Control de Lectura: Aprendizaje Supervisado

Pregunta 1 (25 Puntos)

Sea Y una variable de respuesta cuantitativa y contamos con p predictores $X_1, X_2 ... X_p$ se asume que hay una relación entre Y y $X = (X_1, X_2 ... X_p)$ que se puede representar de una forma general como:

$$Y = f(X) + \epsilon \tag{1}$$

1. Describí cada componente de la ecuación anterior.
2. ¿Porqué nos interesa estimar f?
3. ¿La precisión de $\hat{Y}$ como predicción de Y depende de dos cantidades, ¿Cuáles son estas cantidades y en base a ellas como podemos mejorar la precisión de nuestro modelo?
4. ¿Qué es el compromiso entre sesgo y varianza?
5. ¿Porqué es recomendable separar los datos al menos en conjunto de entrenamiento y testeo?

Pregunta 2 Verdadero y Falso (18 Puntos)

Inidica en cada caso con V si es Verdadera y F si es Falsa.

1. Un método **no paramétrico** asume una forma para f (de la Ecuación 1) intentando que su estimación sea lo más cercana posible a los datos observados.
2. Aprendizaje estadístico supervisado implica que para alguna observación de las predictoras (x_i) con $i = 1, 2, ... n$ hay una respuesta asociada y_i.
3. En un problema de clasificación de dos clases (clase 1 y clase 2), el clasificador de Bayes corresponde en predecir la clase 1 si $P(Y = 1/X = x_0) > 0.5$ y la clase 2 en otro caso.
4. En el método de vecino más cercano donde K representa el número de vecinos cuanto menor es el número de vecinos más flexible es el modelo.

5. El proceso de evaluar el desempeño predictivo del modelo se llama selección de modelos.
6. Bootstrap se usa principalemente como un método de estimación de parámetros

Pregunta 3 (20 Puntos)

En un modelo de regresión lineal simple:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

1. ¿Cómo se interpreta $\hat{\beta}_0$, $\hat{\beta}_1$?
2. ¿Cómo se interpreta el siguiente intervalo de confianza?:

$$[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)]$$

3. ¿Cuál es el estadístico utilizado para la prueba de significación de β_1? Explicitá su forma. (No es necesario explicitar la forma del varianzas o errores estándar en caso de ser necesario)
4. En un modelo de regresión lineal múltiple con tres predictotes (X_1, X_2, X_3):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

¿Qué implica el supuesto de aditividad y que se podría hacer en caso que el mismo no se cumpla?

Pregunta 4 (15 Puntos)

En un problema de regresión describa el procedimiento que se debe realizar para calcular alguna medida de performance usando 5-fold-CV. En la descripción seleccioná una medida concreta de performance.

Pregunta 5 (12 Puntos)

En el capítulo 5 de remuestreo se presenta el siguiente gráfico respecto a bootstrap.

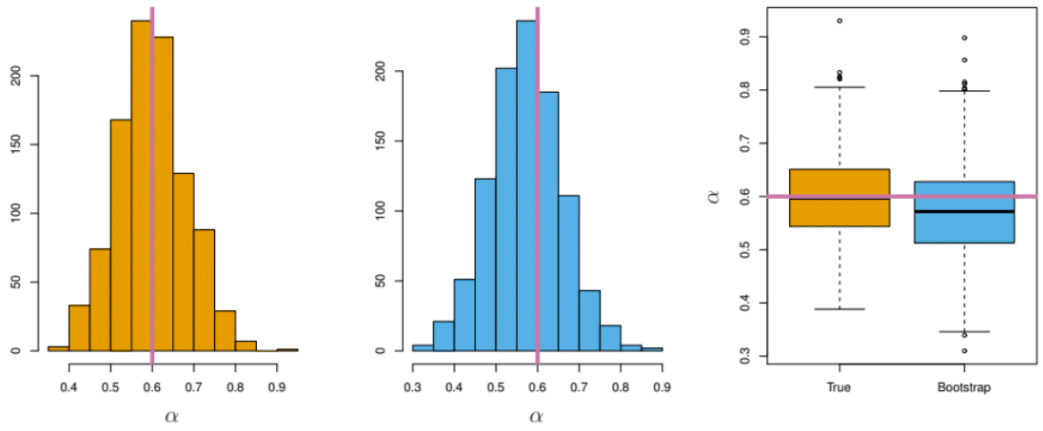

FIGURE 5.10. Left: *A histogram of the estimates of* α *obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of* α *obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of* α *displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of* α.

¿Que tratan de ejemplificar con dicho ejemplo respecto al método de remuestreo basado en bootstrap?

Pregunta 6 (10 Puntos)

1. ¿Dentro de que métodos se encuentran Ridge y Lasso?
2. ¿Cuál es la principal diferencia entre ambos métodos?