Análisis Inteligente de Datos: Tarea 2

Felipe Valdivia
Anibal Pérez

Introducción

Se utilizarán los métodos de regresión lineal estudiados en clases para analizar:

- prostate-cancer, datos de pacientes luego de aplicada prostatectomía.
- Datos sobre Utilidades en estreno de Películas.

DESCRIPCIÓN DATASET PROSTATE-CANCER

- 97 registros.
- 9 Columnas.
- 8 serán predictores.
- 1 de respuesta (lpsa).
- Datos de Train y Test.

REGRESION LINEAL ORDINARIA (LSS)

- MSE en datos de entrenamiento (67)

  MSE = 0.439

- MSE en datos de prueba (30)

  MSE = 0.521

- MSE en Cross-Validation K = 5

  MSE = 0.957

- MSE en Cross-Validation K = 10

  MSE = 0.757

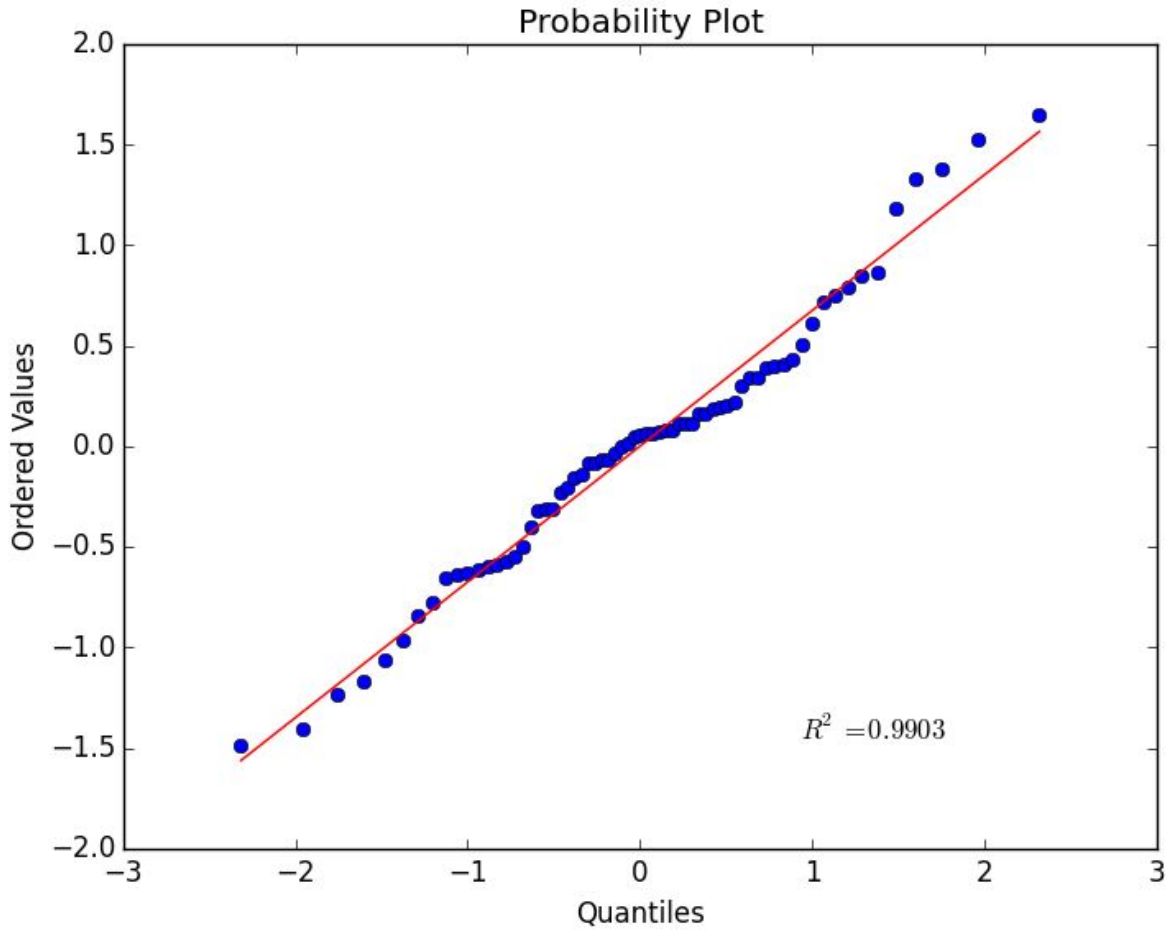
REGRESION LINEAL ORDINARIA (LSS)

- t-Student con 58 gl(67-8-1):
- a=0.05 -> 1.6716
- lcavol
- lweight
- svi
- age, gleason, pgg45 no tienen significancia.

