Tarea 1: Métodos Lineales para Regresión

...

Juan Pablo Castillo
Álvaro Rojas

Introducción

Se utilizaran métodos de regresión lineal multivariado estudiados en cátedra para analizar los siguientes dataset:

- Prostate-cancer: datos de pacientes luego de haberles aplicado prostatectomía radical.
- Datos sobre utilidades en estreno de Películas en USA.

Tópicos principales: Regresión Lineal Ordinaria (LSS), Selección de atributos y Regularización.

Descripción del Dataset “Prostate-Cancer”

- 97 datos con 9 atributos cada uno
- 4 atributos de tipo entero y 5 de tipo decimal

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
count	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000	97.000000
mean	1.350010	3.628943	63.865979	0.100356	0.216495	-0.179366	6.752577	24.381443	2.478387
std	1.178625	0.428411	7.445117	1.450807	0.413995	1.398250	0.722134	28.204035	1.154329
min	-1.347074	2.374906	41.000000	-1.386294	0.000000	-1.386294	6.000000	0.000000	-0.430783
25%	0.512824	3.375880	60.000000	-1.386294	0.000000	-1.386294	6.000000	0.000000	1.731656
50%	1.446919	3.623007	65.000000	0.300105	0.000000	-0.798508	7.000000	15.000000	2.591516
75%	2.127041	3.876396	68.000000	1.558145	0.000000	1.178655	7.000000	40.000000	3.056357
max	3.821004	4.780383	79.000000	2.326302	1.000000	2.904165	9.000000	100.000000	5.582932

Normalización de datos

- Se realiza una normalización de datos respecto a la característica “**lpsa**”. Ajustando a las demás a esta escala.
- Permite que los valores de los distintos atributos tengan un peso adecuado para el momento de la regresión lineal.
- Evita que las variables con mayor orden de magnitud opaquen la importancia de las demás

Pesos y Z-score

- El atributo con mayor correlación dado su peso es lcavol(0.676016), lo que la hace la variable con mayor correlación con la respuesta. A esta le siguen svi(0.303623) y lweight(0.261694)
- Los predictores con un Z-score absoluto menor a 2.002(t^{67-9}) no tienen la suficiente relación con los datos de salida.
- Los predictores age(-1.383823), lcp(-1.850749), gleason(-0.145411) y pgg45(1.722793) no son significantes.

Atributo	Peso	Z-Score
lcavol	0.676016	5.319828
lweight	0.261694	2.726973
age	-0.140734	-1.383823
lbph	0.209061	2.038046
svi	0.303623	2.447876
lcp	-0.287002	-1.850749
gleason	-0.021195	-0.145411
pgg45	0.265576	1.722793
Intercept	2.464933	27.359253

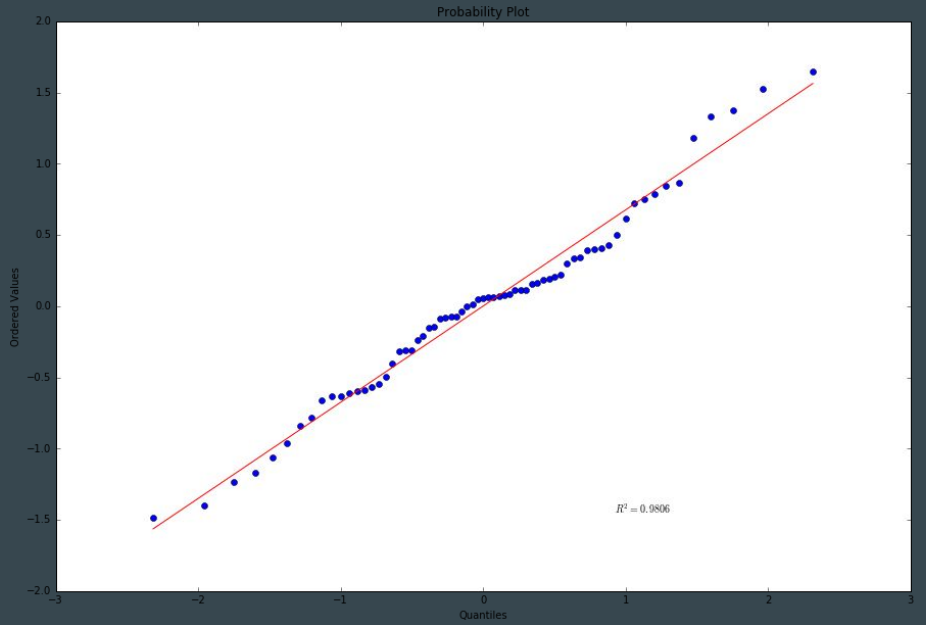
Estimación del error del modelo con Cross-Validation

```
El error de predición real usando el test_set es 0.521274
El error de prediccion usando CV con K=5 es 0.956515
El error de prediccion usando CV con K=10 es 0.757237
```

