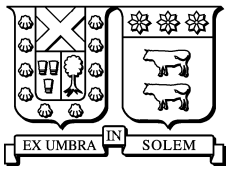
Análisis Inteligente de Datos

Tarea 2

Felipe Flores - Anibal Pérez

24 de Junio de 2016



1. Introducción

El siguiente informe abordará la implementación y análisis de los resultados obtenidos por distintos métodos lineales de regresión, como lo es la regresión por mínimos cuadrados, regresión Ridge y regresión Lasso. De la misma manera, se abordará validación cruzada, métricas de selección de atributos, y regularización con algunas de estas regresiones. Todo esto para abordar la predicción y medición de errores de estas, sobre un dataset de una enfermedad muy común en el grueso de los hobmres de edad, el cáncer de próstata.

Se finaliza el análisis con un análisis predictivo basado en regresiones de Ridge y Lasso, para obtener coeficientes de correlación relevantes en la estimación de utilidades para un gran número de películas en sus estrenos.

2. Regresión Lineal Ordinaria (LSS)

2.1. Descripción de los datos

El dataset original cuenta con 11 columnas, 1 de autoincremento, 8 utilizadas como predictores y 1 que indica la respuesta (log prostatic specific antigen) [1], se listan las 10 principales a continuación :

- **lcavol:** Variable numérica continua, que representa el logaritmo del volumen de cáncer medido en la próstata del paciente.
- **lweight:** Variable numérica continua, correspondiente al logaritmo del peso de la próstata del paciente.
- **Age:** Variable numérica discreta correspondiente a la edad del paciente.
- **lbph:** Variable numérica continua, correspondiente al logaritmo de la cantidad de hiperlasia prostática benigna, representa el agrandamiento de la próstata (lo que produce las constantes ganas de orinar, por ejemplo, en este tipo de pacientes).

- **svi:** Variable numérica binaria, toma el valor 0 si no hay invasión de la vesícula seminal, 1 en el caso contrario. Es un dato histológico de mal pronóstico en pacientes de este tipo.
- **lcp:** Variable numérica continua, corresponde al logaritmo de la penetración capsular en el paciente.
- **gleason:** Variable numérica discreta, representa la escala de Gleason medida en el paciente, da el grado de agresividad del cáncer de próstata en el paciente, obtenida por una biopsia, de 7 hacia arriba representa cáncer con agresividad intermedia/alta.
- **pgg45:** Variable numérica discreta, corresponde al porcentaje del indicador de gleason con grado 4 de 5 observado en el paciente.
- **lpsa:** Variable numérica continua, corresponde al logaritmo del nivel de antígeno prostático específico medido en el paciente.
- **train:** Variable categórica nominal, indica si la fila corresponde a un dato de entrenamiento o de prueba, con T y F respectivamente.

2.2. Implementación

a Luego de construir el dataframe, es necesario quitar de este las columnas que no entreguen valor alguno al modelo de regresión que se quiere alcanzar, es por esto que se elimina la primera columna sin nombre (unnamed), que contiene el contador de registro. Otra columna que no entregará valor al análisis numérico será la que indica si el dato es de entrenamiento o de prueba, es por esto que se copian los datos de esta en un arreglo que se asociará por índice al número de registro, este arreglo será utilizado posteriormente para separar los datos de entrenamiento de los de prueba. Finalmente se elimina del dataframe la columna recién mencionada.

b Con los comandos utilizados en esta parte, se puede apreciar que el dataset a utilizar queda con **97 registros y 9 columnas; 5 flotantes y 4 enteras; no falta ningún registro y utiliza un espacio de 7,6 KB**. De las columnas con logaritmo no mucho se puede decir, pero en las continuas se aprecia un promedio en la edad, por ejemplo, de casi 64 años; más de este apartado puede ser presenciado al ejecutar el archivo **lss.py**.

