Pontificia Universidad Católica de Chile
Facultad de Ciencias Económicas y Administrativas

Análisis de Big Data:
EAA361A

Laboratorio 7

Profesor: Cristián Vásquez Ayudante: Pablo González

Ejercicio 1: Selección de Variables y Regularización

Para este ejercicio utilizaremos el conocido set de datos "Auto MPG". Como usted sabe, esta base de datos contiene información sobre el rendimiento en millas por galón de diferentes modelos de automóviles de la década de los 70 y 80, teniendo en cuenta sus cilindros, desplazamiento, potencia, peso, año, aceleración y origen. Utilizando sus conocimientos de R y lo aprendido en clases, responda las siguientes preguntas:

1. Determine el mejor modelo según SSE que contenga 7 variables.
2. ¿Cuál de los modelos tiene el mayor $\bar{R}^2$?
3. Genere una tabla que le permita comparar ambos modelos y comente a qué se pueden deber estas diferencias. *(Hint:* La librería Stargazer permite realizar dicha tabla)
4. Realice una regresión Lasso y compare con el modelo anteriormente generado.
5. Propuesto: Separe el set de datos en dos muestras (entrenamiento y testeo) en proporción 3:7. Posteriormente realice una regresión Lasso como se le mostró anteriormente y, utilizando la muestra de testeo, calcule el R^2 del modelo.

Ejercicio 2: Modelo Logit

Utilizando la versión base de datos "binary", que contiene información acerca de 400 estados de admisión teóricos a la universidad, responda la siguiente pregunta:

1. Asumiendo que contamos con un modelo que contiene las variables GRE y RANK para predecir el ingreso a la universidad, evalúe la necesidad de agregar la variable GPA (el promedio del postulante) al modelo.
2. Realice una regresión logarítmica de ADMIT sobre GRE, RANK y GPA e interprete los coeficientes obtenidos.
3. Calcule la razón de probabilidades para cada una de las variables. Predictoras.
4. Genere una muestra aleatoria con 500 datos, de forma que la variable RANK se distribuya uniformemente y las variables GPA y GRE distribuyan normalmente con media y desviación estándar iguales a los datos originales. Luego, utilizando dichos datos, estime la probabilidad de ingreso a dicha universidad por parte de los alumnos.
5. ¿Qué porcentaje de los postulantes serían admitidos si solo aquellos con probabilidad de ingreso mayor o igual a 0,5 fueran admitidos?
6. Propuesto: Utilizando una muestra de entrenamiento determine el punto de corte óptimo para construir un clasificador binario. Aplique el clasificador en la muestra de validación y reporte la matriz de confusión. Comente los resultados

* Datos extraídos de https://stats.idre.ucla.edu/stat/data/binary.csv

Ejercicio 3: Caso Pokémon

Luego de una larga travesía, Ash Ketchum (サトシ para los amigos) vuelve a casa con su Pokédex completa. Ésta es luego entregada al profesor Oak, quien le pide a usted, que, utilizando lo aprendido en el curso, analice la base de datos y genere un árbol de decisión que le permita determinar si un Pokémon es legendario o no en función de las siguientes variables:

Ataque	Defensa	Velocidad
Puntos de vida	Ataque especial	Defensa especial
Felicidad Base	Altura	Peso

0. Elimine de la base de datos aquellas variables que no necesitará para su modelo (No se preocupe, el profesor Oak cuenta con un respaldo)
1. Separe los datos en una muestra de entrenamiento y otra de validación con probabilidad 0.7 y 0.3 respectivamente. (*Hint*: Recuerde definir una semilla)
2. Siguiendo las instrucciones del profesor Oak, y utilizando la muestra de entrenamiento, genere un árbol de decisión y grafíquelo.
3. Determine un parámetro de costo de complejidad óptimo y pode el árbol en función de este.
4. Utilizando la muestra de validación, calcule la cantidad de verdaderos negativos, verdaderos positivos, falsos negativos y falsos positivos. Utilizando dichos valores, entregue una matriz de confusión para su modelo.
5. Finalmente, calcule las siguientes métricas: (1) Tasa de reconocimiento, (2) Tasa de error, (3) Sensibilidad, (4) Especificidad y (5) Precisión. Comente.

** Datos extraídos de www.kaggle.com/rounakbanik/pokemon/data con fecha 23/03/2020*