PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

TAREA 2
Métodos Estadísticos para el Manejo de Grandes Volúmenes de Datos
EYP3707

Profesora	:	Ana María Araneda
Ayudante	:	Bianca del Solar
Fecha de entrega	:	21 de mayo de 2022, 23:59 horas.

Como solución a esta tarea, usted debe subir a Canvas:

- Un archivo .pdf con figuras y respuestas a las preguntas.
- Archivos de código R o Phyton. No se evaluarán tareas cuyos códigos no corran o no sean coherentes con las respuestas entregadas.

Para cerrar el tópico de algoritmos supervisados de predicción, en esta tarea utilizaremos la base de datos `FINAL_USO.csv`, de la Tarea 1, para implementar un método de reducción de dimensionalidad y uno de árboles de regresión.

Recuerde que la base de datos contiene información diaria relacionada al valor de transacción del oro en el mercado, entre el 15 de diciembre de 2011 y el 31 de diciembre de 2018. La base de datos incluye variables como el precio del crudo, índice S&P500, índice Dow Jones, bonos a 10 años del gobierno estadounidense, tasa de cambio de dólar estadounidense a euros, y valores de metales preciosos, como plata y platino, y de otros metales, como paladio y rodesio, entre otros. Todas ellas intentan explicar el comportamiento el precio de cierre diario del oro. La base también incluye información sobre el movimiento del valor del oro durante un mismo día (como valores mínimo y máximo, entre otros) sin embargo se le pide no utilizarla para explicar el valor de cierre.

Al igual que en la Tarea 1, la variable que interesa predecir corresponde al valor de cierre de las transacciones del oro (ajustadas) en dicho día, expresados como porcentaje del mínimo valor observado en dicho período (aproximadamente 1.050 USD la onza, el 17 de diciembre de 2015), `Adj.Close`.

Utilice la misma separación de los ejemplos en las bases de entrenamiento, de validación y de testeo utilizada en la Tarea 1.

PARTE I: Reducción de dimensionalidad por descomposición en valores singulares

Considere la matriz de variables de entrada, X, de $n \times p$, y su descomposición en valores singulares dada por:

$$X = UDV^t.$$

Como estudiamos en clases, esta descomposición puede ser utilizada para reducir dimensionalidad y evitar el problema de sobreajuste. Para esto, dado un valor de $r \leq p$, se propone entrenar el modelo lineal utilizando una matriz de variables de entrada reducida dada por:

$$X_r = U_r D_r,$$

donde U_r corresponde a la matriz compuesta por las primeras r columnas de la matriz U y D_r corresponde a la matriz diagonal de $r \times r$ que contiene los primeros r valores singulares de X, cuando ellos han sido ordenados de manera decreciente.

Una vez entrenado el modelo lineal en base a X_r, la predicción de registros futuros, con variables de entrada en X_{val} (o en X_{test}) se realiza transformando previamente estas variables en la forma:

$$X_{val}(r) = X_{val}\ V_r,$$

donde V_r corresponde a la matriz que contiene las primeras r columnas de la matriz V. Las predicciones se obtienen como de costumbre, como:

$$\hat{Y} = X_{val}(r)\ \theta,$$

donde θ corresponde al valor del parámetro entrenado con la matriz X_r.

1. Calibre el valor de r utilizando la base de datos de validación y utilizando el error cuadrático medio como criterio de optimalidad. Para ello, obtenga un gráfico que muestre este error versus la dimensionalidad utilizada r. Justifique su elección.

2. Obtenga el error cuadrático medio en la base de datos de testeo, utilizando el valor de r que determinó en el apartado anterior.

Nota: no olvide normalizar las variables de entrada.

PARTE II: Árboles de regresión y Random Forests

1. Entrene un árbol de regresión sin podar, utilizando todas las variables de entrada (no es necesario normalizar en este caso).

2. Utilice el método de validación cruzada en la base de entrenamiento para calibrar el parámetro de penalización por complejidad, α.

3. Obtenga una representación gráfica del árbol podado y, de acuerdo a ella, prediga el valor diario del oro (ajustado) para el primer registro en su base de datos de testeo. Explique.

4. Obtenga el error cuadrático medio del árbol podado en la base de datos de testeo.

5. Entrene ahora un algoritmo de bosques aleatorios utilizando 100 árboles, y obtenga su error cuadrático medio en la base de datos de testeo.

6. Compare los errores cuadráticos medios obtenidos a través del modelo lineal en la Parte I, a través de un único árbol, y a través del bosque aleatorio, en la Parte II.