Análisis Inteligente de Datos - Tarea 2

Iván González - Diego Salazar

Regresión Lineal Ordinaria

Dataset: Prostate Cancer

Conjunto de Datos

- Dataset compuesto de 97 muestras: 67 de entrenamiento y 30 de pruebas.
- Atributos predictores (8):
  - Lcavol: Logaritmo volumen de cáncer. Variable de punto flotante.
  - Lweight: Logaritmo del peso de la próstata. Variable de punto flotante.
  - Age: Edad del paciente. Variable entera.
  - Lbph: Logaritmo de la cantidad de hiperplasia prostática benigna. Variable de punto flotante.
  - Svi: Invasión vesículo seminal. Variable entera.
  - Lcp: Logaritmo de la penetración capsular. Variable de punto flotante.
  - Gleason: Calificación de Gleason. Variable entera.
  - Pgg45: Porcentaje de Gleason 4 ó 5. Variable entera.
- Respuesta: lpca (nivel de antígeno prostático cancerígeno). Variable de punto flotante.

Preprocesamiento de datos

- Estandarización
  - Predictores con varianza igual a 1. Útil cuando se quiere que los coeficientes de la regresión sean comparables entre sí, especialmente cuando los predictores son cantidades físicas distintas o tienen una escala distinta.
- Agregar columna de 1s al dataset
  - Desplazar el intercepto fuera del origen (distinto de cero)

Pesos y Z score

Atributo	Peso	Z score
intercept	2.465	27.359
lcavol	0.676	5.320
lweight	0.262	2.727
svi	0.304	2.448
lbph	0.209	2.038
pgg45	0.266	1.723
gleason	-0.021	-0.145
age	-0.141	-1.384
lcp	-0.287	-1.851

- Usando la distribución de probabilidad t-student, con una significancia del 5% y con 67 - 9 = 58 grados de libertad, se tiene un intervalo de confianza igual a [-1.672, 1.672].
- Aquellas variables cuyo z score se encuentre dentro del intervalo , no existirá suficiente evidencia que demuestre su relación con la respuesta.