- Con k=5 se utiliza una partición 80/20 entre entrenamiento y validación respectivamente. En cambio con k=10 se utiliza una partición 90/10.
- Es intuitivo pensar que si se entrena con mayor cantidad de datos se obtendrá un menor error, pero esto es a costa del sobreajuste.
- Si se compara con el error real, se observa que los resultados obtenidos con CV no son del todo buenos, ya que incluso se supera el valor de error de la regresión lineal. Esto denota que el modelo tiene un sobreajuste respecto a los datos de entrenamiento, por lo tanto tiene una mala predicción respecto a los datos de prueba.

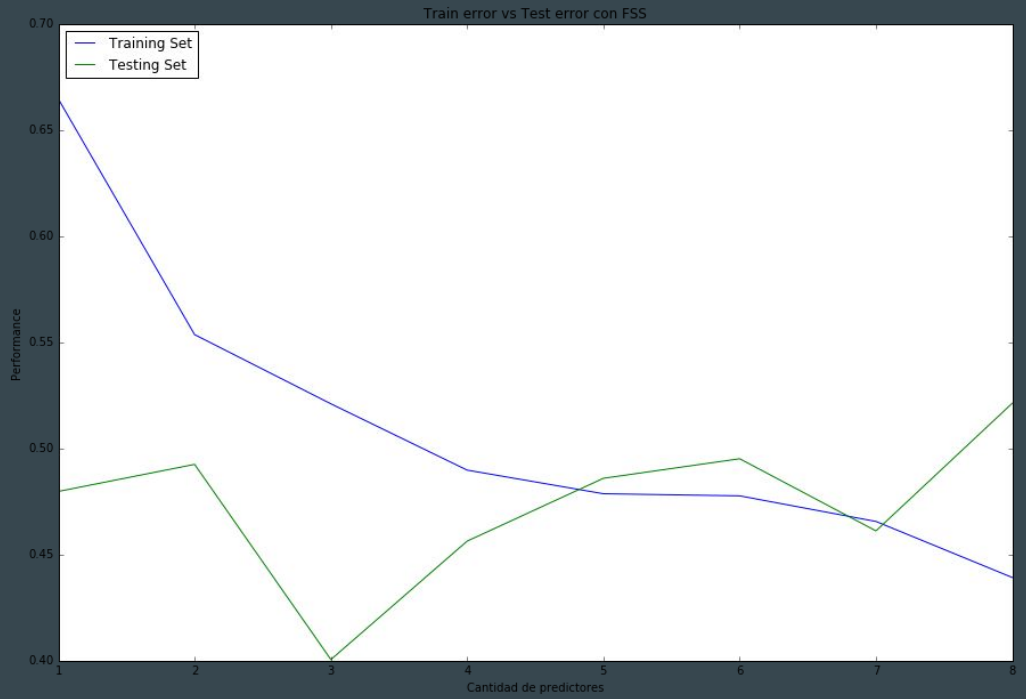
Estimación del error de cada dato de entrenamiento

- Los errores se calcularon como la distancia de cada dato real al estimado.
- Se aprecia una correlación entre las variables con el resultado que se busca.
- Se comprueba que la hipótesis de normalidad es acertada.



