Universidad del Rosario
Aprendizaje automático de máquina I - Grupo 1
Proyecto 1

Éste proyecto puede trabajarse de forma individual o en parejas. Deberá cargarse en formato .ipynb y .html a más tardar el sábado 4 de marzo a e-aulas.

El objetivo del proyecto es comparar el desempeño de diferentes modelos de aprendizaje automático en un problema de regresión. A saber, la predicción del precio de casas en una región de Estados Unidos, con base en 12 atributos.

Nota: Es importante que comenten mediante el uso de celdas de texto todo procedimiento llevado a cabo en cada celda de código. Por favor incluir también comentarios en las celdas de código de ser necesario. Los pasos a seguir son los siguientes:

1. Cargar el dataset HousingData, el cual se proveerá junto con un diccionario de los atributos. La variable objetivo es MEDV.

2. Examinar y realizar una breve descripción de los tipos de variable o cualquier observación que considere pertinente. Remueva filas con valores ausentes si en algún punto lo considera necesario.

3. Separe el conjunto de datos en atributos X y variable objetivo y. Luego separe éstos en conjuntos de entrenamiento y testeo, con una proporción de 75-25.

4. Entrene un modelo de regresión lineal usando el conjunto de entrenamiento. Determine su coeficiente de determinación (R^2 o, en el caso de SKLearn el *score*). Extraiga los coeficientes o parámetros predichos por el modelo usando el atributo *coef_*

5. Mencione cuáles atributos obtuvieron un coeficiente con mayor magnitud. Éstos son los que más afectan a la variable objetivo.

6. Cree una lista de valores para el parámetro alpha en los algoritmos de regularización. Usando la validación cruzada de Ridge, busque un valor óptimo para le parámetro α.

7. Realice una regresión de Ridge usando el parámetro α hallado en el punto anterior.

8. Repita los dos puntos anteriores para regresión de Lasso.

9. ¿Aplicar regresión de Ridge y Lasso mejoró el score de ésta regresion?

10. Aplique árboles de decisión de regresión para predecir la variable objetivo con distintos valores de los parámetros max_leaf_nodes y max_depth. Quédese con el modelo que dé un mayor score.

11. Grafique el árbol con el cuál obtuvo un mayor score. Describa el modelo que define este árbol. ¿Qué atributos son los más importantes para éste modelo? Compare esta lista de atributos con los obtenidos en el ítem quinto.¿Se utilizan todos los atributos en éste modelo? Si no, ¿cuáles no utiliza?

12. Ahora aplicaremos el método de vecinos más cercanos para regresión. Utilize el KNeighbors regressor de SKLearn para predecir nuestra variable objetivo usando diferentes valores del pará el parámetro k. Quédese con el mejor parámetro y mencione su score.

13. Finalmente, compare los resultados obtenidos con los distintos métodos. ¿Cuál tuvo un mejor desempeño? ¿Cuál tuvo un peor desemempeño? Mencione cualquier otra conclusión que extraiga de estos resultados o de los datos, por ejemplo: ¿por qué podría haberse comportado mejor un modelo que otro?