Tarea 3 - Introducción a la Ciencia de Datos

Junio de 2020

1 Problema 1

Para este problema utilizaremos el conjunto de datos Aquienvoto.uy, que se entregará adjunto. Este dataset contiene los resultados de una encuesta de 20 preguntas realizada en 2019, que intentan capturar ideologías políticas, con el objetivo de predecir qué precandidato votaba la persona, a partir de la similitud de sus respuestas con las de otros encuestados. Como el sitio permitía, además, ingresar el candidato que la persona realmente pensaba votar, el dataset contiene las respuestas y el candidato elegido.

Para este trabajo, se entrega el dataset, y un mapeo de los id de los candatos a su nombre, y el partido al que pertenecen.

Se pide:

a) Preprocese los datos para agregar a cada instancia el partido político del candidato.

b) Utilice Principal Component Analysis para visualizar los datos del dataset. Diferencie las respuestas de votantes de diferentes candidatos o partidos. ¿Puede decirse que las preguntas agrupan de alguna forma a los votantes de partidos? ¿Qué otras conclusiones podría sacar de los datos visualizados?

c) Agrupe las instancias utilizando k-means, intentando con, al menos, 2, 3 y 11 clusters y visualice los resultados. Comente.

d) Utilice SVM para intentar predecir, a partir de las respuestas, el partido al que pertenece el encuestado. Divida el dataset de forma adecuada en entrenamiento y evaluación. Ajuste los parámetros del método utilizando validación cruzada. Reporte los valores de accuracy, precisión, recall, y medida F1 sobre el conjunto de evaluación. Muestre la matriz de confusión. Comente los resultados.

La entrega deberá realizarse utilizando un Notebook Jupyter. Pueden utilizar las bibliotecas de su preferencia en Python o R, pero deberán documentar claramente en el informe los pasos realizados en el ánalisis, las decisiones tomadas y los resultados obtenidos.

2 Problema 2

Considere el código que acompaña esta tarea, que implementa una versión de clustering del estilo *k-means*, utilizando descenso por gradiente.

a) Ejecute el código, buscando los motivos por los que no funciona correctamente, y corríjalo. Describa y explique lo sucedido.

b) ¿Cuál es la función de costo (o ajuste a datos) que se está considerando? Interprete sus elementos.

c) Utilizando un set de datos generados fijos, ejecute varias veces el código, y compare los resultados obtenidos. Explique lo sucedido.

3 Problema 3

a) Elija un juego de datos. Describa sus características y sus posibles problemas de calidad.

b) ¿Qué problema se puede resolver utilizando las herramientas presentadas en el curso?[1]

c) Describir el proceso que seguiría para llegar a responder lo indicado en el item anterior. Por ejemplo, qué métodos se podrían aplicar, con qué cuidados, qué visualizaiones usaría, tanto para explorar los datos como para explicar resultados. Describa detalladamente todos los elementos del proceso.

[1] Puede ser por ejemplo, clasificar textos automáticamente de acuerdo al autor, detectar enfermadades en imágenes médicas, agrupar clientes de acuerdo a sus preferencias de compra, etc.