c Se deben normalizar los datos antes de realizar un análisis sobre estos, esta parte es crucial dado que el escalador estándar de scikit-learn deja el dataframe con media nula y varianza unitaria, paso importante para poder obtener coeficientes de regresión con sentido entre sí, pues normaliza las magnitudes de las distintas columnas a pesar de su distinto significado lógico/físico. Se desnormaliza (se copia del dataframe no normalizado) la columna de respuesta lpsa, dado que está no será utilizada por la matriz de datos principal.

d Para formar la matriz de datos principal **X**, se deja fuera la columna de respuesta lpsa, pues claramente presentaría una correlación perfecta al querer predecir con respecto a ella misma. En este paso se guarda el número de datos totales, **paso que no se considera necesario, dado que el número de datos requerido para los algoritmos es, o el número de datos de entrenamiento, o el número de datos de prueba**, los cuales pueden ser obtenidos obteniendo las dimensiones de la matriz **Xtrain**, como bien de la matriz **Xtest**, respectivamente.
Se utiliza el regresor lineal de scikit-learn LinearRegression, a este se le deben pasar los argumentos necesarios para como estén trabajados los datos, en este paso los datos ya se encuentran normalizados y con un intercepto igual a 1, por lo que **se debe pasar como argumento a esta funcion fit-intercept=False, para que no calcule intercepto al modelo**. Finalmente se entregan al modelo lineal los **datos de entrenamiento para ajustar el modelo al**

conjunto de datos que más lo representa, y así poder abordar correctamente la siguiente pregunta.

e A continuación se muestra el Cuadro 1, con los pesos (valor del coeficiente b) y Z-Score de los atributos pertenecientes al modelo de datos de entrenamiento:

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

Cuadro 1: Tabla de pesos y z-score para atributos de prostate-cancer, bajo regresión lineal de mínimos cuadrados básica.

De la anterior tabla se puede desprender que variables están más correlacionadas con la respuesta lpsa, de la siguiente manera:

- Se sabe que el Z-Score equivale a una división entre una normal y una Chi-Cuadrado, lo que entrega una t-student, esta tiene $(N-d-1)$ grados de libertad, N igual al número de datos de entrenamiento (67), y d igual al número de predictores o atributos del modelo (8).
- Se debe buscar entonces en la tabla t-Student [2], el valor entregado para 58 grados de libertad, dada una significancia de $alpha = 0{,}05$, el cual es **1.6716**, esto significa que cualquier valor que esté en el rango $[-1{,}6716, 1{,}6716]$ no tendrá significancia con la respuesta.
- Es así como se identifica mirando el Cuadro 1, que los atributos **lcavol (volumen de cáncer en la próstata), lweight (peso de próstata) y svi (invasión de vesícula seminal)** son los **más correlacionados** con la respuesta en este nivel de significancia.
- Por otro lado, se **descartaría la significancia** de las variables **age (edad) y gleason (medición de agresividad del cáncer) y pgg45, pues no guardan correlación** con la respuesta en este nivel de significancia.

f Como se podrá apreciar en la ejecución del código **lss.py**, el Mean Squared Error para una validación cruzada con K=5 y K=10, entrega los valores **0.957 y 0.757** respectivamente. Si esto es comparado con los MSE obtenidos por los datos de entrenamiento y de prueba, con valores **0.439 y 0.521** respectivamente, se puede desprender que a mayor K, mejor interpreta la validación cruzada el modelo, pero aún así estos dos valores probados no lograron mejorar el error cuadrático que los datos de entrenamiento o de prueba entregaron, por lo que se cree debe probarse con valores mucho más grandes de K, llegando probablemente a que cuando este sea cercano al número de datos, el MSE llegará a ser el mismo que calculado sin Cross-Validation.

g Dado que el error de predicción es la diferencia entre el predictor y el valor real de respuesta, se puede obtener la diferencia para todos los datos y luego graficarlos, para así compararlos con los cuantiles de una distribución normal, como se puede apreciar en la Figura 1. De esta se desprende en primer lugar un muy parecido comportamiento, y en segundo lugar, un

coeficiente de correlación bastante elevado, lo que lleva claramente a concluir que la hipótesis de normalidad de los residuos es efectivamente cierta.

