Tarea 1

Entrega 23 de Octubre

Ejercicio 1

En clase mencionamos que estimar el error de un modelo utilizando los mismos datos con los que el modelo fue estimado subestima el error, es intuitivamente razonable y vimos ejemplo con simulaciones.

El objetivo de este ejercicio es demostrar teoricamente el optimismo del error de entrenamiento para un modelo de regresión lineal.

Sea (x_i, y_i) con $i = 1, \dots, n$ una muestra de *entrenamiento* y $(x_i^{'}, y_i^{'})$ con $i = 1, \dots, m$, una muestra de *validación* que asumimos i.i.d. Además, las variables explicativas son p dimensionales tal que $\beta \in R^p$. Los coeficientes estimados de la regresión lineal en el conjunto de entrenamiento son

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

donde x es una matriz nxp siendo la fila i-ésima x_i y y es un vector n-dimensional con i-ésimo componente y_i.

Se pide demostrar:

$$E[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2] \leq E[\frac{1}{m}\sum_{i=1}^{m}(y_i^{'} - \hat{\beta}^T x_i^{'})^2] \tag{1}$$

Implica probar que el error esperado de entrenamiento es siempre menor o igual al error esperado de validación, implicando que el error de entrenamiento es optimista.

Para demostrar la (1), proponemos recorrer los siguientes pasos:

1. Argumentar que el error esperado de validación es el mismo si tenemos m observaciones o solamente 1. Esto es

$$E[\frac{1}{m}\sum_{i=1}^{m}(y_i^{'} - \hat{\beta}^T x_i^{'})^2] = E[(y_i^{'} - \hat{\beta}^T x_1^{'})^2]$$

para esto es importante tener en cuenta que $E()$ incluye la aleatoriedad de las muestras de entrenamiento y validación, y se puede utilizar la regla de esperanzas iteradas: $E(Z) = E_t\left[E_v(Z|t)\right]$ donde E_t considera aleatoriedad en la muestra de entrenamiento y E_v en la de validación.

Si lo anterior es cierto, entonces podemos asumir sin pérdida de generalidad que $m = n$ (entrenamiento y validación tienen el mismo número de observaciones).

2. Ahora considera dos variables aleatorias:

$$A = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2 \quad B = \frac{1}{n}\sum_{i=1}^{n}(y_i^{'} - \tilde{\beta}^T x_i^{'})^2$$

donde $\tilde{\beta}$ denota el coeficiente de regresión lineal estimado en el conjunto de validación. Argumenta que A y B tienen la misma distribución entonces $E(A) = E(B)$

3. Argumentá que la variable aleatorio B definida en la parte anterior es siempre menor o igual al error de validación

$$B = \frac{1}{n}\sum_{i=1}^{n}(y_i^{'} - \tilde{\beta}^T x_i^{'})^2 \leq \frac{1}{n}\sum_{i=1}^{n}(y_i^{'} - \hat{\beta}^T x_i^{'})^2$$

4. Utiliza el resultado de la parte 3 para concluir

$$E[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2] \leq E[\frac{1}{n}\sum_{i=1}^{n}(y_i^{'} - \hat{\beta}^T x_i^{'})^2]$$

Ejercicio 2

Para cada parte de la a) a la d) indica si debemos esperar en general que la performance de un método de aprendizaje estadístico flexible sea mejor o peor que un método inflexible. Justifica tu respuesta

a) El tamaño muestral n es extremadamente grande y el número de pedictores p es pequeño.

b) El númeto de predictores p es extremadamente grande y el número de observaciones n es pequeño.

c) La relación entre los predictores y la respuesta es marcadamente no lineal.

d) La varianza del término de error $\sigma^2 = V(\epsilon)$ es extremadamente alta.

Ejercicio 3

¿Cuáles son las ventajas y desventajas de una aproximación muy flexible vs una menos flexible para un problema de regresión o clasificación? Bajo que circunstancias puede una aproximacioón más flexible ser preferida a una menos flexible. ¿Cuando una aproximación menos flexible es preferible?

Ejercicio 4

Cuidadosamente explica la diferencia entre un el método de vecino más cercano para clasificación y regresión.

Ejercicio 5

Suponga que contamos con un conjunto de datos con 100 observaciones ($n = 100$) que contienen un único predictor y una respuesta cuantitativa. Se ajusta un modelo de regresión lineal para los datos así como un regresión cúbica $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

a) Suponé que la verdadera relación entre X e Y es lineal, $Y = \beta + 0 + \beta_1 X + \epsilon$. Considerar la suma de cuadrado de los residuos de entrenamiento (SCR) para la regresión lineal y la SCR para la regresión cúbica. Esperarías que una sea menor que la otra, que sean iguales o no hay suficiente información para concluir. Justifica tu respuesta

b) Responde el punto anterior utilizando el conjunto test en vez del entrenamiento para calcular la SCR

c) Suponé que la verdadera relación entre X e Y no es lineal, pero no sabemos que tan lejana está de la linealidad. Considerar la SCR de entrenamiento para la regresión lineal y también la SCR de entrenamiento para la regresión cúbica. ¿Esperaríamos que una sea mayor a la otra, que sean iguales o no hay suficiente información para concluir? Justifica tu respuesta

d) Responder la anterior usando el conjunto test en vez el de entrenamiento

Ejercicio 6

En un problema de regresión donde Y es la variable de respuesta y X la variable predictora, se cuenta con 20 observaciones que se presentan en la siguiente tabla.

El objetivo de este ejercicio es obtener el MSE por validación cruzada usando 5 pliegues. En este problema la relación entre Y y X es ajustada con un modelo de regresión lineal simple tal que $f(X) = \beta_0 + \beta_1 X$ En lo que sigue se presentan las ID de las observaciones seleccionadas aleatoriamente en cada pliegue.

	Y	X
1	8	6
2	9	8
3	14	12
4	10	9
5	10	9
6	15	13
7	11	11
8	6	6
9	7	5
10	8	9
11	13	13
12	11	10
13	11	11
14	10	10
15	8	8
16	15	15
17	11	11
18	4	3
19	12	11
20	8	7

- Pliegue 1: 4, 3, 19, 16
- Pliegue 2: 2, 15, 7, 18
- Pliegue 3: 9, 14, 12, 20
- Pliegue 4: 17, 6, 8, 10
- Pliegue 5: 1, 5, 13, 11

Utiliza la información antes detallada para explicitar en cada pliegue los cálculos que debes realizar para obtener el MSE 5-CV (estimación del MSE por validación cruzada usando 5 pliegues). En este ejercicio es importante el proceso para llegar al resultado final no simplemente obtener el MSE 5-CV.

Ejercicios ISLR

- Capítuo 5 Ejercicio 8
- Capítulo 7 Ejercicios: 1, 2, 3, 4 y ejercicio 9 y 10 usando `tidymodels`