TABLA DE PESOS Y Z-SCORE (TRAIN SET)

Atributo	Coeficiente b	Z-Score
Intercept	2.4649	27.3592
lcavol	0.6760	5.3198
lweight	0.2616	2.7269
age	-0.1407	-1.3838
lbph	0.2090	2.0380
svi	0.3036	2.4478
lcp	-0.2870	-1.8507
gleason	-0.0211	-0.1454
pgg45	0.2655	1.7227

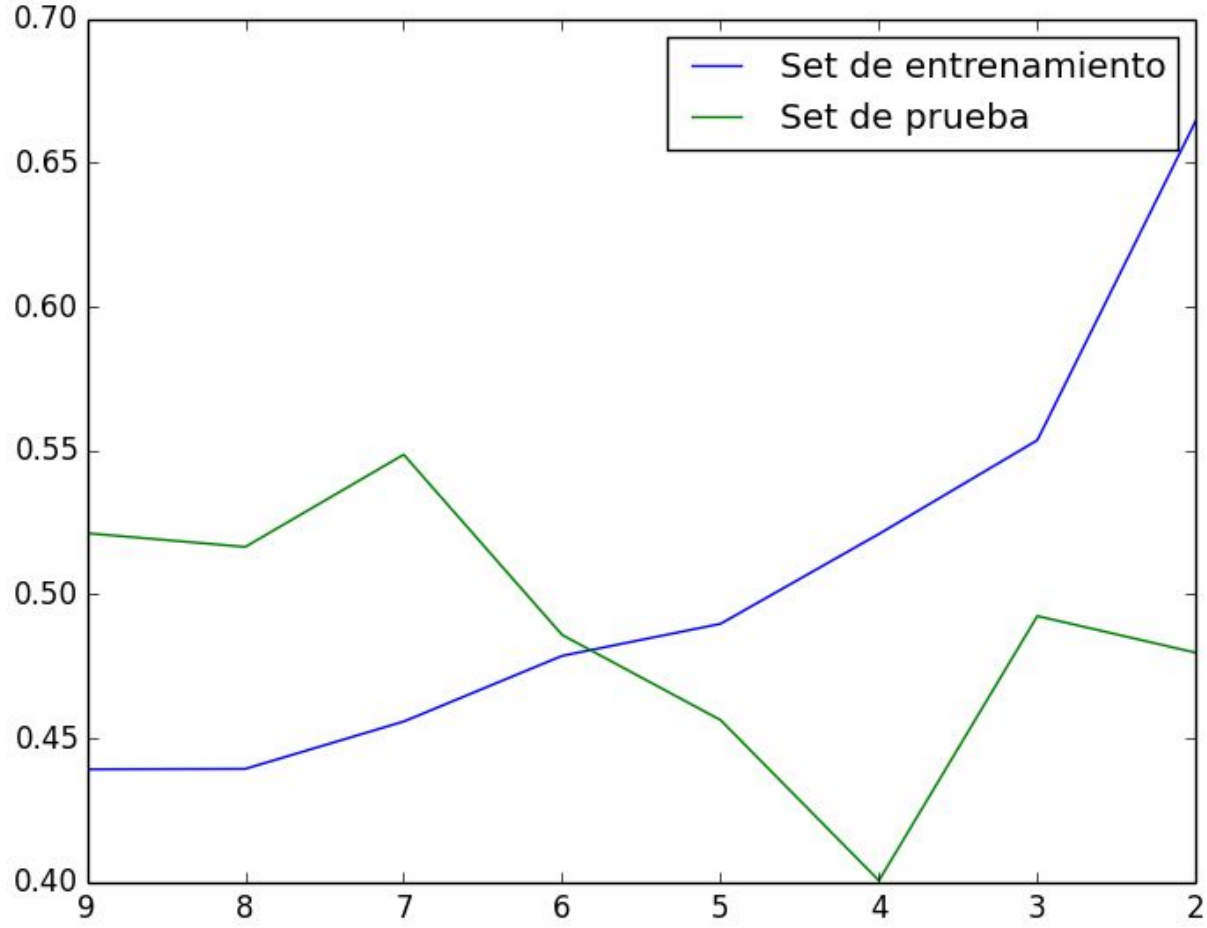
NORMALIDAD DE RESIDUOS



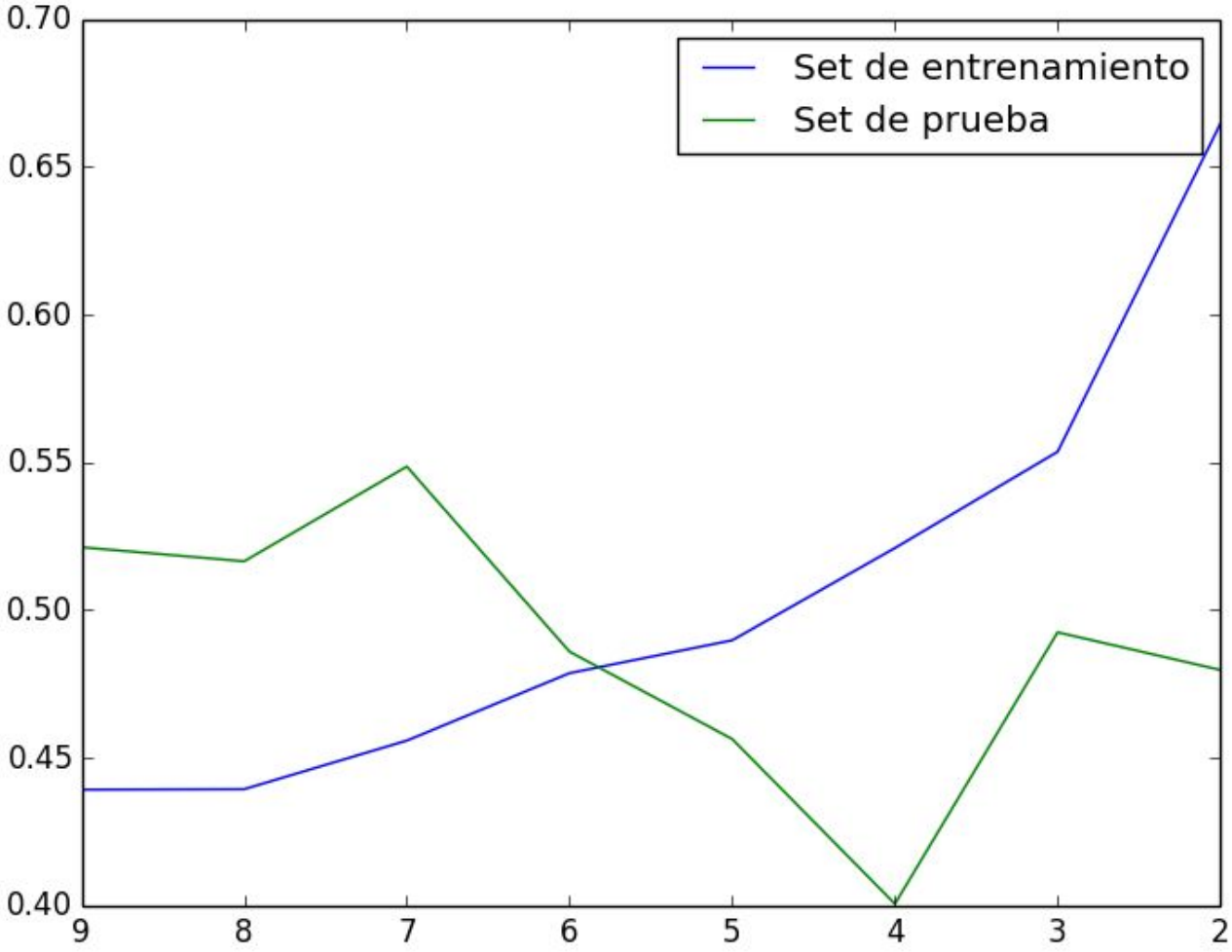
Selección de atributos

Se pide realizar Backward Step-wise y Forward Step-wise para selección de atributos, la primera consiste en partir con todas las variables y eliminar una a una aquella que tiene el menor peso, determinado por el Z-score, mientras que la segunda, busca partir desde la menor cantidad de parámetros hasta llegar a la mayor, agregando en cada iteración aquel más importante. La técnica se aplica al mismo dataset entregado para la pregunta 1 y se obtienen los siguientes resultados

