Presentación de la Asignatura Tecnologías para el Análisis de Datos Masivos

Máster de Análisis de Datos Masivos UIB: Juan Gabriel Gomila & Ricardo Alberich

17/10/2020

¿Quiénes somos?

Figure 1: Juan Gabriel Gomila

Figure 2: Ricardo Alberich

¿Quiénes somos? Juan Gabriel Gomila

- Departamento de Ciencias Matemáticas e Informática e la UIB
- Profesor asociado del área de Ciencia de la Computación e Inteligencia artificial
- Licenciado en Matemáticas por la UIB
- CEO Frogames
- Y más cosas...
- Email

¿Quiénes somos? Ricardo Alberich

- Departamento de Ciencias Matemáticas e Informática e la UIB
- Profesor Titular del área de Ciencia de la Computación e Inteligencia Artificial
- Licenciado en matemáticas por la Universidad de Valencia
- Doctor en informática por la UIB
- Email

Asignatura 11630 - Tecnologías para el Analisis de Datos Masivos

- Guía docente(català)
- Cronograma: Horarios de clase
- Espacio discord de la asignatura
- Espacio moodle de la UIB de la asignatura

Temas de la asignatura

Todo será de carácter práctico y aplicado. Ya se profundizará en otras asignaturas según los itinerarios que hayáis elegido.

Grandes temas (no necesariamente en este orden) son tecnologías para:

- Tema 1 Repaso de estadística descriptiva, inferencia y Introducción a R y RStudio Este será un tema transversal al curso cubierto con los materiales del curso online de Udemy los anexos.
- Tema 2. Manipulación de datos con Tidyverse. Este tema se tratará sólo en R pero se extensible al lenguaje python
- Tema 3. Machine Learning y Aprendizaje Estadístico
- Tema 4 Protección de datos y Legislación.

Contenidos Tema 3 (y 1)

- Parte 1 - Preprocesamiento de datos
- Parte 2 - Regresión: Regresión Lineal Simple, Regresión Lineal Múltiple, Regresión Polinomial, SVR, Regresión en Árboles de Decisión y Regresión con Bosques Aleatorios
- Parte 3 - Clasificación: Regresión Logística, K-NN, SVM, Kernel SVM, Naive Bayes, Clasificación con Árboles de Decisión y Clasificación con Bosques Aleatorios
- Parte 4 - Clustering: K-Means, Clustering Jerárquico
- Parte 5 - Aprendizaje por Reglas de Asociación: Apriori, Eclat

Contenidos Tema 3 (y 2)

Continuación

- Parte 6 - Reinforcement Learning: Límite de Confianza Superior, Muestreo Thompson
- Parte 7 - Procesamiento Natural del Lenguaje: Modelo de Bag-of-words y algoritmos de NLP
- Parte 8 - Deep Learning: Redes Neuronales Artificiales y Redes Neuronales Convolucionales
- Parte 9 - Reducción de la dimensión: ACP, LDA, Kernel ACP
- Parte 10 - Selección de Modelos & Boosting: k-fold Cross Validation, Ajuste de Parámetros, Grid Search, XGBoost

Materiales del curso (enlaces gratuitos a materiales)

- Curso completo de Estadística descriptiva - RStudio y Python
- Curso completo de R para Data Science con Tidyverse
- Machine Learning de la A a la Z: R y Python para data science

Enlaces a manuales cursos y libros de Análisis de Datos

Manuales de R básicos

1. AprendeR1. Manual de R parte 1. Descriptiva.
2. AprendeR2 version 2. VERSION 2 Manual de R parte 2. Inferencial.
3. AprendeR2 version 1. VERSION 2 Manual de R parte 2. Inferencial.

Enlaces a repositorios de datos

1. Kaggle
2. Sciencie (Nature)
3. Data Repositories
4. UCI (Machine learningrepository)

Material cursos probabilidad y estadística básicos

1. Probabilidad y variables aleatorias para ML con R y Python bookdown
   - Presentaciones
   - Presentaciones repositorio
   - Ejercicios, en Rmd, pdf
   - Shiny
2. Curso completo de estadística inferencial con R y Python bookdown. EN CONSTRUCCIÓN
   - Presentaciones
   - Presentaciones repositorio
   - Shiny intervalos de confianza y contrastes hipótesis. EN CONSTRUCCIÓN
3. Página de itinerarios del profesor Juan Gabriel Gomila

Libros del entorno de R y Rstudio

En general los libros gratuitos de (https://bookdown.org/)

Dos libros, entre otros, básicos que introducen a las librerías tidyverse del entorno de Rstudio son

1. R for Data Science
2. Tidy text
3. Hay muchos más libros en bookdown.org

Metodología

- Teoría y materiales explicados en vídeos online. (Home work)
- Las clase en linea en el discord de la asignatura y BBcollaborate de moodle de la UIB
- En las clases online, repasaremos conceptos de los videos y haremos más ejemplos, profundizaremos y resolveremos consultas y dudas,
- **Recomendación** las primeras semanas sólo tenéis esta asignatura habrá MUCHO TRABAJO y tendréis que invertir muchas horas (15 horas de clase y unas 45 más de trabajo). El objetivo es conseguir un nivel MÍNIMO para poder cursar el resto de asignaturas del máster con garantías de éxito.

Actividad de evaluación

- Estadística descriptiva: R + python: 20%
- Manipulación de Datos: Tidyverse: 20%
- Trabajo Final (individual): Machine Learning : 50% nota mínima 5/10) recuperable.
- Cuestionarios, interacción en clase y trabajos cursos online: 10%