Control de Lectura Solución: Aprendizaje Supervisado

Pregunta 1: Capítulo 2

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.

Y es la variable de respuesta cuantitativa, $f(X)$ es el componente sistemático que representa la información que tiene X de Y y ϵ el componente aleatorio que tiene media cero y es independiente de X.

2. ¿Porqué nos interesa estimar f?

Estimo f con objetivo predictivo o para hacer inferencia.

3. ¿La precisión de $\hat{Y}$ como predicción de Y depende de dos cantidades, ¿Cuales son estas cantidades y en base a ellas como podemos mejorar la precisión de nuestro modelo?

Error reducible y el irreducible, el irreducible es el error asociado al componente aleatorio y ne no voy a mejorar aunque tenga el mejor modelo y el error irreducible es el que voy a poder mejorar si mejoro el modelo.

4. ¿Qué es el compromiso entre sesgo y varianza?

Para ejemplificarlo pueden mencionar el problema de regresión y que el MSE esperado de test se puede descomponer para un valor de x_0 en la varianza de $\hat{f}(x_0)$ el sesgo al cuadrado de $\hat{f}(x_0)$ y la varianza del error entonces para disminuir el error de test hay que seleccionar un modelo que tenga bajo sesgo y baja varianza

al mismo tiempo. En general cuando el modelo es más flexible tiene mayor variza y menor sesgo y aquí el trade-off.

5. ¿Porqué es recomendable separar los datos al menos en conjunto de entrenamiento y testeo?

Porque el error de generalización es mejor estimado con el conjunto de test, no podemos usar los mismos datos para entrenar el modelo que para evaluarlo ya que el error de entrenamiento es optimista.

Pregunta 2 Verdadero y Falso

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

Pregunta 3

En un modelo de regresión lineal simple:

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

1. ¿Cómo se interpreta $\hat{\beta}_0$, $\hat{\beta}_1$?

$\hat{\beta}_1$ es el efecto promedio en Y cuando se aumenta en una unidad en X

Si $X = 0$ se espera que Y sea $\hat{\beta}_0$ en promedio

2. ¿Cómo se interpreta el siguiente intervalo de confianza?:

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

Si tomamos repetidas muestras y calculamos el IC para cada muestra, el 95% de los intervalos va a contener el verdadero valor del parámetro β_1

Hay un 95% de confianza que el intervalo contenga al verdadero valor de β_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)

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

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?

El efecto de X_j en Y es independiente del resto de las predictoras. Si no se cumple se deberían incluir interacciones entre las variables predictoras.

Pregunta 4

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.

Supongo que tenemos n observaciones, dividimos los datos en 5 grupos con 5<n: estimo f con n-n/5 y se predice con n/5, calculo el MSE y repito el procedimiento 5 veces. Luego tengo 5 valores de MSE y calculo el promedio de los 5 MSE y obtengo el MSE por CV.

Pregunta 5

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

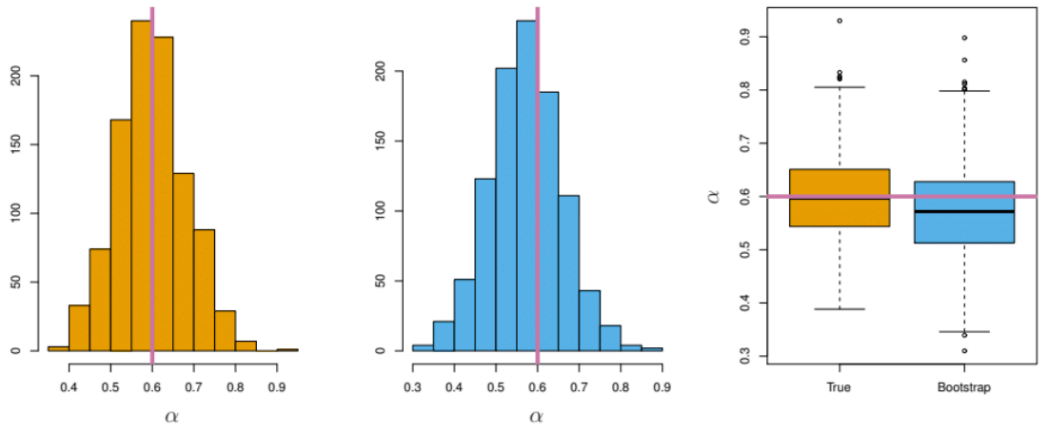

FIGURE 5.10. Left: *A histogram of the estimates of* α *obtained by generating 1,000 simulated data sets from the true population.* Center: *A histogram of the estimates of* α *obtained from 1,000 bootstrap samples from a single data set.* Right: *The estimates of* α *displayed in the left and center 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?

Se comparan los resultados en base a simulaciones de la población para estimar α y su variabilidad. A su vez se realiza lo mismo en base a réplicas bootstrap. En el gráfico se indica el valor verdadero de α (línea rosada) y se muestra la distribución de su estimación usando el método de simulación y el método basado en bootstrap. Podemos ver que el método basado en bootstrap es bueno para aproximar bien la variabilidad de $\hat{\alpha}$ pero no es bueno para estimar α. Por eso es que se usa bootstrap como herramienta para evaluar propiedades estadísticas

Pregunta 6

1. ¿Dentro de que métodos se encuentran Ridge y Lasso?

Métodos de regularización o regresión regularizada.

2. ¿Culál es la principal diferencia entre ambos métodos?

La penalidad utilizada, en le caso de Ridge usa una penalidad cuadrática y Lasso usa penalidad en valor absoluto.

- **Ridge penalidad**: $\lambda \sum_{j=1}^{p} \beta_j^2$
- **Lasso penalidad**: $\lambda \sum_{j=1}^{p} |\beta_j|$

La penalidad en Ridge puede llevar a que los coeficientes estimados sean cercanos a cero pero no iguales a cero como en el caso de Lasso. Por lo que Lasso sirve como un método de selección de variables.