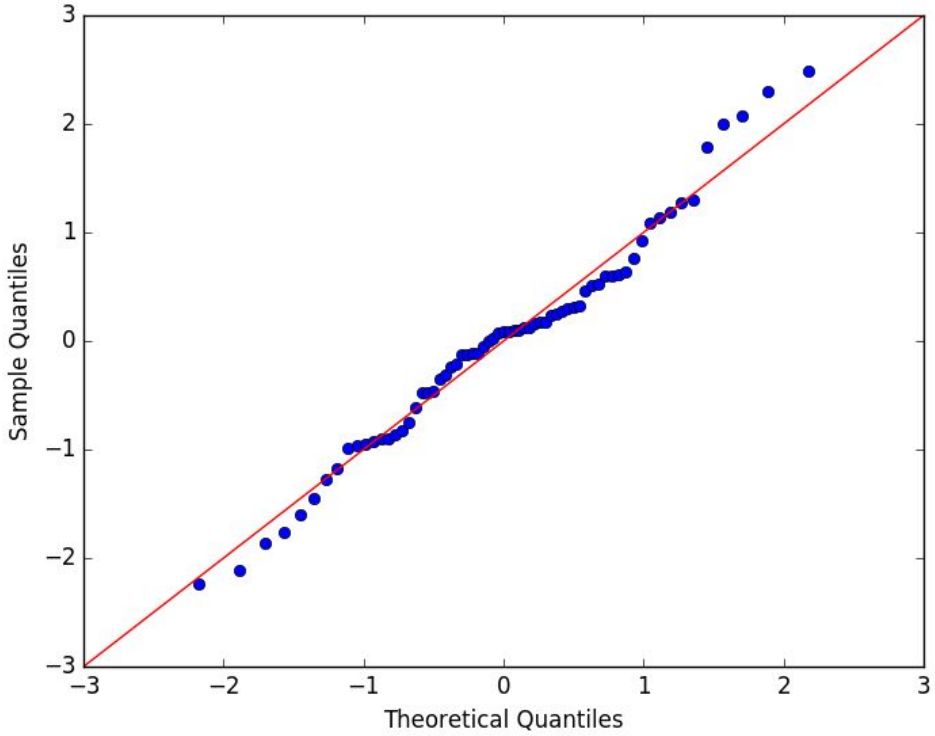
Error de predicción y Cross Validation

- Con k folds, con k = 5, 6, 7, 8, 9, 10:

