Notebook

Dedicación: 1h
Grupo grande/Teoría: 1h

Primeros pasos con Python

Descripción:
a. Introducción
b. Obteniendo ayuda
c. Tipos básicos
d. Mutable y mutable
e. Operador de asignación
f. Control del flujo de ejecución
g. Manejo de excepciones

Dedicación: 1h
Grupo grande/Teoría: 1h

Funciones y Programación Orientada a Objetos

Descripción:
a. Definición de funciones
b. Entrada y salida
c. Biblioteca Estándar
d. Programación orientada a objetos

Dedicación: 1h
Grupo grande/Teoría: 1h

Introducción a NumPy

Descripción:
a. Visión de conjunto
b. Matrices
c. Operaciones en arrays
d. Arrays avanzados (ndarrays)
e. Notas sobre el rendimiento (\%timeit en ipython)

Dedicación: 2h
Grupo grande/Teoría: 2h

Matplotlib

Descripción:
a. Introducción
b. Figuras y subplots
c. Ejes y control adicional de las figuras
d. Otros tipos de gráficos
e. Animaciones

Dedicación: 2h
Grupo grande/Teoría: 2h

Introudcción a Pandas

Descripción:
contenido castellano

Dedicación: 2h
Grupo grande/Teoría: 2h

Scikits de Python

Descripción:
a. Introducción
b. scikit-timeseries

Dedicación: 1h
Grupo grande/Teoría: 1h

scikit-learn

Descripción:
a. Conjuntos de datos
b. Generadores de muestras
c. Aprendizaje no supervisado
d. Aprendizaje supervisado
i. Análisis Discriminante Lineal y Cuadrático
ii. Vecinos más cercanos
iii. Máquinas de soporte vectorial (Support Vector Machines)
e. Selección de características

Dedicación: 8h
Grupo grande/Teoría: 8h

Introducción práctica a Scikit-learn

Descripción:
a. Resolver un problema de caras principales (eigenfaces)
i. Objetivos
ii. Descripción de los datos
iii. Clases iniciales
iv. Importación de datos
b. Análisis no supervisado
i. Estadísticas descriptivas
ii. Análisis de componentes principales
iii. Clustering
c. Análisis supervisado
i. K-Vecinos más cercanos
ii. Clasificación con soporte vectorial
iii. Validación cruzada

Dedicación: 5h 30m
Grupo grande/Teoría: 5h 30m

Introducción a Zeppelin, Scala y Programación Funcional

Descripción:
a. Inmutable y Mutable
b. Listas y mapas, filtros, reducciones
c. Map reduce
d. Otras colecciones, Streams

Dedicación: 5h
Grupo grande/Teoría: 5h

Arquitectura Spark y Spark Core

Descripción:
a. Arquitectura Spark: en particular, Spark Core
b. Contexto de chispa
c. Tipos de operaciones: transformaciones y acciones
d. RDD: Conjuntos de Datos Distribuidos Resistentes
e. Clausura de una función

Dedicación: 5h
Grupo grande/Teoría: 5h

Spark: MLlib

Descripción:
a. Descripción del MLlib.
b. Labeled Points y features
c. Ejemplo de regresión lineal

Dedicación: 5h
Grupo grande/Teoría: 5h

Spark SQL

Descripción:
a. Lectura de un archivo.
b. Spark Data Frame.
do. Selección, filtros, agrupamiento, clasificación.
re. Operaciones de ventana
do. SQL

Dedicación: 7h 30m
Grupo grande/Teoría: 7h 30m

SISTEMA DE CALIFICACIÓN

- 1/4 Examen escrito del primer módulo.
- 1/4 Examen escrito del segundo módulo.
- 1/2 Práctica final en bases de datos grandes que integran conceptos de ambos módulos.

BIBLIOGRAFÍA

Básica:
- Zaharia, M.; Karau, H.; Konwinski, A.; Wendell, P. Learning spark lightning-fast big data analysis. 2015. O'Reilly Media, ISBN 9781449358624.
- Swartz, Jason. Learning scala: practical functional programming for the JVM [en línea]. 2014. O'Reilly Media, [Consulta: 28/06/2023]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=1888253. ISBN 9781449367930.
- Langtangen, H.P. A Primer on scientific programming with Python [en línea]. Springer, 2011 [Consulta: 28/06/2023]. Disponible a: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-3-642-30293-0. ISBN 9783642183652.
- Shapiro, B.E. Scientific computation: Python hacking for math junkies. Sherwood Forest Books, 2015. ISBN 9780692366936.
- Baumer, Benjamin; Kaplan, Daniel; Horton, Nicholas. Modern data science in R. Primera. Boca Raton: CRC, 2017. ISBN 9781498724487.

Complementaria:
- Spector, P. Concepts in computing with data (Stat 133, UC Berkeley) [en línea]. Berkeley, 2011 [Consulta: 28/06/2023]. Disponible a: http://www.stat.berkeley.edu/~s133/.