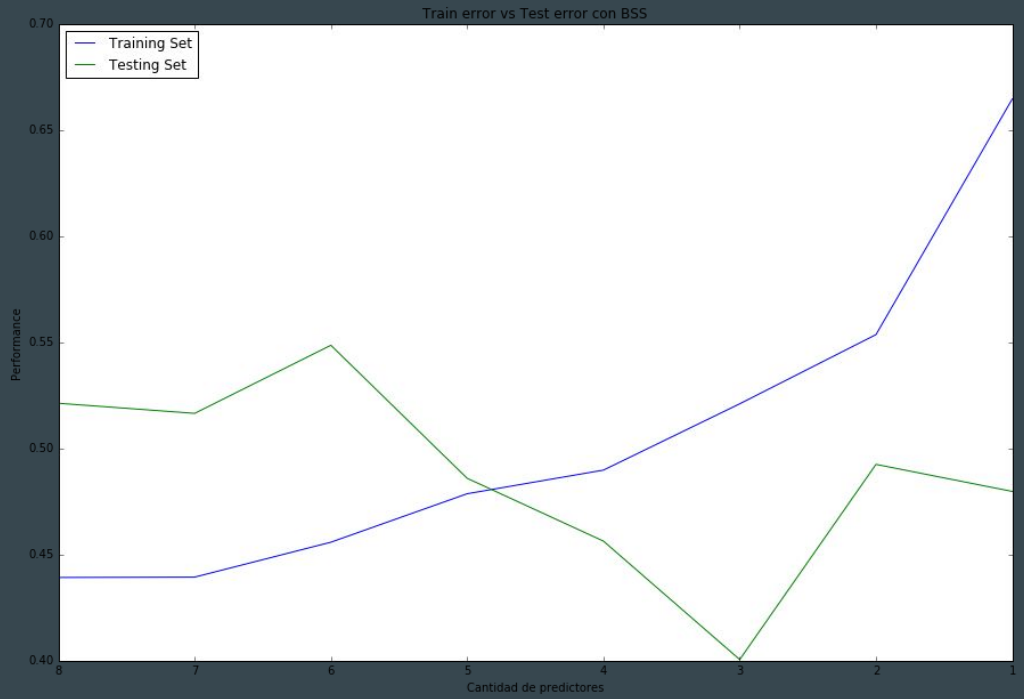
Selección de Atributos (Forward Step-wise Selection)

- Los valores son escogidos según el que minimiza el error.
- El error se calculó según la distancia de los z-score.
- El mínimo error para el training son las 8 variables.
- el mínimo error para el testing son 3 variables (lcavol, lweight,Svi)



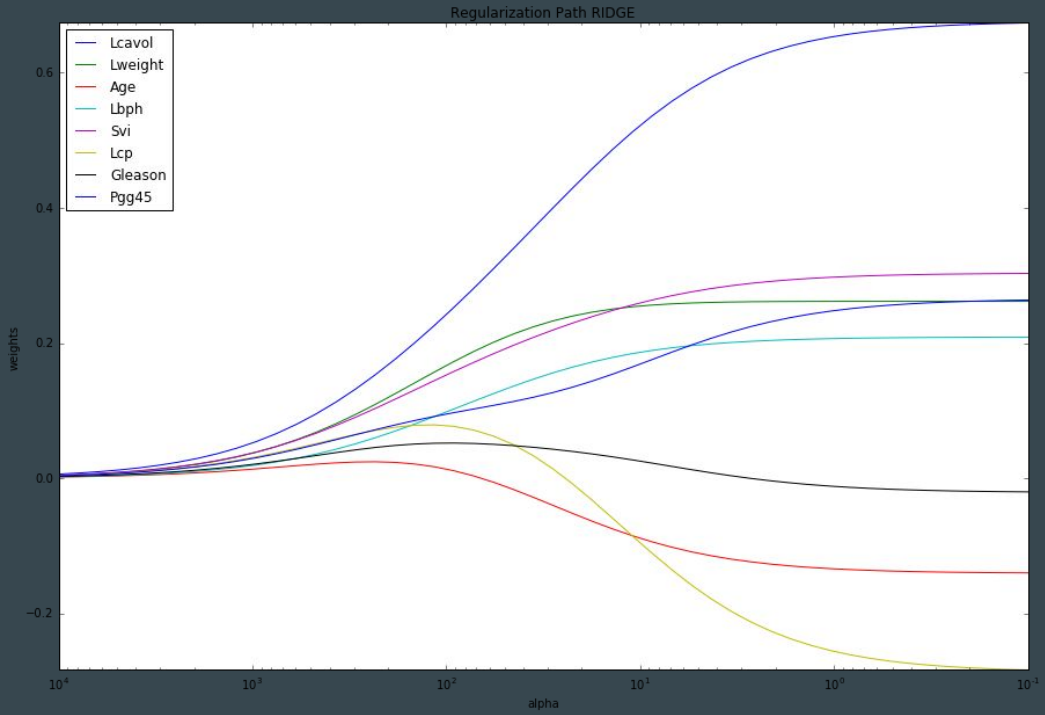
Selección de Atributos (Backward Step-wise Selection)

