FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Tarea 2 - EYP3407

Profesor: Mauricio Castro
Ayudante: Leonardo Medina

1. **(30%)** Asumiendo que $\mathbf{X}$ es ortogonal, es decir, $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, obtenga el sesgo y varianza de los estimadores de mínimos cuadrados, ridge y LASSO. Compárelos y comente sus hallazgos.

2. **(70%)** Descargue el conjunto de datos *hitters*. Esta base corresponde a los datos de una liga de baseball entre las temporadas de 1986 y 1987. Para una descripción de los datos puede ver el enlace `https://cran.r-project.org/web/packages/ISLR/ISLR.pdf`. La variable de respuesta para este problema es **Salario**. Como la distribución del Salario es sesgada, se debe tomar la transformación $Y = \log(\text{Salario})$.

   (a) **(20%)** ¿Cuáles son las características más importantes para predecir el salario de los jugadores?

      (i) Ajuste y visualize métodos de regularización vistos en clase (LASSO, Elastic-Net) incluyendo LASSO adaptativo.

      (ii) ¿Cuáles son los mejores predictores seleccionados por cada método? ¿Son diferentes? Si lo son, ¿por qué?

   (b) **(50%)** ¿Cuál método es mejor para predecir el salario de los jugadores? Para hacer la decisión considere una set de entrenamiento (60%), validación (20%) y test (20%). Si los métodos considerados tienen parámetros que calibrar, entonces se debe ajustar el modelo con el set de datos de entrenamiento, se debe elegir los parámetros a calibrar minimizando el error de predicción en el set de validación y se debe reportar la predicción final con el set de datos de testeo. Se debe repetir este procedimiento 10 veces y reportar los resultados promedio (*Nota*: puede ver más detalles sobre el set de validación en la página 176 del libro "An Introduction to Statistical Learning" de James, Witten, Hastie y Tibshirani).

      (iii) Compare el MSE promedio obtenido en los set de datos de testeo considerando (a) mínimos cuadrados, (b) regresión Ridge, (c) LASSO, (d) Elastic-Net y (e) LASSO adaptativo.

      (iv) Visualize los resultados obtenidos para comparar los modelos. Muestre los resultados solamente para el mejor parámetro de calibración elegido.

(v) ¿Qué métodos generan el mejor error de predicción? ¿Por qué estos métodos funcionan bien? ¿Los métodos eligen el mismo subconjunto de variables? Explique y amplía sus respuestas.

Aspectos a evaluar en la presentación oral (si corresponde): Cada grupo deberá presentar sus principales hallazgos en una presentación de 10 minutos. Solo se evaluará la presentación (calidad, duración, claridad de la exposición, contenido) y si el código funciona o no en vivo (programación).