K	5	6	7	8	9	10
MSE	0.957	0.957	0.895	0.880	0.819	0.757

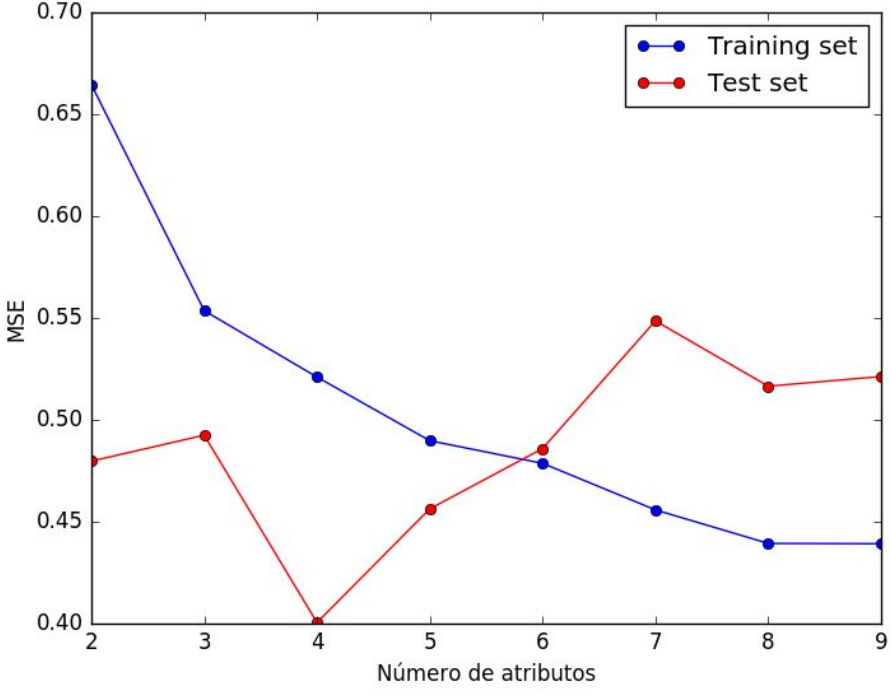
- MSE set de pruebas: 0.521

Hipótesis de normalidad de los errores



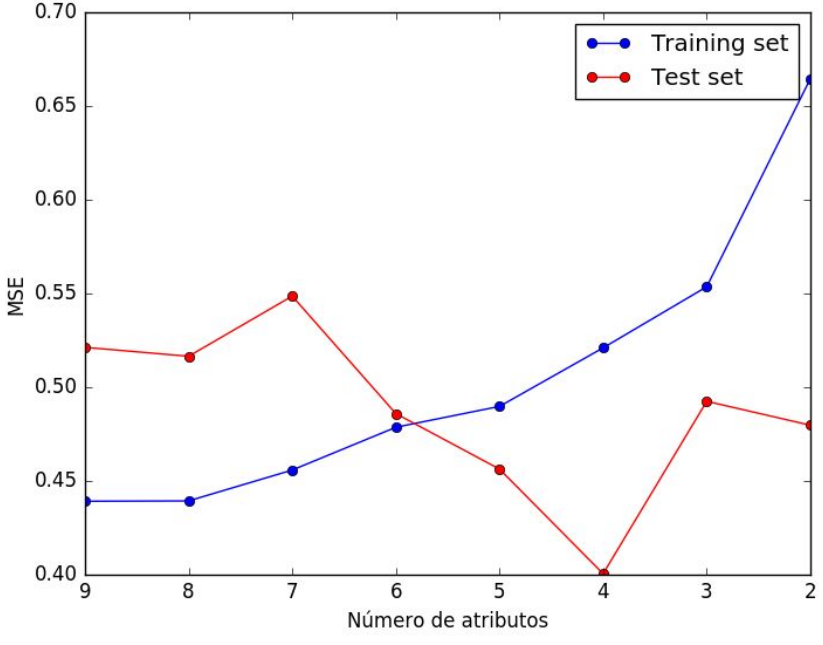
Selección de atributos

Forward stepwise selection



- Set de testing, mínimo MSE con cuatro atributos:
  - **lcavol**, **lweight**, **svi** más el **intercepto**