- Los valores son dejados según los que minimizan el error.
- El error se calculó según la distancia de los z-score.
- El mínimo error para el training son las 8 variables.
- el mínimo error para el testing son 3 variables (lcavol, lweight,Svi)



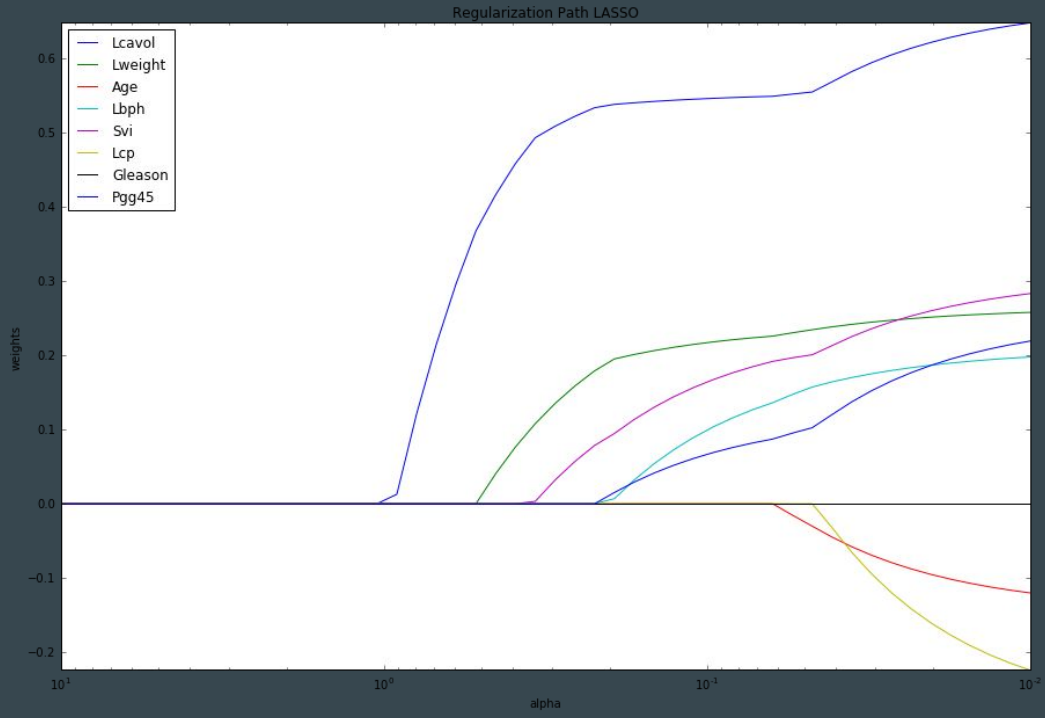
Regularización (Ridge Regression)

- A medida que el valor de alpha aumenta las variables sufren una penalización paulatina.
- El atributo de mayor importancia es Lcavol, que requiere de alpha 10000 para perder significancia.
- Las otra variables sufren menor penalización pero llegan a la insignificancia antes.



