TADM 20_21. MADM. Trabajo Final

Nombre y Apellidos

2021

Contenidos

1 Trabajo final 50% nota final (se puntúa sobre 100) **1**
1.1 Instrucciones 1

2 ¿Qué es lo que se pide? **1**
2.1 Parte 1 Contexto del problema y modelo de datos (30/100) 1
2.2 Parte 2: Análisis exploratorio (EDA). (30/100) 2
2.3 Parte 3: Aprendizaje estadístico (Machine Learning) (40/100) 2

1 Trabajo final 50% nota final (se puntúa sobre 100)

El proyecto final consiste en elegir un `data.set` (buscar en las fuentes que hemos utilizado a lo largo del curso) que consista en:

- Una o más tablas de datos (se valorará más de dos tablas, pero no es necesario)
- El conjunto de datos global debe contener variables de varios tipos: númericas, factores, de texto, de tiempo... etc.
- Se tendrá en cuenta la justificación del código mostrándolo para su evaluación, su claridad, la contextualización del problema así como la investigación adicional de librerías no vistas durante el curso.

1.1 Instrucciones

- Entregad de forma individual.
- Se puede hacer con R o python.
- Hay que entregar el Rmd/notebook junto con su salida en html/pdf
- Máxima longitud: 10 páginas en pdf. Aproximadamente
- Hay que cuidar la presentación, ortografía y redacción.
- Fecha entrega 23:55 horas 5 de febrero (viernes) de 2021 a través de la actividad de moodle correspondiente.

2 ¿Qué es lo que se pide?

El proyecto de evalúa en tres partes.

2.1 Parte 1 Contexto del problema y modelo de datos (30/100)

1. Contextualiza a partir de la información del origen de los datos de que disponemos. Qué datos contiene cada uno de los ficheros y para qué nos pueden resultar importantes en nuestro proyecto. Construir el modelo de datos para vuestro problema.

2. Realiza la limpieza de datos, si es necesario, y validación de las variables (análisis de valores perdidos, valores nulos, valores aberrantes, datos fuera de rango...)

3. Justifica para cada una de las variables de vuestro proyecto el tipo de dato que mejor se ajusta a cada una de ellas: numérico, ordinal, categórico.... Transforma cada variable al tipo adecuado de dato.

2.2 Parte 2: Análisis exploratorio (EDA). (30/100)

En las siguientes preguntas aplica todo lo que hemos visto acerca de la documentación en el EDA: Título de gráficos, etiquetas de los ejes, coloreado con información, leyendas, tablas bien presentadas (knitr)... Establece en dos o tres etapas un Análisis Exploratorio (EDA) global de estos datos que incluya algunos de estos aspectos: 1. Gráficos con ggplot2 de las v.a. discretas, continuas 2. Gráficos de dos variables segmentando una en función de otras, que tenga sentido. Por ejemplo continua versus factor. 3. Tablas de resumen de estadísticos globales. 4. Tablas de resumen de estadísticos por grupos. 5. Ampliación de alguna técnica de inferencia estadística.

2.3 Parte 3: Aprendizaje estadístico (Machine Learning) (40/100)

En esta última parte aplica algunas de las técnicas de Machine Learning que hemos visto.

En cada caso de machine learning, si es necesario, divídelos en conjuntos de datos de **training** y de **testing**.

Sobre todo desde el punto de vista de la parte de la tecnología de aplicación.

1. Para alguna de las variables continuas, aplica al menos dos de los algoritmos de regresión. Se valorará tener el modelo, su explicación, así como una representación gráfica del resultado.
2. Para alguna de las variables discretas, aplica al menos dos de los algoritmos de clasificación. Se valorará tener el modelo, su explicación, así como una representación gráfica del resultado.
3. Reduce la dimensión de tu dataset con LDA o PCA (puede resultar interesante para reducir el número de variables a 2 o 3 para así representarlo en uno, dos o tres gráficos del plano).
4. Crea un clustering por k-means o jerárquico utilizando como entrada resultado de la reducción de la dimensión anterior, para segmentar las observaciones en un número de grupos. Se valorará tener el modelo, su explicación, así como una representación gráfica del resultado.

- BONUS: utiliza tensorflow como librería de redes neuronales para resolver el problema desde un punto de vista diferente al de la estadística clásica.
- BONUS 2: utiliza alguno de los algoritmos que hayáis aprendido durante el resto del máster, justificando el por qué utilizáis ese algoritmo.