laSalle
UNIVERSITAT RAMON LLULL

Estadística

Presentación de la Asignatura

Curso 2020-2021

laSalle
UNIVERSITAT RAMON LLULL

Estadística
Máster Universitario en Ingeniería de Datos Masivos (Big Data)
Curso 2020-2021

DESCRIPCIÓN DE LA ASSIGNATURA

En esta asignatura se imparten conceptos básicos de estadística que permitirán al alumno obtener los fundamentos de la analítica de datos. La asignatura cubre diferentes aspectos de introducción a la inferencia estadística, clasificación y predicción mediante técnicas de Machine Learning con una visión aplicada al tratamiento de grandes volúmenes de datos. Se abordan distintas técnicas de modelado y *clustering* haciendo hincapié en la interpretación de los resultados, en proporcionar medidas de incertidumbre y en su utilidad en la toma de decisiones. Durante toda la asignatura se usará el paquete estadístico R.

El objetivo de esta asignatura es proporcionar al alumno las capacidades para poder extraer información eficiente de un conjunto de datos de medio/gran volumen. Por lo tanto, el alumno sabrá realizar e interpretar una descriptiva de un conjunto de datos, será capaz de construir modelos de respuesta continua y binaria interpretando correctamente los coeficientes resultantes, conocerá distintas técnicas para hacer agrupaciones de instancias en base a sus similitudes y, además, sabrá aplicar distintos algoritmos de *machine learning* conociendo las fortalezas y debilidades de cada uno de ellos.

Los resultados de aprendizaje de esta asignatura son:

- **ADM_RA5**: El alumno conoce los conceptos de análisis estadístico relacionados con la inferencia, regresión, *clustering* para grandes volúmenes de datos.
- **ADM_RA6**: El alumno es capaz de diseñar y programar análisis complejos mediante una tecnología de análisis estadístico disponible en el mercado.

CONTENIDOS

1. *Introducción a la inferencia estadística*
2. *Clustering no supervisado*
3. *Clustering supervisado*
4. *Visualización de datos y paquetes de R*

ORGANIZACIÓN DE LA ASIGNATURA

Profesor titular: Jordi Cortés (jcortes@salle.url.edu)

Horario lectivo

	Lunes	Martes	Miércoles	Jueves	Viernes
Semana del 19 de octubre de 2020		**19:00 – 22:00**		**19:00 – 22:00**	
Semana del 26 de octubre de 2020		**19:00 – 22:00**		**19:00 – 22:00**	
Semana del 9 de noviembre de 2020		**19:00 – 22:00**		**19:00 – 22:00**	
Semana del 16 de noviembre de 2020		**19:00 – 22:00**		**19:00 – 22:00**	
Semana del 23 de noviembre de 2020		**19:00 – 22:00**		**19:00 – 22:00**	

CONTENIDOS DETALLADOS

1. Introducción a R **(3 horas)**
   - 1.1. Visión general
     - 1.1.1. Instalación de R y RStudio
     - 1.1.2. Interfaz de RStudio
     - 1.1.3. Primeros pasos: instrucciones, objetos, funciones, instalar paquetes, ayuda, cierre de la sesión.
   - 1.2. Organización de la información
   - 1.3. Acceso y modificación de datos
   - 1.4. Importación y exportación de datos
     - 1.4.1. Lectura
     - 1.4.2. Escritura
   - 1.5. Estadística descriptiva numérica y gráfica
     - 1.5.1. Tipos de variables
     - 1.5.2. Descriptiva univariante
       - 1.5.2.1. Variables categóricas: tablas y diagramas de barras
       - 1.5.2.2. Variables numéricas: estadísticos, histogramas y diagrama de barras
     - 1.5.3. Descriptiva bivariante
       - 1.5.3.1. Categórica vs Categórica: Tablas y diagramas de mosaico
       - 1.5.3.2. Categórica vs numérica: estadísticos y boxplots estratificados
       - 1.5.3.3. Numérica vs numérica: correlación y diagramas de dispersión
     - 1.5.4. Datos ausentes: "missings"
2. Introducción a la Inferencia Estadística **(6 horas)**
   - 2.1. Nociones básicas
     - 2.1.1. Definiciones de población, muestra, observaciones, parámetro, estadístico y estimador
     - 2.1.2. Muestreo aleatorio simple
     - 2.1.3. Esperanza y varianza de una variable aleatoria
     - 2.1.4. Funciones de probabilidad y de distribución: Normal, t-Student, Chi-cuadrado, F de Fisher. Cálculo de probabilidades
     - 2.1.5. Distribución de la media muestral. Teorema central del límite
     - 2.1.6. Distribución de la varianza muestral
     - 2.1.7. Distribución de la proporción muestral
   - 2.2. Estimación de un parámetro. Puntual y por intervalo
     - 2.2.1. Estimación de una media
     - 2.2.2. Estimación de una varianza
     - 2.2.3. Estimación de una probabilidad
   - 2.3. Comparación de 2 poblaciones
     - 2.3.1. Comparación de 2 medias
       - 2.3.1.1. Independientes
       - 2.3.1.2. Apareadas
     - 2.3.2. Comparación de 2 varianzas
     - 2.3.3. Comparación de 2 proporciones
   - 2.4. **Ejemplo práctico**
3. Modelos de Regresión **(6 horas)**
   - 3.1. Regresión lineal simple y múltiple
     - 3.1.1. Especificación del modelo
     - 3.1.2. Estimación puntual i por intervalo de los parámetros de la recta
     - 3.1.3. Validación del modelo. Análisis de residuos
     - 3.1.4. Predicción para la el valor esperado y una observación puntual
     - 3.1.5. Capacidad predictiva: R^2
     - 3.1.6. **Ejemplo práctico**
   - 3.2. Regresión logística
     - 3.2.1. Estimación del modelo
     - 3.2.2. Estimación puntual i por intervalo de los parámetros del modelo

