Aprendizaje de Máquinas
Tarea RLM & SVM

ITESM Campus Guadalajara

Escuela de Posgrados de Ingeniería y Ciencias
Maestría en Ciencias de la Computación **Dr. Luis Eduardo Falcón Morales**

Nombre(s): _____ **Matrícula(s):** _____

Esta tarea está basada en una base de datos de precios de casas de California que se encuentran en dos archivos: el un conjunto de datos de entrenamiento (17,000) y el conjunto de prueba (3,000):

https://www.kaggle.com/vikramtiwari/california-housing-dataset-ml-crash-course

Descarga ambos archivos: *california_housing_train.csv*, *california_housing_test.csv*

1) Utilizando los datos de entrenamiento sin hacerles algún preprocesamiento:
   a) Obtener el modelo de regresión lineal múltiple (*RLM*).

   b) Obtener los coeficientes del modelo y la significancia de cada uno de dichos coeficientes.

   c) Utilizando ahora el conjunto de datos de prueba:
      i. Obtener la media de la suma de los cuadrados de los errores (*MSSE*) y el coeficiente de correlación ajustado (R^2_{adj}).
      ii. Usando el vector de predicciones del modelo para los precios medios de las casas y la variable *median_house_value*:
         I. Obtener la gráfica de dispersión entre dichas variables. Teóricamente ¿qué gráfica se esperaría obtener?
         II. Obtener el coeficiente de correlación de Pearson entre ambas variables.
         III. Obtener el coseno del ángulo entre ambas variables. **NOTA:** Tanto el coeficiente de correlación de Pearson como el coseno del ángulo se utilizan usualmente como métricas para medir el desempeño en los modelos de regresión lineal.

   d) Reporta tus conclusiones. Este resultado se tomará como el modelo base para ser comparado con el resto de los ejercicios.

2) En los siguientes incisos deberás usar solamente el archivo con los datos de entrenamiento.. Es decir, en este ejercicio no debes usar todavía los datos de prueba del archivo *california_housing_test.csv*.

   a) Realiza un análisis descriptivo de los datos: mínimo, máximo, medianda, media, primer y tercer cuartil, si existen datos perdidos, etc.

   b) Obtener los diagramas de caja e histogramas de todas las variables. Reporta tus conclusiones a partir de dichos gráficos.

c) A partir de los gráficos anteriores, deberás proponer algún ajuste en relación a los datos extremos (outliers) de las variables que consideres adecuado. Justifica tu razonamiento aún cuando consideres que no se requiere ningún ajuste.

d) De acuerdo a los histogramas de las coordenadas geográficas dadas en las variables *longitude* y *latitude*, observamos que ambas distribuciones son claramente multimodales.
   i. ¿Qué significa que sean ultimodales?
   ii. ¿Por qué no es de extrañar que estas variables sean multimodales?
   iii. Los autores que generaron dicha base de datos comentan que agruparon los datos de acuerdo a diferentes zonas habitacionales de California mediante dichas coordenadas geográficas. (https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html)
   Completa la siguiente tabla con la cantidad de casas que se encuentran en las siguientes cuatro zonas (los límites se obtuvieron de manera aproximada a partir de los histogramas). Más delante se usará esta información, pero por el momento solamente completa la tabla:

	Longitude < -120	Longitude >= -120
Latitude >= 34.7		
Latitude < 34.7		

e) A partir de los histogramas obtenidos previamente, aplica en las variables que así lo consideres, alguna transformación para obtener histogramas con una distribución aproximadamente acampanada. Muestra los histogramas de las variables transformadas.

f) A partir del resumen descriptivo de los datos preprocesados en el inciso anterior y usando solamente las variables independientes, aplica una transformación para que cada variable quede centrada en el origen y todas en un mismo rango aproximadamente. ¿Por qué no incluimos en este caso a la variable dependiente de salida?
NOTA: Recuerda que los datos de validación y prueba deberánn transformarse en el origen y rango obtenidos con los datos de entrenamiento en este inciso.

g) Obtener los coeficientes de correlación de Pearson de cada par de variables. Con base a estos resultados simplifica las variables que consideres adecuado. Justifica tu razonamiento.

h) Usando los datos preprocesados hasta el inciso anterior, realiza una partición aleatoria en un conjunto de datos de entrenamiento del 80% y otro de validación del 20%. Reporta los coeficientes de regresión del modelo, la significancia de cada coeficiente, el promedio de la suma del cuadrado de los residuos MSSE y el coeficiente de correlación ajustado. Además, medir el desempeño de las predicciones del modelo con la métrica del coeficiente de Pearson y el coseno del ángulo. Escribe tus conclusiones comparándolo con el resultado del inciso (a).

i) Prueba otras particiones de los conjuntos de entrenamiento y validación y en caso de obtener un mejor resultado en relación al inciso anterior, reporta aquí su resultado. Utiliza los valores de R^2_{adj} y el coseno del ángulo para comparar los modelos. Para

fines de este ejercicio, continuaremos con la partición de 80%/20% en las siguientes preguntas.

j) Aplicar el modelo de máquinas de vectores de sporte (SVM) al problema. Deberás indicar el mejor modelo obtenido incluyendo al menos el valor del coeficiente asociado a las variables de holgura C, el modelo utilizado (lineal, o el de algún kernel no-lineal) y el valor del hiperparámetro gamma, en caso de haber usado un kernel no-lineal. Reporta las métricas en relación a las predicciones del modelo mediante el coeficiente de correlación de Pearson y el coseno del ángulo.

k) En la siguiente liga: (https://doi.org/10.1016/S0167-7152(96)00140-X) se encuntra el artículo original de los autores que generaron la base de datos California_Housing. Recurre a la base de datos del Tecnológico de Monterrey (https://biblioteca.tec.mx/inicio) para bajar dicho artículo (lo encuentras en la base de datos de las revistas de ScienceDirect). Aunque el conjunto original de datos utilizado por los autores en dicho artículo es ligeramente diferente a los de Kaggle que estamos utilizando, obtener el modelo no-lineal propuesto por los autores en la ecuación (8) de la página 295, usando el conjunto de entrenamiento y realiza la validación. Reporta las métricas en relación a las predicciones del modelo mediante el coeficiente de correlación de Pearson y el coseno del ángulo.

l) Incluye un nuevo factor (variable/columna) llamada “zona” para dividir el conjunto de casas en 3 regiones. Para fines de este ejercicio las llamaremos A, B y C, correspondientes a “latitude<34.7”, “(latitude>=34.7)&(longitude<-120)” y “(latitude>=34.7)&(longitude>=-120)”, respectivamente. Incluye dicho factor, verifica la correlación de Pearson con respecto a la latitude y longitude y obtén el modelo de RLM, reportando el desempeño del modelo en relación a las predicciones del modelo mediante el coeficiente de correlación de Pearson y el coseno del ángulo.

3) Con respecto al mejor modelo obtenido en los incisos anteriores, obtener el desempeño final de tu modelo mediante los datos de prueba del archivo *california_housing_test.csv*. Reporta el desempeño del modelo en relación a las predicciones del modelo mediante el coeficiente de correlación de Pearson y el coseno del ángulo.