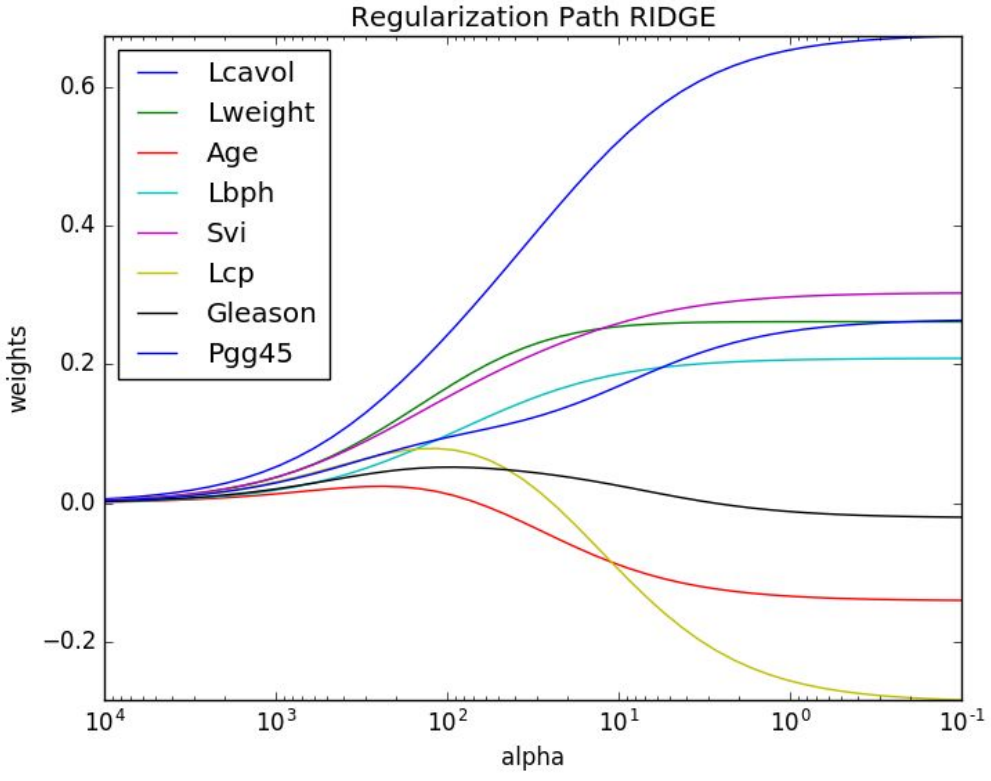
Backward stepwise selection



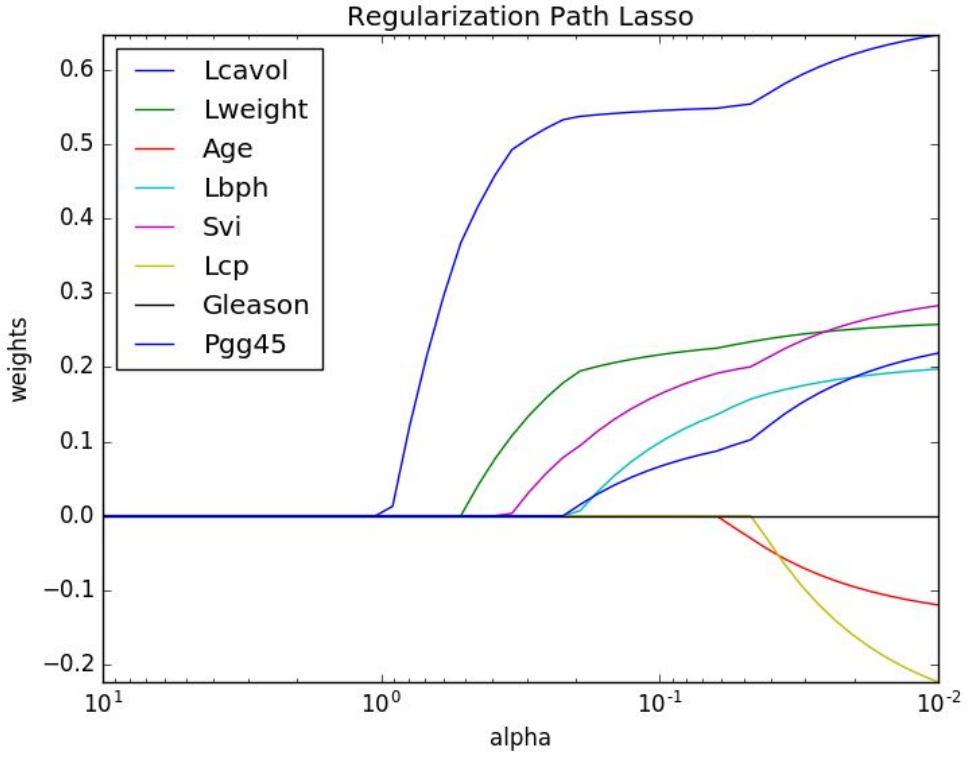
- Set de testing, mínimo MSE con cuatro atributos:
  - **lcavol**, **lweight**, **svi** más el **intercepto**

