Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2020 - S2

02 de Febrero, 2021

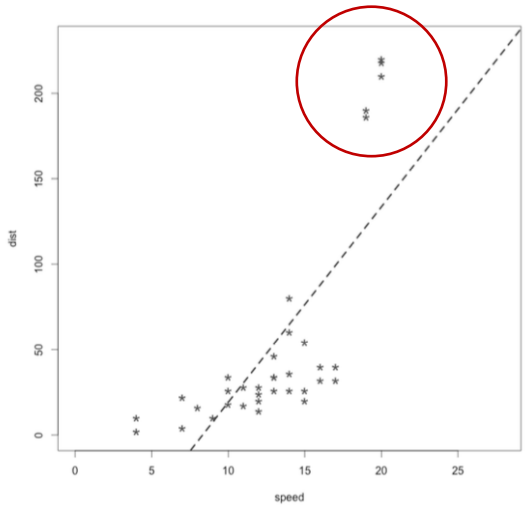
1.

a) ¿Qué es un outlier?

b) ¿Qué tratamiento se le debe dar a este tipo de valores?

c) Suponga que tiene un dataset con 100 observaciones y sabe que la distribución de valores en la muestra es normal con una media de 50 y una distribución estándar de 5, ¿calificaría alguno de los siguientes valores como outlier?: 20, 29, 32, 41, 45, 55, 58, 62, 67, 74. ¿Cuáles? Justifique.

d) ¿Tendría algún impacto en el análisis quitar los outliers del siguiente dataset? Justifique.



2. ¿Por qué es necesaria la ingeniería de atributos?

3. Luego de generar predicciones con un modelo predictivo usted nota una gran diferencia de performance entre su set de entrenamiento y su set de validación. ¿Qué fenómeno es este? ¿Cómo puede evitarlo?

4. ¿Es mejor tener demasiados falsos positivos (error tipo I) o demasiados falsos negativos (error tipo II)?. ¿Que hubiese priorizado minimizar en el trabajo obligatorio del curso? Justifique.

5. Una empresa "A" fue contratada por una empresa inmobiliaria "B" para realizar un proyecto de Data Science. La misma al finalizar el proyecto, entre otras cosas, entrega un modelo predictivo que tiene una precisión del 90%. Mucho tiempo ha pasado desde esto y la empresa "B" ha crecido significativamente aumentando su número de clientes. La misma ha experimentado una disminución progresiva de performance del modelo que le fue entregado. ¿Por qué usted cree que esto ha ocurrido? ¿Qué cree usted que debe hacerse?

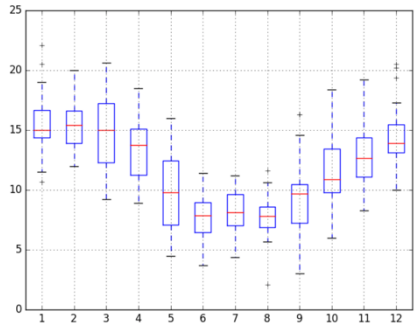
6. ¿En qué situaciones se debe utilizar una regresión logística vs. una regresión lineal? ¿Cómo se interpretan los coeficientes de la regresión (ß) y el resultado (y) en una y en otra? Proporcione un ejemplo de caso de uso para ambos modelos.

7. Suponga un modelo de regresión lineal que estima el precio de un auto dado su kilometraje.

a) ¿Qué significa que el coeficiente ß de la variable independiente tenga un valor-p > 0.1?
b) ¿Qué información nos agregan los intervalos de confianza asociados a los coeficientes estimados?
c) Suponga que al agregar el año del auto a la regresión como variable independiente el coeficiente ß del kilometraje cambia vs. la regresión anterior. ¿Era esperable? ¿A qué se debe este cambio?

8.

a) Suponga que está analizando la venta de helados en Uruguay en un período de 30 años. Su colega le entrega el siguiente gráfico, con los meses en el eje X y las unidades promedio vendidas en el eje Y. ¿Qué componente de la serie de tiempo podría analizar a partir del gráfico? ¿Qué conclusiones se desprenden?



b) Si le dijeran que año tras año el consumo de helados en Uruguay ha aumentado sistemática, ¿qué componente de la serie de tiempo esperaría ver afectado? ¿Cuál sería el impacto del componente? ¿Cómo se podría modelar?

c) Si supiera que cada 10 años hay una crisis económica que afecta directamente la venta de helados, ¿qué componente de una serie de tiempo estaría analizando? ¿Cómo se podría modelar?

d) Suponga que tiene un modelo que predice la venta de helados del próximo mes, ¿en qué componente entraría el efecto de una pandemia mundial como el COVID-19? (antes de poder modelarla).

9. Explique los problemas asociados al clustering de datos poniendo ejemplos que ilustren los problemas.

10. Explique cómo determinaría la similitud entre los libros de una biblioteca incluyendo el pre-procesamiento necesario.