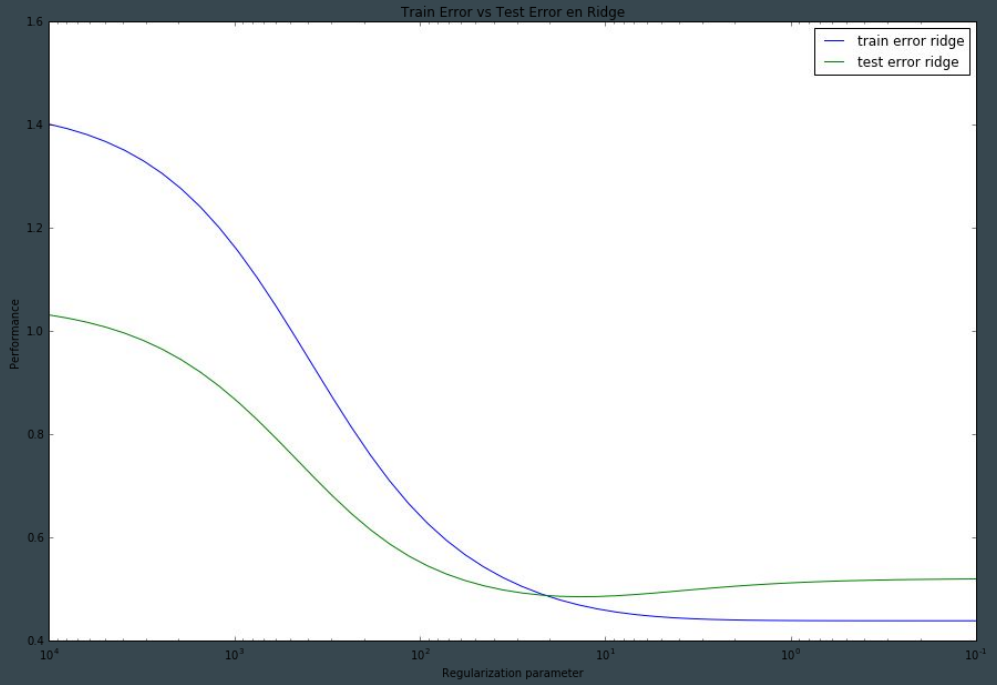
Regularización (Lasso)

- La importancia y pérdida de significancia de los atributos es similar a Ridge Regression.
- Los pesos de las variables disminuyen de una forma más brusca.
- Con lasso la selección de variables es más notoria a medida que se penaliza, por lo que en la práctica es bastante efectivo



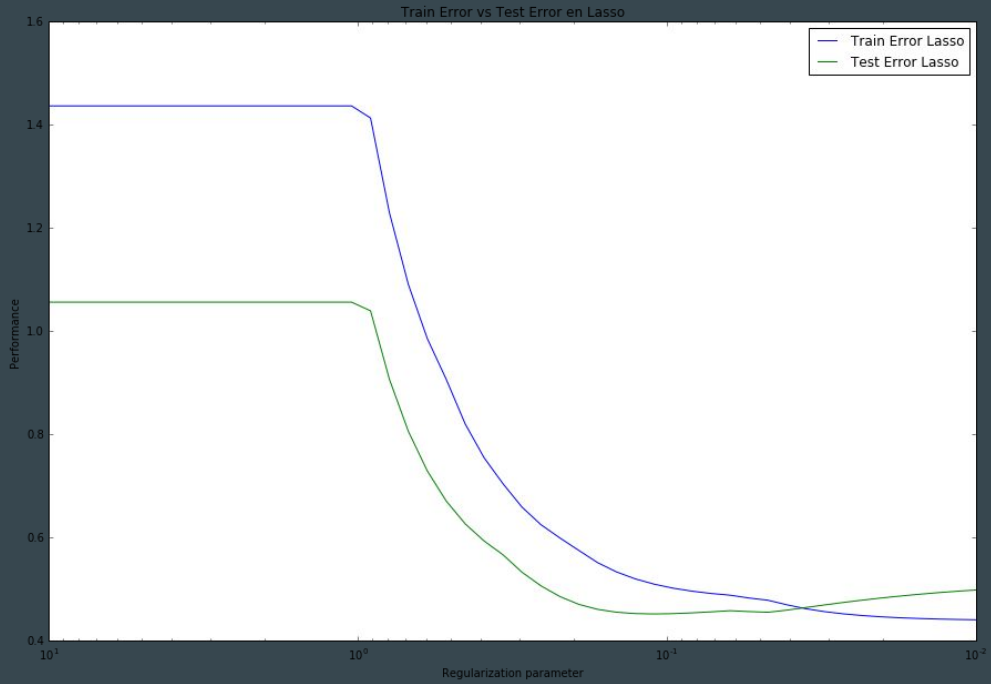
Errores de entrenamiento y pruebas (Ridge Regression)

- Las variables pueden perder su importancia, por lo que penalizar demasiado a las variables es contraproducente para el modelo.
- Para este caso en particular lo mejor es aplicar un alfa cercano a 10 para disminuir el test error.



Errores de entrenamiento y pruebas (Lasso)

- Pierde su utilidad cuando sobrepasa el valor 1 debido a que ya castigó los pesos de todas las variables.
- Comparando con los errores anteriores, se observa nuevamente la diferencia en el orden de magnitud de los alphas para los cuales se castiga a los atributos.



