Master of Science in Data Science

Métodos de aprendizaje de máquinas en Data Science

Tarea 1

Clusters espaciales

El archivo dataTaxi.csv contiene la información de 452,166 viajes de radiotaxi de una empresa chilena. Los datos tienen 10 variables dadas por:

- **LatitudPAB:** Coordenada de la latitud cuando un pasajero se sube a bordo del taxi.
- **LongitudPAB:** Coordenada de la longitud cuando un pasajero se sube a bordo del taxi.
- **LatitudOM:** Coordenada de la latitud cuando un pasajero se baja del taxi.
- **LongitudOM:** Coordenada de la longitud cuando un pasajero se baja del taxi.
- **Horas:** Hora de inicio del viaje
- **DOW**: Dia de la semana
- **Hora_Inicio**: Hora de cuando inicia el viaje
- **Hora_Fin**: Hora de cuando finaliza el viaje
- **DistKilometros**: Distancia en kilómetros recorridas por el taxi.

Como Santiago está subdividido en varias comunas, la empresa quiere analizar si existe un patrón común entre los viajes. Por lo mismo, el objetivo principal de esta tarea es analizar los datos, obtener clusters representativos de los datos, y crear un nuevo proceso de clusterización basado en k-means.

Problema 1) Realice una limpieza y selección de variables de los datos. Justifique cada paso realizado. Atención, los datos seleccionados para entrenamiento en este proceso serán definitivos para el resto de los Problemas. A la vez, variables descartadas en este problema si pueden ser utilizadas en los problemas donde se tiene que explicar los clusters encontrados.

Problema 2) Aplique alguno de los algoritmos visto en clases para determinar clusters de trayectoria.

Problema 3) Explique los clusters encontrados en el problema 2) para una persona que no entiende lo que es un proceso de clusterización.

Problema 4) Modifique k-means para que una vez que sea aplicado con un valor dado de k, analice cada cluster y, según alguna regla definida por usted, determine que clusters están mal definidos/incorrectos. Posteriormente, cada cluster incorrecto sepárelo en dos clusters y vuelva a correr k-means con el nuevo número de clusters y centroides. Repita este proceso hasta que todos los clusters encontrados sean considerados correctos.

Problema 5) Explique los clusters encontrados en el problema 4) para una persona que no entiende lo que es un proceso de clusterización.

Problema 6) Compare los clusters encontrados en el Problema 2 y Problema 4. Posteriormente, justifique cual de los métodos seleccionaría para una implementación dentro de la empresa.

Cada problema tendrá un punto. La fecha de entrega es el 14 de Septiembre a las 23:59 horas.

La tarea se puede realizar hasta en grupos de 2 personas.

El entregable final puede ser:

- A través de webcursos => un archivo jupyter notebook comentando todo los procesos realizados. Caso contrario, se deberá entregar un archivo .py y un informe en formato pdf.

- A través de GitHub => el ayudante realizará una ayudantía explicando el uso e importancia de GitHub (22 de Agosto). Las personas que entreguen de esta manera tendrán una bonificación de 5 décimas en la nota final (si la nota final queda sobre 7.0, se perderán los puntos otorgados).