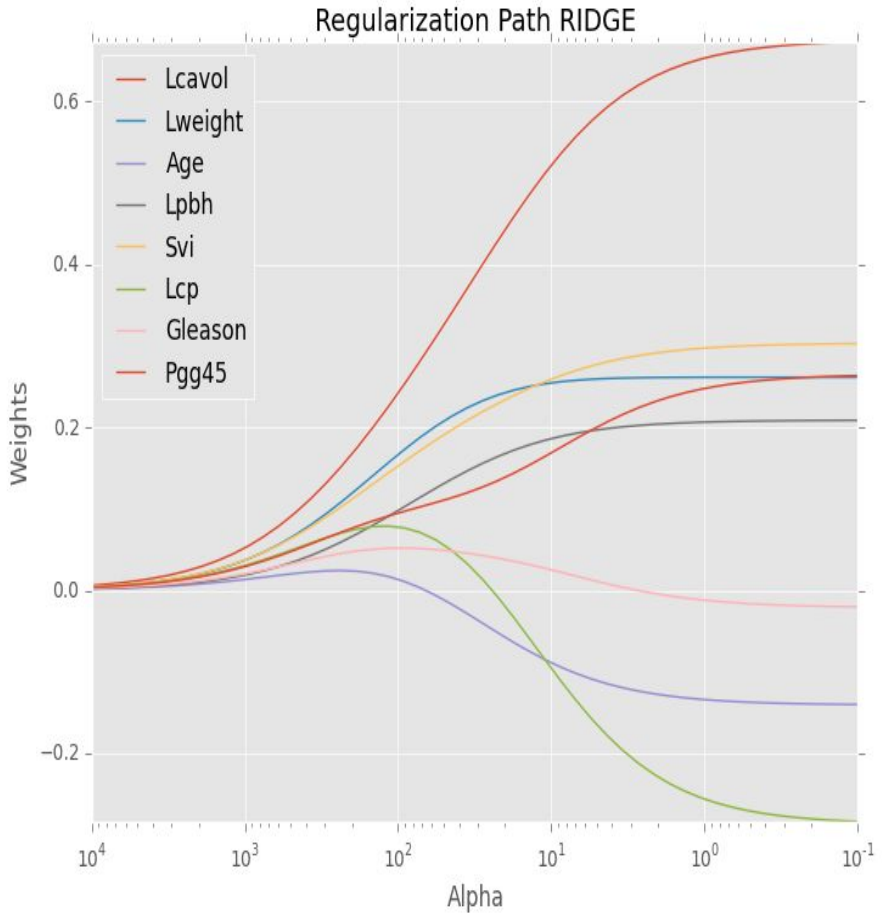
Error en función del número de parámetros(FSS)



Error en función del número de parámetros(BSS)