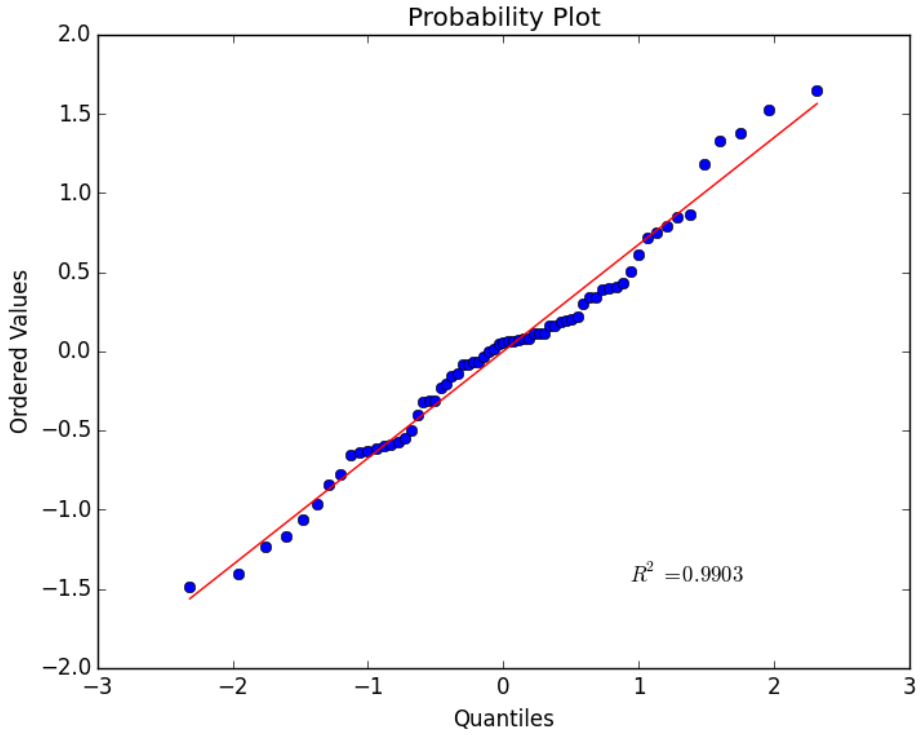


Figura 1: Datos de entrenamiento con linea de tendencia ajustada a la de una normal, con coeficiente de correlación.

3. Selección de Atributos

a Al realizar FSS, agregando atributos uno por uno se obtiene el siguiente gráfico, donde el eje x corresponde al número de atributos y el eje y corresponde al error cuadrático medio, tanto para los datos de entrenamiento, como para los datos de prueba. Como es de esperar, el error de entrenamiento disminuye al utilizar un número mayor de atributos, logrando su mínimo con 9 atributos y esto se debe a que al seleccionar un número tan grande de atributos, se tiene un overfitting, por lo cual es necesario contrarrestar con el error de prueba, donde se aprecia que existe un mínimo al seleccionar 4 atributos.

