Modelado y Predicción Estadística
Regresión con regularización

Beatriz Coronado Sanz

1. Apartado 1

Para obtener el modelo de regresión múltiple de la variable *varobj* frente al resto de variables usamos la función *lm(varobj˜.,datos)*. Obtenemos los coeficientes del polinomio para aproximar *varobj* a partir del resto de las variables de nuestros datos. Obtenemos también la desviación típica de cada variable y el resultado del contraste de hipótesis para cada una de las variables en función del nivel de significación α ($\alpha = \{0, 0.001, 0.01, 0.05, 0.1, 1\}$).

Analizamos el problema de multicolinealidad de dos formas distintas: usando el determinante de la matriz de covarianzas de los datos y con el cálculo del VIF para cada una de las variables. Observamos que el determinante es casi 0 ($2.034153e-134$), por lo que hay multicolinealidad. Con el cálculo del VIF podemos averiguar que variables la producen. Decimos que una variable tiene multicolinealidad si su VIF es mayor que 10. Para nuestros datos, las variables *x04*, *x07* y *x10* tienen VIF > 10 (26.20, 21.71 y 10.2, respectivamente).

2. Apartado 2

Para aplicar las técnicas de regularización elasticnet, LASSO y RIDGE usamos la función *glment(datos,prediccion,α,λ)* donde $\alpha = 0.5$ para elasticnet, $\alpha = 1$ para LASSO y $\alpha = 0$ para RIDGE. λ es el valor de regularización. En nuestro problema daremos a λ una secuencia de valores para obtener una sucesión de modelos aplicando la técnica de regularización escogida. En el

apartado 3 dividiremos el conjunto de datos en entrenamiento y test y seleccionaremos el mejor modelo obtenido en cada técnica para compararlo con el resto de modelos.

3. Apartado 3

Para realizar un análisis comparativo de los 4 modelos vamos a calcular el MSE de cada modelo. Para ello vamos a dividir el conjunto de datos entre entrenamiento y test (50 % de los datos para cada conjunto) y vamos a calcular un conjunto de modelos para cada técnica de regularización con el conjunto de datos de entrenamiento mediante validación cruzada (menos para regresión múltiple que solo calcularemos un modelo) . Después vamos a predecir los datos para el conjunto de test con cada uno de los modelos y elegiremos para cada técnica de regularización el modelo con menor MSE para el conjunto test.

Los resultados obtenidos serán parecidos a los de la siguiente imagen:

Metodo	MSE	Lambda
"Ridge"	"4.06178681352168"	"0.002"
"Lasso"	"4.05959979891097"	"0.001"
"Elastic"	"4.06042636499274"	"0.001"
"RegMult"	"272.227807057485"	""

Figura 1: Tabla con los resultados obtenidos para cada método

Observamos que el MSE para regresión múltiple es muy alto, pero para el resto de métodos casi no se aprecian diferencias entre ellos. Llegamos a la conclusión de que para nuestros datos es mejor un modelo de regresión con regularización que un modelo de regresión múltiple porque a la hora de añadir nuevos valores los errores no son despreciables.