   3.2.3. Validación del modelo. Bondad del ajuste
   3.2.4. Predicción de la probabilidad de un evento.
   3.2.5. Capacidad predictiva: Curva ROC y AUC
   3.2.6. **Ejemplo práctico**
4. Modelos de Clustering no supervisado **(4 horas)**
   4.1. K-means
      4.1.1. Algoritmo
      4.1.2. Elección del número de grupos
      4.1.3. Convergencia
      4.1.4. Concepto de Inercia
      4.1.5. Implementación en R
      4.1.6. Medidas de rendimiento
   4.2. Clusterización jerárquica
      4.2.1. Propiedades de distancias
      4.2.2. Tipos de distancias
      4.2.3. Tipos: aglomerativos, divisivos
      4.2.4. Implementación en R
   4.3. Anexo: Reducción de dimensionalidad (ACP)
   4.4. Ejemplo práctico
5. Modelos de Clustering supervisado **(8 horas)**
   5.1. K-Nearest-Neighbors
   5.2. Naive Bayes
   5.3. Árboles condicionales
   5.4. Random-forest
   5.5. SVM
   5.6. **Ejemplo práctico**
6. Shiny **(3 horas)**
   6.1. Introducción
   6.2. UI
      6.2.1. Estructura
      6.2.2. Widgets
   6.3. SERVER
      6.3.1. Estructura
      6.3.2. Outputs
   6.4. ¿Dónde "colgar" las apps?
   6.5. **Ejemplo práctico**
7. Paquetes de R aplicados al big data **(3 horas)**
   7.1. Alternativa a data.frame: data.table
   7.2. Grandes volúmenes: biglm, bigMemory y bigAnalytics
   7.3. Paralelizar: foreach, snowfall
   7.4. Web scrapping: rvest i Rcurl
   7.5. Reference card

BIBLIOGRAFIA BÁSICA

- James G, Witten D, Hastie T, et al. An introduction to Statistical. Learning with applications in R. First Edition: Springer, 2013.
- Wickham, H. & Grolemund, G. *R for Data Science* (O'Reilly, 2016); http://r4ds.had.co.nz/
- B. Efron and T. Hastie. Computer age statistical inference, volume 5. Cambridge University Press, 2016.
- Chang, W. (2013). *R graphics cookbook.* O'Reilly

METODOLOGIA

Las clases presenciales combinan un 30% de material teórico impartido con el soporte de diapositivas; un 40% de trabajo guiado con ayuda de scripts de R; y un 30% de trabajo individual o en parejas sobre un problema práctico planteado por el profesor. A los alumnos se les proporcionará tanto las diapositivas como los códigos de análisis (completos o incompletos) para presenciar las clases.

SISTEMA DE EVALUACIÓN GLOBAL

- La nota final de la asignatura se computa como la media ponderada de 3 prácticas que los alumnos deberán realizar durante el curso sobre las siguientes temáticas:
  - Modelo lineal (20%)
  - Modelo de respuesta binaria (30%)
  - Clustering (50%)
- Se recomienda realizar las prácticas en parejas.
- Todas las actividades de evaluación de esta asignatura se consideran actividades altamente significativas según la normativa académica (https://www.salleurl.edu/es/normativa-de-copias). Por lo tanto, las copias totales o parciales en cualquier actividad evaluable, se penalizarán con el que está establecido en la normativa académica, tanto en la fuente como la copia sin excepción.
- Cualquier interacción vía correo electrónico con el personal asociado a la asignatura (profesores de teoría, prácticas, monitores, etc.) se realizará estrictamente desde la dirección de correo de la escuela (@ students.salle.url.edu). No se responderá a ninguna dirección de correo ajena a la escuela.