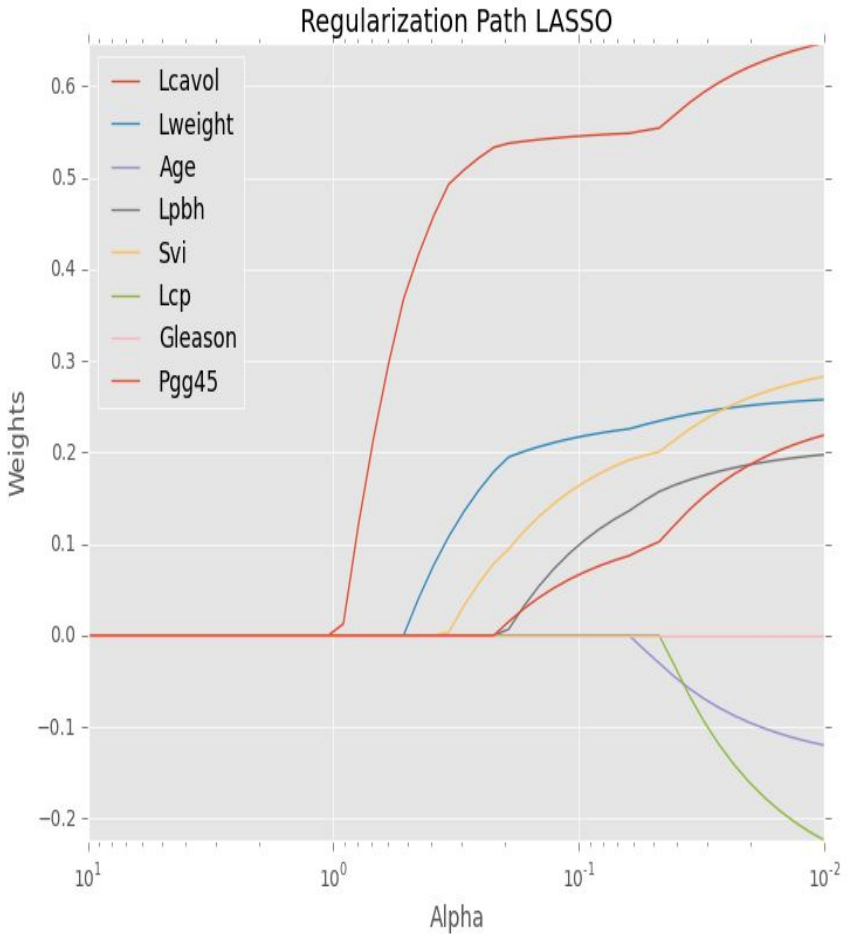


COEFICIENTES DE RIDGE EN FUNCION DE PARAM DE REG



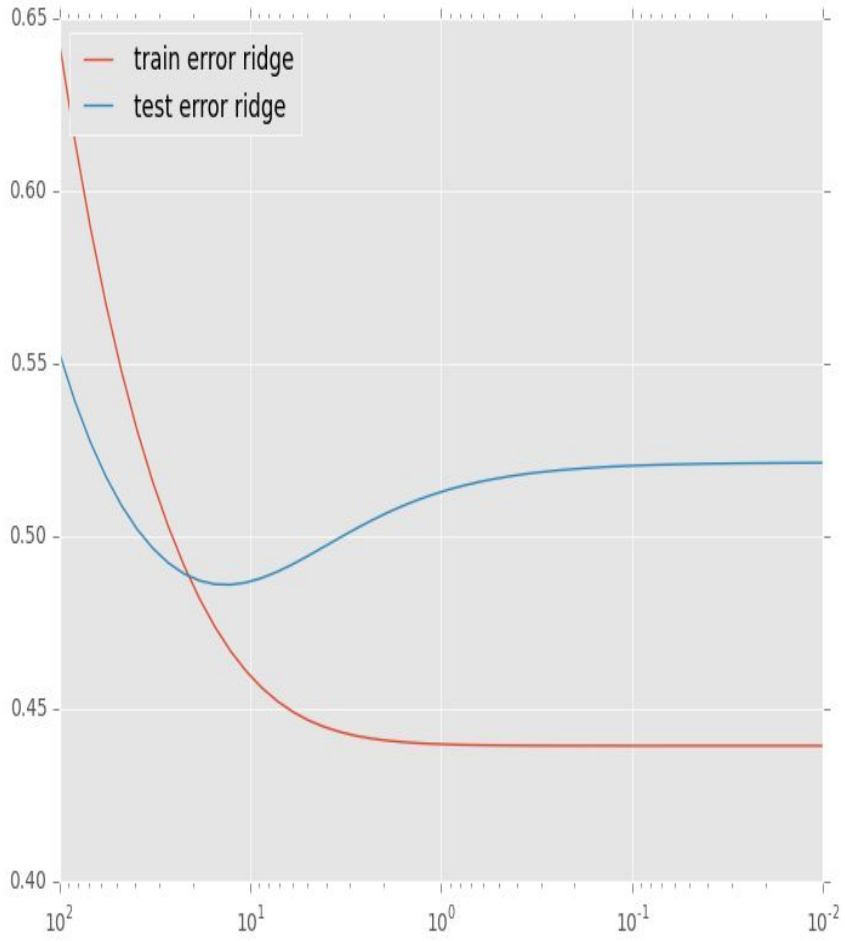
- Tiende a LSS con alpha pequeño.
- Alpha grande sobre-regulariza.
- Luego de 10^3 no hay significancia, pero nunca es 0.
- Poco efecto, hasta 10.

COEFICIENTES DE LASSO EN FUNCION DE PARAM DE REG



- Lcavol, Lweight, svi siguen con mayor peso (a<0.5)
- 0.1<a<1 se hacen lweight, svi = 0.
- a>1 desaparece el modelo.
- a>0.1, lcp y age desaparecen.
- Alpha grande sobre-regulariza.

