Examen Final

Profesor: Alexis Montecinos

Ejercicios de Teoría (40 puntos)

1. (10 puntos) Un cliente lo contrata para la evaluación de cómo el salario del cliente afecta su propensión a utilizar la tarjeta de crédito del supermercado. El modelo es un modelo de clasificación que toma el valor 1 si el cliente ha usado la tarjeta en los últimos tres meses y cero si no. El analista recomienda que dada la precisión y facilidad del algoritmo, un buen candidato para esto es el método del K vecino más cercado. Comente en detalle su recomendación al respecto.

2. (5 puntos) El error cuádratico medio en problemas de clasificación no es muy preciso. Para esto, es mejor utilizar la raíz de este como función de costo, ya que este permite intuitivamente entender la unidad de medida en la estimación. Comente

3. (5 puntos) Explique cuál es la ventaja que posee el support vector classifier sobre el maximal margin classifier.

4. (10 puntos) En un restaurant los dueños poseen el problema que algunos clientes sólo vienen a comer una vez y no vuelven más. Plantee un modelo de clasificación donde pueda predecir si un cliente volverá o no al restaurant. Para esto defina el posible modelo, su variable dependiente y los posibles atributos que le permitirían llevar esto a cabo.

5. (10 puntos) Un análista le recomienda que dada su simplicidad utilice una regresión lineal en un problema de clasificación. Esto a su vez le permitirá medir los efectos de cada atributo en la clase predicha. Comente en detalle su recomendación al respecto.

Ejercicio de Python (60 puntos)

Utilizará una base de datos de precios de arriendo de viviendas. Esta será la variable dependiente. A su vez, la base cuenta con 5 variables explicativas o atributos. Estos son: número de dormitorios, número de baños, superficie construida, superficie de terreno y la distancia a la estación de metro más cercana. Realice lo siguiente:

1. (2 puntos) Importe la base de datos real estate a Python. Para esto es recomendado utilizar el paquete Pandas.

2. (3 puntos) Elimine las observaciones que no poseen datos completos para cada observación.

3. (5 puntos) Realice un trabajo de visualización de cada variable explicativa en su relación con la variable dependiente. Para esto, realice un scatter plot de cada atributo vs el precio de arriendo. Explique en cada caso que relación pareciera existir entre las dos variables.

4. (10 puntos) Estime una regresión lineal donde la variable dependiente es el logaritmo natural del precio de arriendo y las variables explicativas son el número de dormitorios, número de baños, el logaritmo natural de la superficie construida, el logaritmo natural de la superficie de terreno y el logaritmo natural de la distancia a la estación de metro más cercana.

5. (5 puntos) Calcule el ECM de la estimación previa. ¿Qué problema podría tener el calcular este estadístico con toda la muestra?

6. (10 puntos) Realice una estimación de la variable dependiente en logaritmo natural y cada una de las variables independientes utilizadas en 4. Es decir corra una regresión lineal con una constante y una sola variable independiente para cada caso. Realice un gráfico con scatter plot para los valores reales y una recta para el valor predicho en cada caso. Documente en qué caso pareciera que existe un underfitting.

7. (10 puntos) Estime un modelo donde incorpore una constante y la superficie construida como polinomio desde el grado 2 hasta el grado 10 (utilice el precio como variable dependiente). Reporte el ECM para este modelo. Para realizar esto, utilice una muestra de entrenamiento y testeo.

8. (15 puntos) Utilizando la estimación encontrada en el punto anterior, regularice utilizando el método de Ridge Regression. Para esto, cree una grilla de 10 valores para λ entre 0 y 5 y testee con cuál de ellos se encuentra el menor ECM. Es decir, diga con que λ de los 10 de la grilla el modelo posee el mejor ajuste. Para realizar esto, en cada caso, utilice una muestra de entrenamiento y testeo.