PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE MATEMÁTICAS
DEPARTAMENTO DE ESTADÍSTICA

TAREA 1
Métodos Estadísticos para el Manejo de Grandes Volúmenes de Datos
EYP3707

Profesora	:	Ana María Araneda
Ayudante	:	Bianca del Solar
Fecha de entrega	:	29 de abril de 2022

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.

PARTE I: Curva de aprendizaje (*Learning curve)*

Al entrenar un algoritmo de aprendizaje de máquina, es importante monitorear su comportamiento, para determinar si está sufriendo problemas de sesgo o varianza, y tomar así las medidas correspondientes. En este problema se explorará esta idea utilizando una base de datos pequeña, de modo de ilustrar los conceptos principales.

Una *curva de aprendizaje* muestra el comportamiento del error cuadrático medio en la medida en que se agregan más ejemplos a la base de entrenamiento. La idea principal es comparar, en una figura, los errores cuadráticos medios de las bases de entrenamiento y de validación, en la medida en que aumenta el número de ejemplos utilizados para entrenar el algoritmo.

Si al aumentar el número de ejemplos, ambos errores se acercan, sin embargo el valor al que convergen es demasiado grande considerando el problema en cuestión, se está en presencia de sesgo: la hipótesis $h_{\boldsymbol{\theta}}$ que se está utilizando es demasiado simple para explicar el comportamiento de los datos. En este caso, deben considerarse modelos polinomiales o con un mayor número de variables de entrada.

Por otra parte, si los errores en la base de validación y en la base de entrenamiento se estabilizan en la medida en que se agregan más ejemplos en la base de entrenamiento, pero el valor al que converge el error en la base de validación es mucho mayor al de la base de entrenamiento, se está en presencia de problemas de varianza, y una de las soluciones es utilizar regularización.

Para ilustrar estas ideas, utilizaremos un modelo lineal para predecir el volúmen de agua que fluye en una represa, en términos del cambio del nivel del agua en la reserva. La base de datos original ha sido separada aleatoriamente en una base de datos de entrenamiento, una de validación y una de testeo. Todas se encuentran en el archivo `represa.mat`.

1. Ajuste un modelo lineal (sin penalización) utilizando la base de entrenamiento y grafique la recta ajustada sobre el gráfico de dispersión de los datos. ¿Observa problemas se sesgo o varianza? Explique.

2. Para verificar lo anterior, utilizaremos la *curva de aprendizaje* del algoritmo. Para esto, para cada valor de $i = 1, \ldots, m$, con m el número de ejemplos en la base de entrenamiento, usted debe:

   a) Entrenar el modelo lineal en la base de entrenamiento, utilizando únicamente las primeras i observaciones. Note que, para $i = 1$, debe ajustar el modelo constante (o nulo), dado que no es posible ajustar dos parámetros con un único ejemplo.

   b) Obtener el error cuadrático medio en la base de entrenamiento, utilizando únicamente las observaciones que se usaron en el paso anterior para entrenar, es decir, las observaciones $1, \ldots, i$.

   c) Obtener el error cuadrático medio en la base de validación, utilizando todas las observaciones.

   d) Guardar dichos errores.

   Finalmente, debe construir una figura que muestre ambos errores versus el número de ejemplos utilizados en el entrenamiento. De acuerdo a lo explicado previamente sobre las *curvas de aprendizaje*, ¿observa problemas de sesgo o varianza? Explique.

3. Ajustaremos ahora un modelo polinomial de orden 8, es decir, el modelo tendrá 9 coeficientes (uno para cada potencia de la variable de entrada, más el intercepto). Para ello obtenga los vectores con las 7 potencias faltantes y normalícelas, guardado sus medias y sus desviaciones estándar.

4. Ajuste un modelo lineal polinomial de orden 8, sin penalización, utilizando la base de datos de entrenamiento. Grafique la curva ajustada sobre el gráfico de dispersión de los datos. ¿Observa problemas se sesgo o varianza? Explique.

5. De manera similar al procedimiento utilizando en el modelo con una única variable de entrada, obtenga los errores necesarios para construir la *curva de aprendizaje* del modelo polinomial. Construya la figura que muestre ambos errores versus el número de ejemplos utilizados en el entrenamiento. De acuerdo a lo explicado previamente sobre las *curvas de aprendizaje*, ¿observa problemas de sesgo o varianza? Explique.

6. Ajuste ahora el modelo polinomial penalizado por:

$$\frac{\lambda}{2m} \sum_{j=1}^{8} \theta_j^2$$

   para los valores $\lambda = 10$ y $\lambda = 100$. Obtenga las curvas ajustadas y compare sus comportamientos dependiendo del valor de λ utilizado.

7. Utilizando la base de datos de testeo, obtenga el error cuadrático medio del modelo lineal con una única variable de entrada (sin penalizar), el modelo lineal polinomial (sin penalizar) y el modelo lineal polinomial regularizado con el valor de λ que escoja en el paso anterior, $\lambda = 10$ o $\lambda = 100$ (justifique su elección) . Comente sus resultados.

PARTE II:

En esta parte de la tarea se pondrá énfasis en el procedimiento de validación de un algoritmo.

Los datos (tomados desde `https://www.kaggle.com/datasets/sid321axn/gold-price-prediction-dataset`) se encuentran en el archivo `FINAL_USO.csv`, que 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.

La variable que interesa predecir se encuentra bajo el nombre `Adj.Close`. Ella 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).

1. Separe los ejemplos de la base, de manera aleatoria, en una base de entrenamiento (60 %), de validación (20 %) y de testeo (20 %). Utilice una semilla dada (por usted).

2. Entrene un modelo Ridge para diferentes valores del parámetro de regularización. Muestre, a través de un gráfico, la evolución de los coeficientes estimados en la medida en que varía dicho parámetro. Interprete.

3. Repita el procedimiento en el apartado anterior para el modelo Lasso.

4. Para el modelo Ridge, encuentre el valor óptimo del parámetro de regularización, utilizando la base de datos de validación y el método de validación cruzada 10-fold sobre la base de datos de entrenamiento. Compare sus resultados y comente.

5. Repita el procedimiento en el apartado anterior para el modelo Lasso.

6. Entrene los modelos Ridge y Lasso utilizando los valores óptimos del parámetro de regularización que encontró en los apartados anteriores. Evalúe su comportamiento en la base de datos de testeo y comente.