Regularización

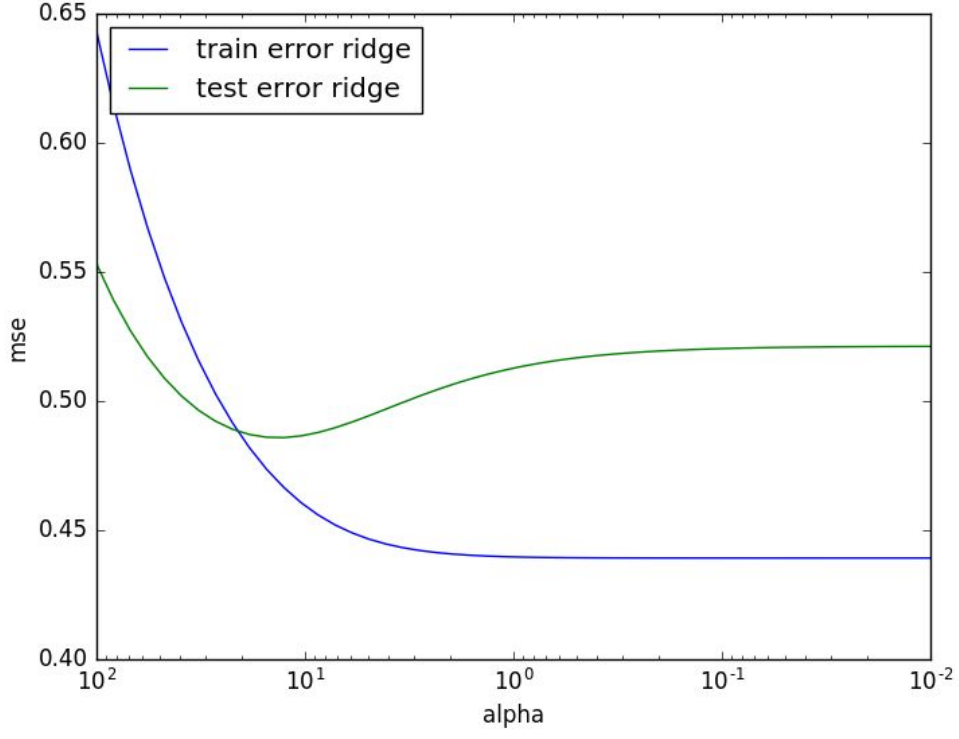
Regularization path Ridge Regression



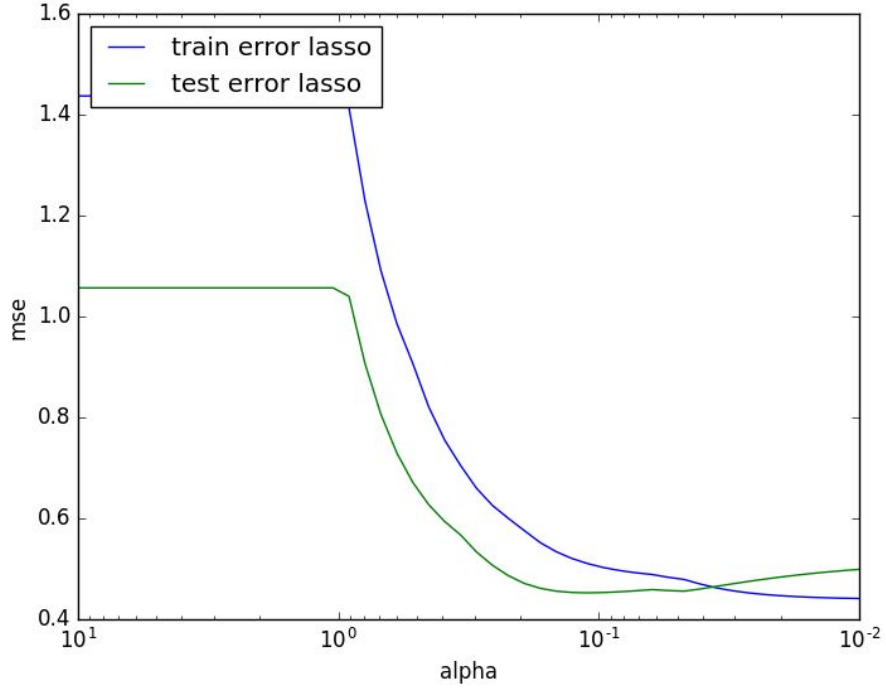
Regularization path Lasso Regression



Ridge Regression: MSE vs Alpha parameter



Lasso Regression: MSE vs Alpha parameter



Mejores parámetros de regularización

Método	Rango ƛ	Mejor válor	mse
Ridge	$[10^{-2}, 10^{2}]$	2.330	0.752
Lasso	$[10^{-2}, 10^{1}]$	0.010	0.759

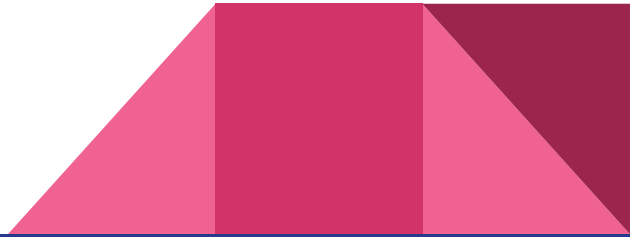
Predicción de Utilidades de Películas

Características de los datos

- Matriz *sparce* de 276222 columnas
  - Una matriz demasiado grande como para manipular con librerías corrientes.
  - Se opera con *csr_matrix* para aprovechar las propiedades de dicha matriz.
  - Para preservar dispersión, ***no se centra y se opera con intercepto no nulo***.
- Cada columna contiene metada o texto asociado a crítica de películas