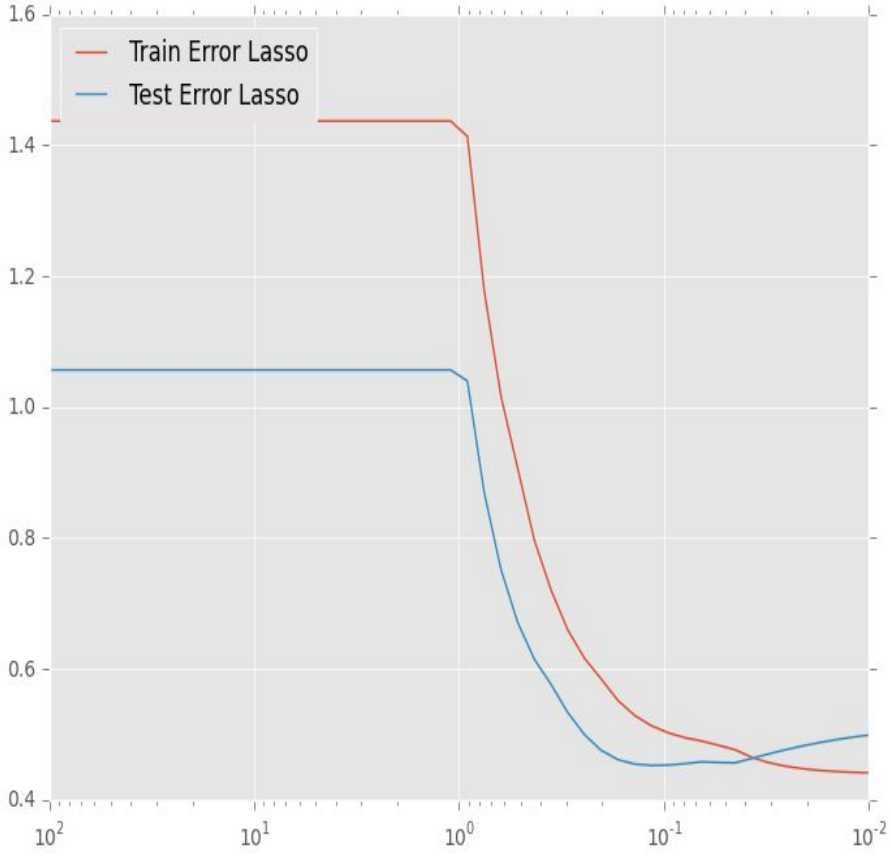
COMPARACION ERROR TRAIN Y TEST, RIDGE REGRESSION



- a>1 errores constantes
  - train ~0.53
  - test ~0.44
- error de train menor al de test (a<30)
- error de test menor al de train (a>30)
- Train siempre baja al disminuir alpha.

COMPARACION ERROR TRAIN Y TEST, LASSO REGRESSION

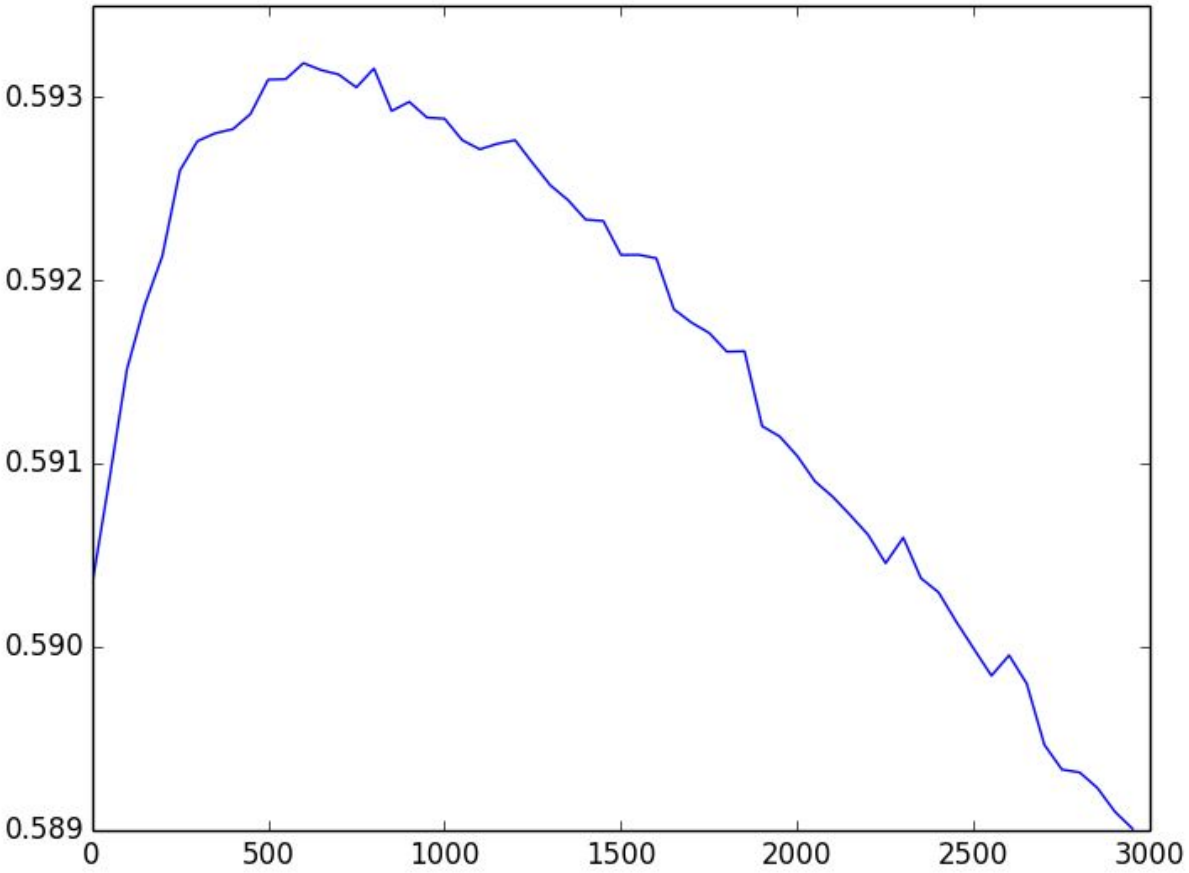
- Baja el MSE a medida que baja alpha.
- a<1 errores disminuyen rápidamente.
- Error en a>1:
  - train ~1.42
  - test ~1.03
- Test más bajo hasta a>0.05.