Estimación de los parámetros de regularización con CV

Algunos valores representativos de lo obtenido son:

Ridge

Alfa	MSE
10000	1.764186
232.995181	1.191415
2.120951	0.751909

Lasso

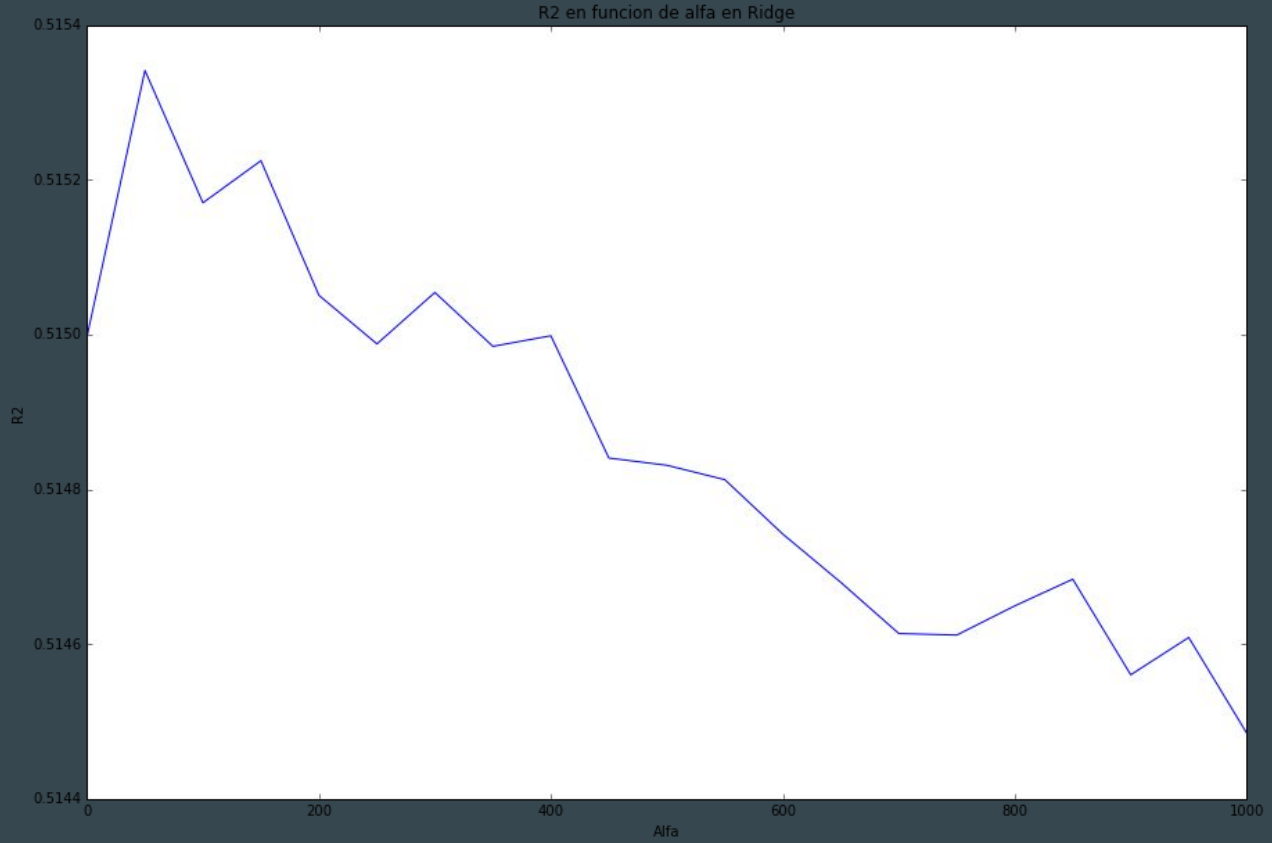
Alfa	MSE
10.000000	1.795596
0.109854	0.882670
0.010000	0.758661