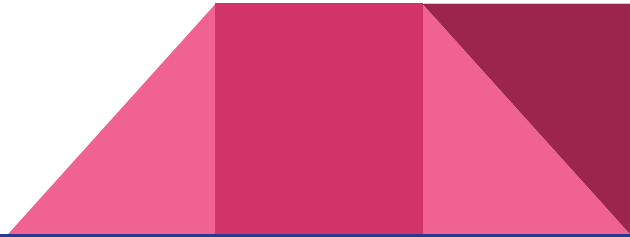
Características de los datos

- Datos volumen de utilidades por lugar de proyección
  - Al poseer un mayor volumen de datos, se comienzan las pruebas aquí.
  - Se hacen pruebas regularizando mediante LASSO y Ridge
  - Se evita el uso de LASSO por por problemas de convergencia y lentitud computacional.



Método general para operar

- Se aplican principios de *machine learning*
  - Se generan varios modelos con Ridge o LASSO a partir de los datos de entrenamiento.
  - Se calcula el MSE asociado a partir de los datos de validación y se elige el modelo con menor MSE.
  - Se calcula el coeficiente R^2 final a partir de los datos de test.



Características de los datos

- Datos volumen de utilidades por lugar de proyección

Método	Alfas (rango)	Alfa (elegido)	MSE (mín)	R^2
Ridge	[0.1 , 0.2 , … ,1.0]	0.900	96214654.36	0.242
Ridge	LogSpaced [-1, 0.9] N=10	3.005	96215832.02	0.242
Ridge	1000	1000	6109517.21	0.242
Lasso	[0.7 , 0.8 , 0.9]	0.7	139681061.77	0.079
Lasso	1000	1000	140348785.28	0.114
Lasso	1,000,000	1,000,000	154630228.53	0.033

Características de los datos

- Datos volumen de utilidades por lugar de proyección
  - No hay resultados satisfactorios con ninguno de los dos métodos.
  - Demasiado lento hacer cada prueba: ¡hasta 2 horas iterando sobre 10 alfas!
  - Se prueba con los datos para el total de utilidades por película.
  - Se aprovecha de que el set es más pequeño para converger más rápido a las respuestas.

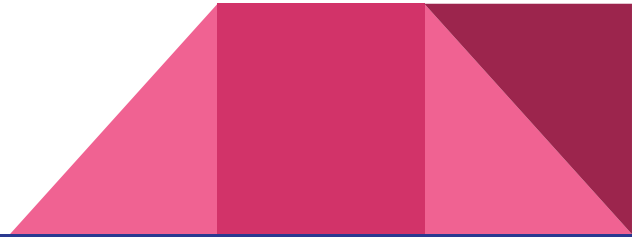
Características de los datos

- Datos volumen total de utilidades

Método	**Alfas (rango)**	**Alfa (elegido)**	**MSE (mín)**	**R^2**
Ridge	[0.1 , 0.2 , … ,1.0]	0.2	150449.52E9	0.515

Características de los datos

- Datos volumen total de utilidades
  - Mayor precisión inmediata
  - Problema: No se puede lograr un resultado mucho mejor:



Características de los datos

- Datos volumen total de utilidades:
  - Haciendo múltiples tests sobre alfas entre 0 y 3000 (sin intercepto para reducir tiempo)
  - Entre 0 y 600 no existe mayor diferencia en coef. de determinación.
  - Sobre 600, R^2 comienza a decrecer
  - No se puede lograr $R^2 \geq 0.75$ con Ridge.
  - Lasso tarda demasiado para hacer Tests suficientes.
  - Falta poder computacional para Resolver problema