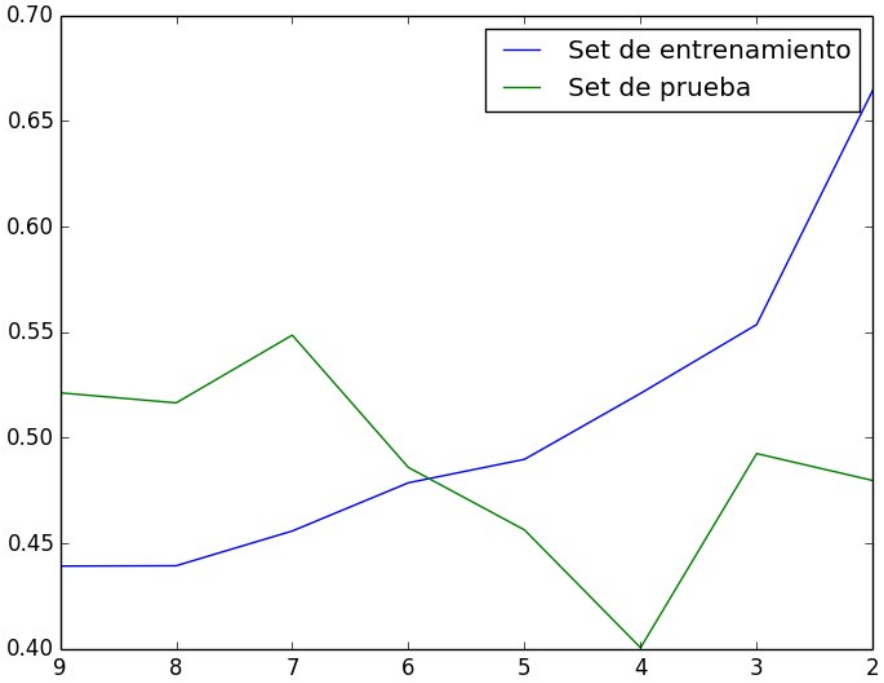


Figura 2: Gráfico de error de prueba y error de entrenamiento por el método de selección de atributos Forward Step-wise, en función del número de atributos.

b Para el caso del método de selección de atributos Backward Step-wise, se aprecia un fenómeno idéntico al anterior, esto debido a que ambos métodos usan el mismo criterio, solo aplicando en orden inverso. Nuevamente, el error de entrenamiento disminuye al utilizar más atributos, y al igual que en el caso anterior, esto es por que existe un sobre ajuste, pero al contrarrestar esto con los datos de prueba, se ve que el error de prueba al pasar de los 4 parámetros comienza a aumentar de manera considerable.

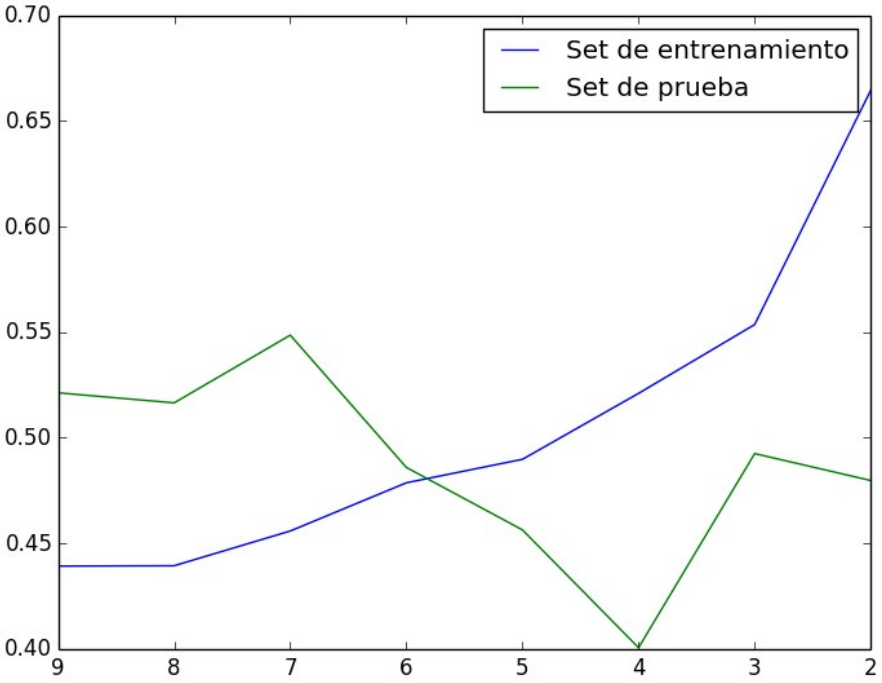


Figura 3: Gráfico de error de prueba y error de entrenamiento por el método de selección de atributos Backward Step-wise, en función del número de atributos.