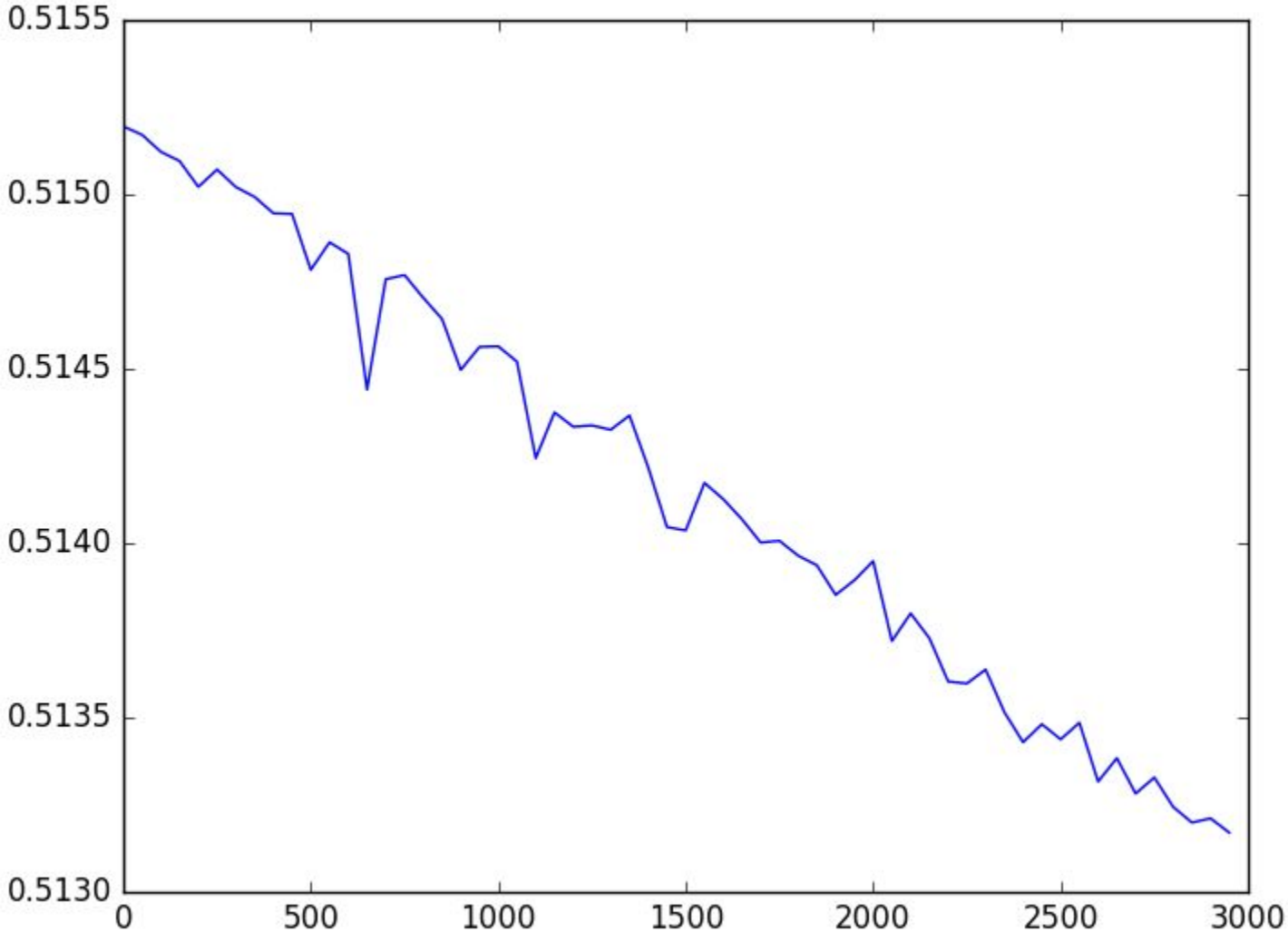


Predicción de utilidades de una película

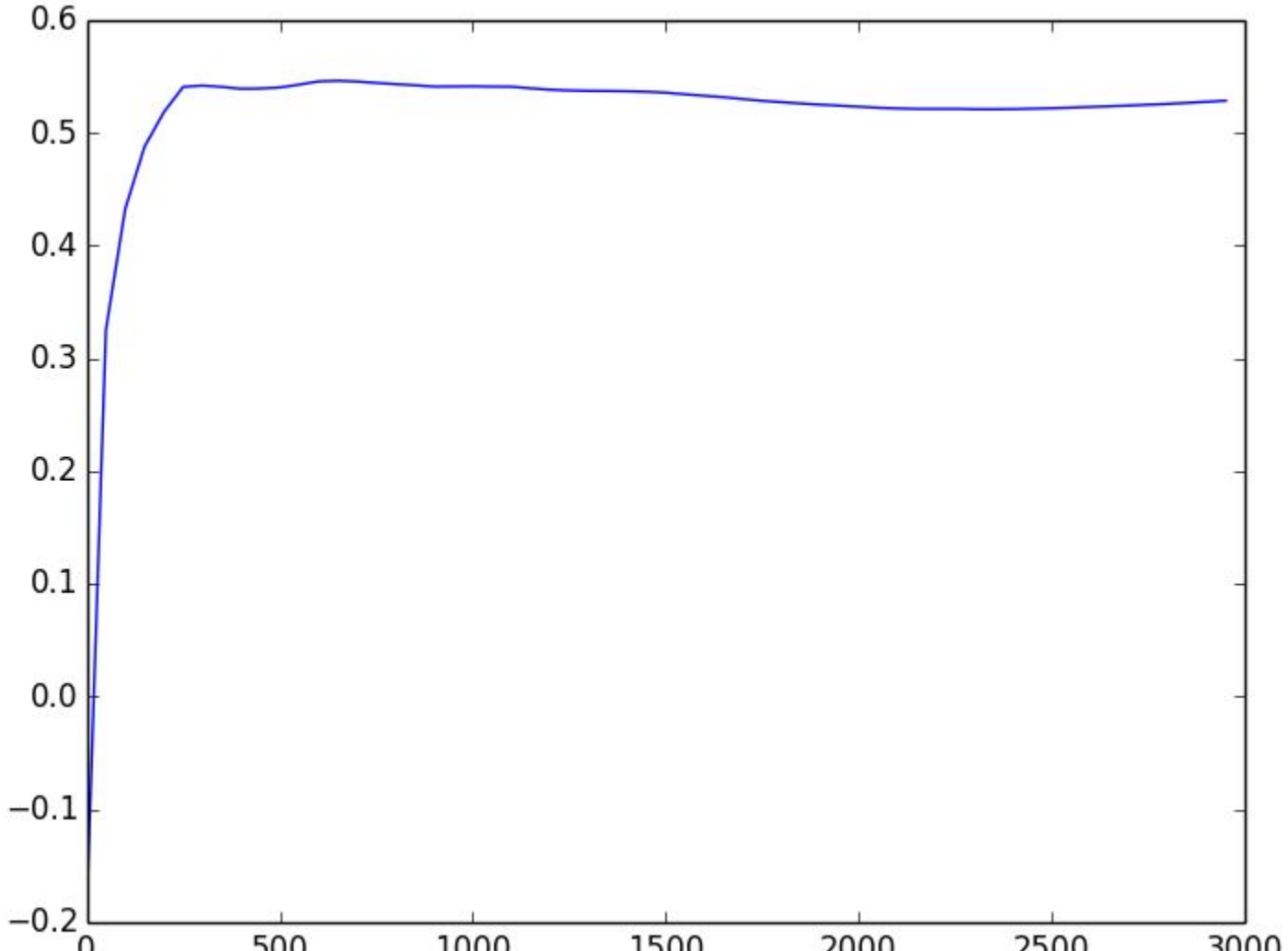
En esta parte, se intenta encontrar un predictor para las utilidades de una película,a través de ciertos atributos como lugar de origen, presupuesto, número de puntos de proyección, etc. Para esto, se probarón 3 métodos, el primero una regresión lineal sin regularizar, el segundo una regresión lineal por el metodo Lasso, y por último, una regresión lineal por el método Ridge.

Coeficiente de determinación en función de alfa (Ridge)




0.5155
0.5150
0.5145
0.5140
0.5135
0.5130
0
500
1000
1500
2000
2500
3000

Coeficiente de determinación en función de alfa (lasso)



GRACIAS POR SU ATENCIÓN!