Predicción de Utilidades de Películas

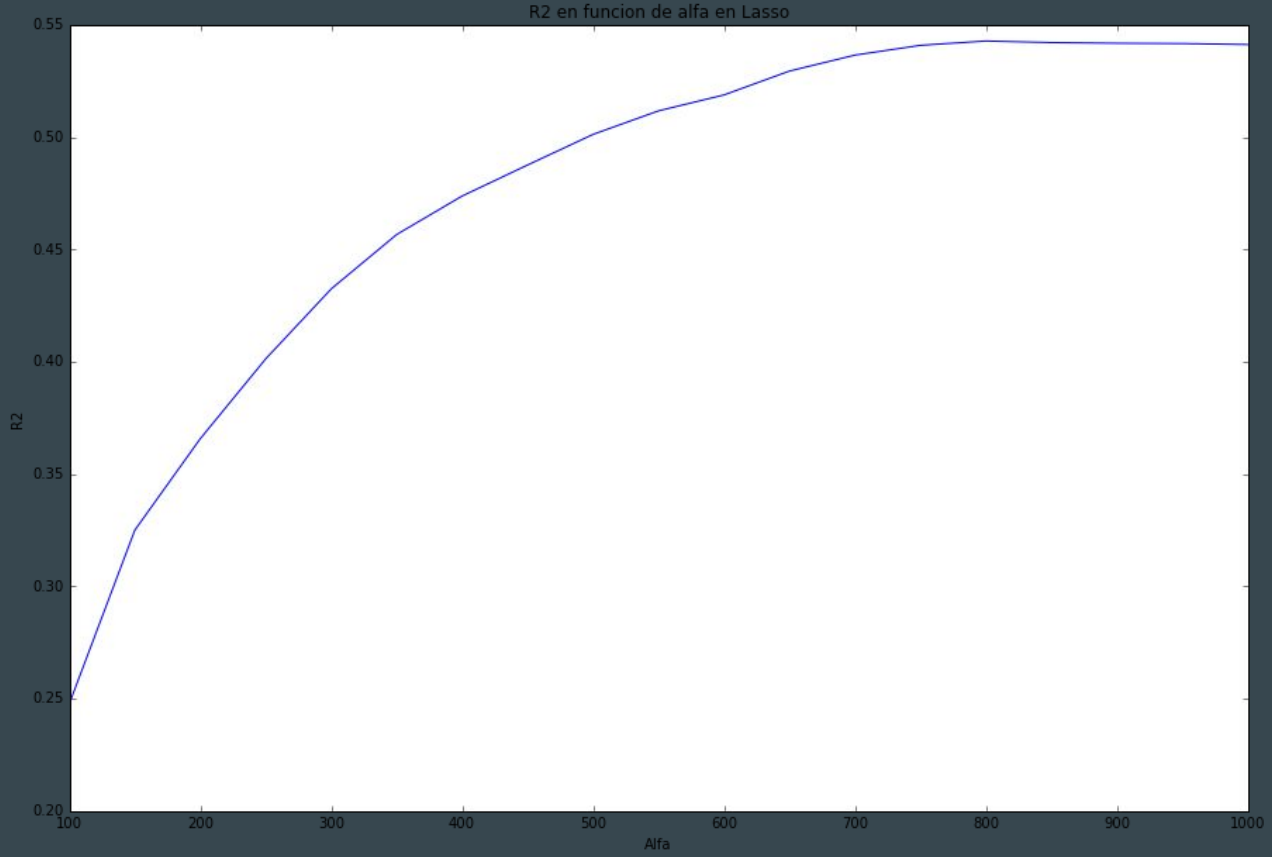
- Se probaron dos métodos para intentar predecir las utilidades considerando un modelo con un R^2 mayor o igual a 0.75. Uno basándose en la regularización de Ridge Regression y el segundo con Lasso.
- Ambos modelos son regresiones lineales simples.
- Los mejores valores de R^2 obtenidos para cada modelo probando diferentes alphas son:

	R^2	Alfa
Ridge regression:	0.515276	50
Lasso	0.542747	800

Predicción de Utilidades de Películas (modelo Ridge)



Predicción de Utilidades de Películas (modelo Lasso)



Comentarios y reflexiones (Conclusiones)

- Notamos al trabajar los datos que los valores no son ideales en la realidad, requiriendo de trabajo y estandarización para solamente comenzar a analizar y estimar con regresiones.
- Siempre existe la posibilidad del mal del sobreajuste, el cual influyó en varias partes del trabajo realizado, alejando los errores de los modelos respecto a los datos de prueba y entrenamiento.
- Cross-Validation es una buena medida de ver que tan bueno quedo entrenado el modelo y también para estimar un buen parámetro (alfa) para la re-parametrización de los predictores.