Tarea RLM con datos de Salaries

2023-04-24

Tarea

Usaremos el conjunto de datos de *Salaries* [*car*], que contiene el salario académico de nueve meses de 2008-09 para profesores asistentes, profesores asociados y profesores en una universidad en los EE. UU. Los datos se recopilaron como parte del esfuerzo continuo de la administración de la universidad para monitorear las diferencias salariales entre los profesores masculinos y femeninos.

La tarea consiste en encontrar el mejor modelo de regresión para el dataset "Salaries" de la biblioteca "car" de "R", seleccionando las mejores variables X. El objetivo es producir la predicción más precisa para la categoría "Prof" de la disciplina "B" de sexo "Male" con 30 "yrs.service" y 33 "yrs.since.phd".

Para llevar a cabo esta tarea, se pueden seguir los siguientes pasos:

1. Cargar el dataset "Salaries" en R usando la función `data()` .
2. Explorar los datos usando funciones como `head()`, `summary()`, `str()`, `cor()` y `pairs()` para tener una idea general de las variables incluidas en el *dataset*, su distribución, su correlación y su posible relación con la variable objetivo, que en este caso es el salario.
3. Seleccionar un conjunto inicial de variables X que puedan ser relevantes para predecir el salario, basándose en la exploración previa de los datos y en el conocimiento experto del dominio. Se puede utilizar la función `lm()` de R para ajustar un modelo de regresión lineal con estas variables.
4. Evaluar la calidad del modelo utilizando medidas de error como el error cuadrático medio (MSE) o el coeficiente de determinación (R-squared). También se puede utilizar la validación cruzada para estimar el error de generalización del modelo.
5. Si el modelo inicial no produce una predicción precisa, se pueden probar diferentes combinaciones de variables X y ajustar nuevos modelos. Para seleccionar las mejores variables X se pueden utilizar métodos de selección de características como el *backward elimination* o el *forward selection*.
6. Evaluar la calidad de cada modelo ajustado y seleccionar el que produce la predicción más precisa en la categoría "Prof" de la disciplina "B" de sexo "Male" con 30 "yrs.service" y 33 "yrs.since.phd".
7. Finalmente, se puede utilizar el modelo seleccionado para hacer la predicción de salario para un individuo con características específicas, como un hombre con 30 años de servicio y 33 años desde su doctorado en la disciplina "B" y la categoría "Prof". Para hacer la predicción, se debe aplicar el modelo seleccionado utilizando los valores específicos de cada variable. Se puede evaluar la precisión de la predicción comparándola con el salario real de la persona en cuestión y produciendo un intervalo de confianza para la predicción.

El entregable es un reporte realizado en `Rmarkdown` con la salida corresondiente (PDF, HTML o DOCX) con lo que se ha indicado.