Master of Science in Data Science

Métodos de aprendizaje de máquinas en Data Science

Tarea 2 K-NN

La NASA mantiene la información de varios cometas y quiere determinar alguna manera de predecir el diámetro de un cometa. Específicamente, han analizado en forma manual una muestra de 100,000 asteroides. Los datos consisten en 26 variables, con distinta información como por ejemplo nombre del asteroide, su periodo orbital, su periodo de rotación, etc. Todos los datos existentes, se encuentran en un puro archivo llamado asteroidTrain.csv. Mientras que las descripciones de cada una de las variables se encuentran en el archivo tareaInformacion.txt

Desafortunadamente, la NASA tiene información bastante variada de cada cometa y todavía no ha evaluado 37.681 asteroides y no tienen tiempo para realizarlo. Por lo mismo, le piden que aplique k-NN para obtener una predicción de estos asteroides (después del proceso de limpieza de datos, el cual debe ser justificado).

Aplique k-NN regresor y evalúe los 37,681 asteroides que la NASA no ha evaluado. En este proceso genere un archivo csv de una sola columna con 37,681 filas donde cada celda tendrá el valor del cometa a predecir (2 puntos). Atención, si ustedes entrega un archivo con un número distinto de filas, de igual manera se evaluará las 37,681 filas.

El punto de evaluación final será una competencia entre todas las tareas basados en los MSE más bajos y altos obtenido por cada grupo. El puntaje final será una regresión lineal entre un modelo **muy básico** y el mejor puntaje.

Para esta entrega usted deberá entregar 2 archivos

1. Un archivo rmd que muestre todo el proceso de selección de variables y limpieza de datos aplicados. Además, deberá mostrar la búsqueda de hiperparametros. Este archivo ya deberá haber sido ejecutado y cuando se cargué uno deberá ver todo el proceso de ejecución.

2. Un archivo csv con las 37,681 estimaciones realizadas para el modelo.

Los puntajes asignados a cada tarea corresponden a:

- Limpieza de datos y selección (1 punto, código y justificación)
- Búsqueda de hiperparámetros y aplicación de kNN para los datos (3 puntos)
- Análisis de los resultados (2 puntos, código y justificación).

La fecha de entrega es el 10 de Octubre a las 15:30 horas.
La tarea se puede realizar hasta en grupos de 3 personas.
¡Mucha suerte!