

Lista de Ejercicios 2

Machine Learning para CCSS

20 de octubre, 2024

Instrucciones

1. **Temas abordados:** Esta lista de ejercicios se enfoca en los siguientes temas: Bootstrap & Regularization Methods.
2. **Formación de grupos:** Se permite la formación de grupos de hasta 5 integrantes. La composición de los grupos se mantendrá constante para la lista de ejercicios 3.
3. **Puntuación de ejercicios:** La lista contiene 5 ejercicios. Cada ejercicio vale 4 puntos.
4. **Formato de entrega:** La resolución de los ejercicios debe presentarse en un archivo jupyter-notebook con todas las celdas ejecutadas.
5. **Fecha límite de entrega:** La fecha límite para la entrega es el Domingo 28 de octubre a las 11:59 pm. Un representante del equipo debe subir su solucionario a la actividad correspondiente en la plataforma Canvas. Los nombres y códigos de todos los participantes deben ser incluidos en el solucionario.

Pregunta 1

Considere el conjunto de datos de viviendas de Boston:

- a) Defina una función llamada `estimate_mean` que tome como argumentos un DataFrame de Pandas `data`, una serie de índices `idx` y el nombre de una columna `x`. Esta función debe calcular y devolver la media de la columna `x` para las filas correspondientes a los índices dados en el DataFrame. Utilice esta función y estime la media poblacional de la columna `medv` en el conjunto de datos proporcionado. Llame a esta estimación $\hat{\mu}$. *Hint:* Para estimar la media poblacional, asegúrese de incluir todos los índices del conjunto de datos en el cálculo.
- b) Proporcione una estimación del error estándar de $\hat{\mu}$. Interprete este resultado. *Hint:* Puede calcular el error estándar dividiendo la desviación estándar de la columna `medv` (`.std()`) por la raíz cuadrada del número de observaciones.
- c) Ahora estime el error estándar (SE) de $\hat{\mu}$ usando el método Bootstrap. ¿Cómo se compara esto con su respuesta de (b)? *Hint:* Puede modificar la función `boot_SE` para adaptarla a la función definida en (a).
- d) Basado en su estimación bootstrap de (c), proporcione un intervalo de confianza del 95% para la media de `medv`. *Hint:* Puede aproximar un intervalo de confianza del 95% usando la fórmula $\hat{\mu} - 2SE(\hat{\mu})$, $\hat{\mu} + 2SE(\hat{\mu})$.
- e) Defina una función para estimar la mediana de una variable. Use esta función y estime la mediana poblacional de la columna `medv`. Llame a esta estimación $\hat{\mu}_{med}$. *Hint:* puede seguir el mismo enfoque que en (a).
- f) Estime el error estándar (SE) de $\hat{\mu}_{med}$ mediante el método Bootstrap. *Hint:* Puede seguir el mismo enfoque en que (c).

Pregunta 2

Suponga que estima los coeficientes de regresión de un modelo de Regresión Lineal minimizando:

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

para un valor particular de λ . Indique si los ítems son verdaderos o falsos y justifique su respuesta.

a) A medida que λ se incrementa desde 0, la métrica RSS (Residual Sum of Squares) de entrenamiento cambiará de la siguiente forma:

- I. Aumentará inicialmente, y luego comenzará a disminuir en forma de U invertida.
- II. Disminuirá inicialmente, y luego comenzará a aumentar en forma de U.
- III. Aumentará constantemente.
- IV. Disminuirá constantemente.
- V. Permanecerá constante.

b) Repita a) para la métrica RSS de prueba.

c) Repita a) para la varianza.

d) Repita a) para el sesgo (bias) al cuadrado.

Pregunta 3

Suponga que estima los coeficientes de regresión de un modelo de Regresión Lineal minimizando:

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

para un valor particular de λ . Indique si los ítems son verdaderos o falsos y justifique su respuesta.

a) A medida que λ se incrementa desde 0, la métrica RSS (Residual Sum of Squares) de entrenamiento cambiará de la siguiente forma:

- I. Aumentará inicialmente, y luego comenzará a disminuir en forma de U invertida.
- II. Disminuirá inicialmente, y luego comenzará a aumentar en forma de U.
- III. Aumentará constantemente.
- IV. Disminuirá constantemente.
- V. Permanecerá constante.

b) Repita a) para la métrica RSS de prueba.

c) Repita a) para la varianza.

d) Repita a) para el sesgo (bias) al cuadrado.

Pregunta 4

Prediga el número de solicitudes recibidas (columna **Apps**) utilizando la dataset College:

- a) Divida el conjunto de datos en un conjunto de entrenamiento y un conjunto de validación.
- b) Ajuste un modelo lineal utilizando mínimos cuadrados en el conjunto de entrenamiento e informe el error en el conjunto de prueba.
- c) Ajuste un modelo de regresión Ridge en el conjunto de entrenamiento, con λ elegido por Cross-Validation. Reporte el error en el conjunto de prueba y los coeficientes de las variables.
- d) Ajuste un modelo de regresión Lasso en el conjunto de entrenamiento, con λ elegido por Cross-Validation. Informe el error en el conjunto de prueba y los coeficientes de las variables.
- e) Muestre un pandas Dataframe que resuma los resultados. ¿Cuál modelo seleccionará para su investigación? Justifique.

Pregunta 5

En este ejercicio predecirá las ventas de autos de juguete (variable **Sales**) de la dataset Carseats:

- a) Implemente los modelos de regularización explorados: Ridge y Lasso. Presente y discuta los resultados.
- b) Proponga un modelo (o conjunto de modelos) que parezca funcionar bien en este conjunto de datos y justifique su respuesta. Asegúrese de que está evaluando el desempeño del modelo utilizando el error del conjunto de validación, la validación cruzada o alguna otra alternativa razonable, en lugar del error de entrenamiento.
- c) ¿Su modelo elegido involucra todas las variables de la base de datos? ¿Por qué o por qué no?