BUSINESS INTELLIGENCE PARA LAS FINANZAS

Profesor:David Díaz Solís
Ayudantes: Álvaro Gutiérrez Vargas

TAREA # 1
Primavera 2018

Parte 1 (50 puntos)

Utilizando data sintética cree un gráfico que muestre la complejidad computacional *"O()"* (tiempo de calculo en cada iteración) para la estimación de un modelo lineal como el de la Ecuación (1) cuando la cantidad de columnas (m) de la Matriz X va aumentando.

$$Y_i = c + \sum_{m=1}^{M} X_{i,m}\theta_m + \varepsilon_i \quad \forall i \in \{1, ..., N\} \tag{1}$$

Genere un código que realice un For o loop con las siguientes indicaciones:

1. El valor de la cantidad de filas (N) en la ecuación (1) debe estar fijo y ser igual a 100. Por otro lado, número de columnas (m) debe partir en 100 y llegar hasta 1.000.000 cuando acabe el loop.
2. El número de columnas (m) en la regresión se debe duplicar en cada iteración.
3. Su ciclo debe contener un quiebre que detenga el código si alguna de las iteraciones supera los 90 segundos de duración.
4. En cada iteración el *loop* deberá imprimir en pantalla el tiempo ocupado en el calculo.
5. Al final de las interaciones su gráfico debe representar en el eje de las abcisas el número de columnas de la regresión y en el eje de la ordenada la cantidad de tiempo de ejecución.

Preguntas

1. (20 puntos)¿Qué grado de complejidad aproximado puede observar en el gráfico resultante? ¿Cómo se compara este resultado con la complejidad computacional de aumentar el número de filas de la matriz X visto en clases?
2. (20 puntos) ¿Existe alguna diferencia al estimar la Ecuación (1) si se ocupa la ecuación normal vs ocupar la librería sklearn? Ayuda: Compare los resultados de tiempos de ejecución obtenidos al ocupar la librería *"sklearn"*, vs ocupar *"numpy"* para resolver la Ecuación (2). (**Indicación**: Ambos cálculos deben estar contenidos en el mismo loop)

$$\hat{\theta} = (X'X)^{-1}(X'Y) \tag{2}$$

3. (10 puntos) Solo ocupando la librería *"sklearn"* repita el ejercicio, pero ahora considerando que las variables explicativas de su modelo se encuentran al cuadrado. es decir, considere que la ecuación corresponde a un polinomio de grado dos descrito en la ecuación (3). Comente sobre qué es lo que ocurre con la complejidad computacional al duplicar el número de columnas en esta especificación y genere un nuevo gráfico que superponga esta curva, con la obtenida en la parte anterior.

$$Y_i = c + \sum_{m=1}^{M} X_{i,m}^2\theta_m + \varepsilon_i \quad \forall i \in \{1, ..., N\} \tag{3}$$

Parte 2 (50 puntos)

1. (5 puntos) Haciendo uso de la base de datos disponible en el siguiente link, consolide toda la información disponible en las hojas del archivo Excel en un solo DataFrame. En su informe deberá detallar si los datos fueron concatenados usando "inner", "outter" o de otra forma y por qué.

2. (10 puntos) Genere las muestras Training, Testing1 y Testing2, siguiendo el formato entregado en el Libro de Excel. Ocupando *"sklearn"* entrene una Regresión Lineal en su muestra de Training y luego calcule el MAE en Testing 1 y Testing 2. Interprete sus resultados. Presente una Tabla con el número de observaciones de cada muestra, los coeficientes estimados, el número de variables explicativas y el MAE de las tres muestras. ¿Cual debería ser mayor *apriori*?

3. (35 puntos) Ocupando un *"loop"* genere una fecha de corte movil entre 24 de Enero de 2013 y el 13 de Diciembre de 2016 que divida las muestras de Training y Testing. En cada iteración deberá recuperar el MAE de ambas muestras (Training y Testing). Finalmente, deberá graficar dos curvas, una para el MAE obtenido en Training y otra con el MAE obtenido en Testing, colocando en el eje X la fecha de corte utilizada para obtener dichas métricas. Interprete su resultado. Interprete su resultado. ¿Cuál de las curvas debería ir aumentando y cual bajando a medida que crece la muestra de testing? De una explicación intuitiva de su resultado.

Instrucciones para entrega

- **Instrucciones de envío**: La evaluación deberá ser enviada en la fecha **14 de Septiembre de 2018 a las 17:30hrs** al correo `algutierre@fen.uchile.cl` un archivo comprimido (7zip o zip) (todos los archivos dentro de las carpetas deben tener por nombre los RUTS de los integrantes (sin número verificador). El asunto del mail debe ser **BI_FIN_"RUT**[1]**"**[ejemplo : BI_FIN_12345678_19854451]

    1. Archivo **".py"**[ejemplo: 12345678_19854451.py]: Este debe contener el código de su tarea. Este archivo **debe** ser intensivo en comentarios. Códigos sin explicaciones tendrán descuento en la nota final.
    2. Archivo **".pdf"**[ejemplo: 12345678_19854451.pdf] con el informe de su tarea. Máximo 5 planas.

- La tarea puede ser desarrollada en grupos de hasta tres personas.

[1] AMBOS RUTS (O PASAPORTE SEGÚN CORRESPONDA) SIN DÍGITO VERIFICADOR