Introducción a Machine Learning para Ciencias Sociales: Proyecto 1

Pavel Coronado - Anzony Quispe

10 de mayo de 2023

El grupo compuesto por un máximo de 5 integrantes deberá generar un rmarkdown file como reporte. Este archivo tendrá todas las respuestas a la tarea y debe ser entregado únicamente por un integrante del grupo. La fecha máxima de entrega es el 17 de mayo a las 11:59 pm . Adicionalmente, el archivo debe contener los nombres y códigos de todos los integrantes. El grupo escogido para el proyecto 1 se mantendrá para el proyecto 2. Las bases de datos mencionadas en las preguntas se encuentran disponibles en la libreria **(ISLR)**. Finalmente, el nombre del archivo debe contener el código de todos los integrantes separados por un guion bajo.

Ejemplo: **proyecto1_20150317_..._..._**.

Cualquier duda respecto al proyecto escribir a **anzony.quispe@gmail.com**.

Bootstrap

1. Para una muestra de tamaño 5 (n=5), ¿Cuál es la probabilidad que la j-enésima observacion se encuentre dentro de la muestra bootstrap?
2. Genere un plot que muestre en el eje X el tamaño de muestra (n) desde 1 a 100000 y en el eje Y la probabilidad que la j-enésima observación se encuentre dentro de la muestra bootstrap. Comente los resultados.
3. Ahora consideraremos el conjunto de datos de vivienda de Boston, del ISLR2 library. Use **boot**. **(View(Boston))**

   a) Con base en este conjunto de datos, proporcione una estimación de la media de la variable **medv**. Nombre a esta estimación **u**.

   b) Proporcione una estimación del error estándar de **u**. Hoot: $SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}$

   c) Ahora estime el error estándar de **u** usando bootstrap. Compare con **(b)**.

   d) Con base en su estimación en **(c)** , proporcione un intervalo de confianza al 95 % de la media de **medv** utilizando bootstrap. Compárelo con los resultados obtenidos usando **t.test(Boston$medv)**.

   e) Con base en este conjunto de datos, proporcione una estimación, **u_med**, para la mediana de la variable **medv**.

   f) Estime el error estándar de **u_med** utilizando bootstrap. Comente sus hallazgos.

Cross Validation

1. Explique cómo se implementa k-fold cross validation.
2. Comente las ventajas y desventajas de k-fold cross validation con respecto a **validation set approach** y **LOOCV**.

Lasso y Ridge

1. Determine cuál de las siguientes proposiciones es verdadera. Justifique su respuesta.

   a) Lasso con respecto a OLS es:

1) Más flexible y, por lo tanto, mejorará la precisión de la predicción cuando el incremento en el sesgo es menor que la reducción en la varianza de las predicciones.
2) Más flexible y, por lo tanto, mejorará la precisión de la predicción cuando el incremento en la varianza es menor que la reducción del sesgo de las predicciones.
3) Menos flexible y, por lo tanto, mejorará la precisión de la predicción cuando el aumento en el sesgo es menor que la disminución en la varianza de las predicciones.
4) Menos flexible y, por lo tanto, mejorará la precisión de la predicción cuando su aumento en la varianza es menor que su disminución en parcialidad de las predicciones.

b) Evalue **(a)** para Ridge con respecto a OLS.

2. Supongamos que estimamos los betas en el siguiente modelo de regresión lineal:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{s.t.} \sum_{j=1}^{p}|\beta_j|\leq s$$

Determine cual de las siguientes proposiciones es verdadera. Justifique su respuesta.

a) A medida que aumentamos s desde 0, el RSS de la data de entrenamiento:
1) Aumenta inicialmente y luego eventualmente comienza a disminuir en forma de U invertida.
2) Disminuye inicialmente y luego eventualmente comienza a aumentar en forma de U.
3) Aumenta constantemente.
4) Disminuye constantemente.
5) Permanece constante.

b) A medida que aumentamos s desde 0, el RSS de la data de test:
1) Aumenta inicialmente y luego eventualmente comienza a disminuir en forma de U invertida.
2) Disminuye inicialmente y luego eventualmente comienza a aumentar en forma de U.
3) Aumenta constantemente.
4) Disminuye constantemente.
5) Permanece constante.

3. Supongamos que estimamos los betas en el siguiente modelo de regresión lineal:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

Determine cual de las siguientes proposiciones es verdadera. Justifique su respuesta para la proposición verdadera.

a) A medida que aumentamos λ desde 0, el RSS de la data de entrenamiento:
1) Aumenta inicialmente y luego eventualmente comienza a disminuir en forma de U invertida.
2) Disminuye inicialmente y luego eventualmente comienza a aumentar en forma de U.
3) Aumenta constantemente.
4) Disminuye constantemente.
5) Permanece constante.

b) A medida que aumentamos λ desde 0, el RSS de la data de test:
1) Aumenta inicialmente y luego eventualmente comienza a disminuir en forma de U invertida.
2) Disminuye inicialmente y luego eventualmente comienza a aumentar en forma de U.
3) Aumenta constantemente.
4) Disminuye constantemente.
5) Permanece constante.

4. Predecir el número de solicitudes recibidas usando las variables en el conjunto de datos de College. **(View(College))**

    a) Divida el conjunto de datos en un conjunto de entrenamiento (70 %) y un conjunto de prueba (30 %).

    b) Ajuste un modelo lineal usando OLS en el conjunto de entrenamiento, y reportar el MSE del conjunto de prueba.

    c) Ajuste un modelo Ridge en el conjunto de entrenamiento, con λ elegido por cross validation. Reporte el MSE del conjunto de prueba.

    d) Ajuste un modelo Lasso en el conjunto de entrenamiento, con λ elegido por cross validation. Reporte el MSE del conjunto prueba, junto con el número de estimaciones de coeficiente distintas de cero.

    e) Muestre un **dataframe** que resuma los resultados.

5. Ahora intentaremos predecir la tasa de criminalidad per cápita en los datos de Boston, del ISLR2 library. **(View(Boston))**

    a) Proponga un modelo (o conjunto de modelos) que performen bien en este conjunto de datos y justifique su respuesta. Asegúrese de que estos modelos evaluen el rendimiento del modelo utilizando cross validation, en lugar de utilizar el error de entrenamiento.

    b) ¿El modelo elegido involucra todas las variables disponibles? Justifique.