4. Regularización

4.0.1. Implementación

a Como se puede observar en la Figura 4, los valores que toma el parámetro de regularización aumenta o disminuyen el peso de los coeficientes de la regresión lineal de Ridge. Se concluye que a mayor valor de alpha, mayor es la sobreregularización del modelo, por lo que sus coeficientes comienzan a tener **pesos practicamente nulos luego de** $alpha = 1000$**, notar que estos solo se acercan al valor cero dada la forma de cálculo de esta regresión (cuadrática)**. Por otro lado, se observa que para alphas menores a 10, el peso de los coeficientes empieza a verse más afectado por el parámetro de regularización, haciéndose cada vez más cercano al valor que estos tendrían si fuera una regresión por mínimos cuadrados. Se observa claramente que las variables lcavol, lweight y svi son las que representan mejor la respuesta, dado su peso notoriamente mayor. Por lo tanto, se concluye que a menor alpha, menor es la regularización.

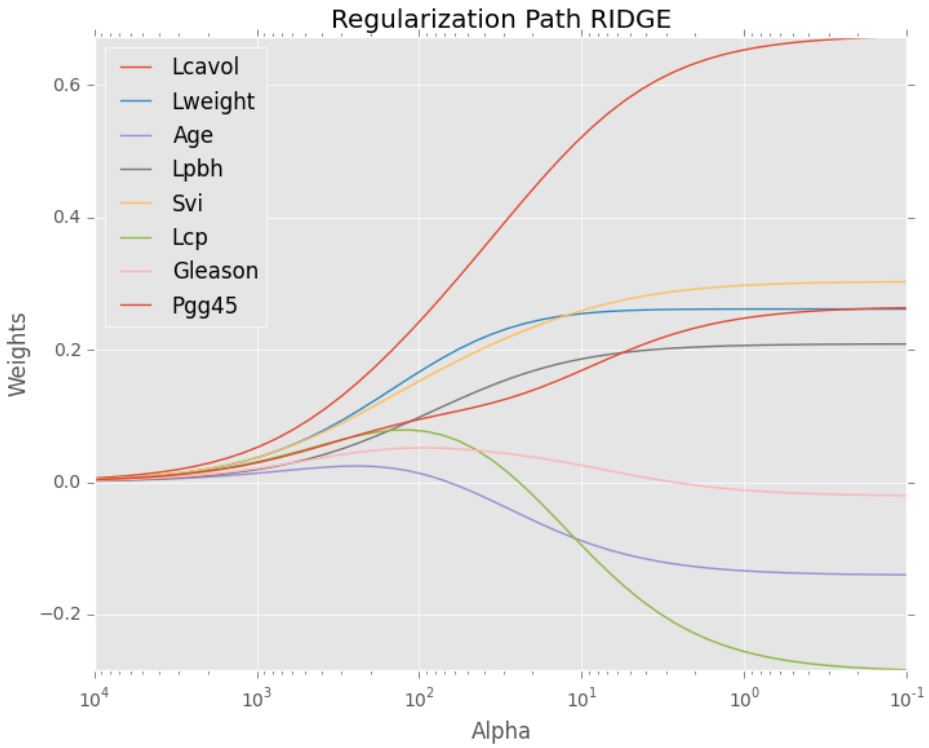


Figura 4: Coeficientes de predicción en función de parámetro de regularización, Ridge Regression .

b El comportamiento de los atributos con mayor peso sobre la respuesta siguen haciéndose notar en la regresión de Lasso, y si se observa la Figura 5, para valores de alpha más pequeños, se comienza a dejar de regularizar (más cercano a LSS), y para valores "grandes"se sobreregulariza, haciendo desaparecer los pesos de las variables del modelo. Se observan algunas diferencias con respecto a Ridge; en primer lugar la magnitud de alpha es mucho más significante para el peso de las variables, por lo que una pequeña variación de este hace que empiecen a tomar mucha importancia, o bien dejen de aparecer en el modelo. **Para alphas mayores a 1, desaparece el modelo completamente, y en este caso realmente se hacen cero los pesos, por la fórmula de cálculo lineal que tiene esta regresión**. Se puede apreciar la poca significancia de las variables age y lcp, pues estan desaparecen para alphas incluso más pequeños que 0,1. Se observa como todo el modelo comienza recién a tomar peso, muy disparado en el caso de lcavol, entre alpha 0,1y1. Las variables lcavol, lweight y svi siguen representando la mayor correlación con la respuesta, lo que confirma su relevancia.

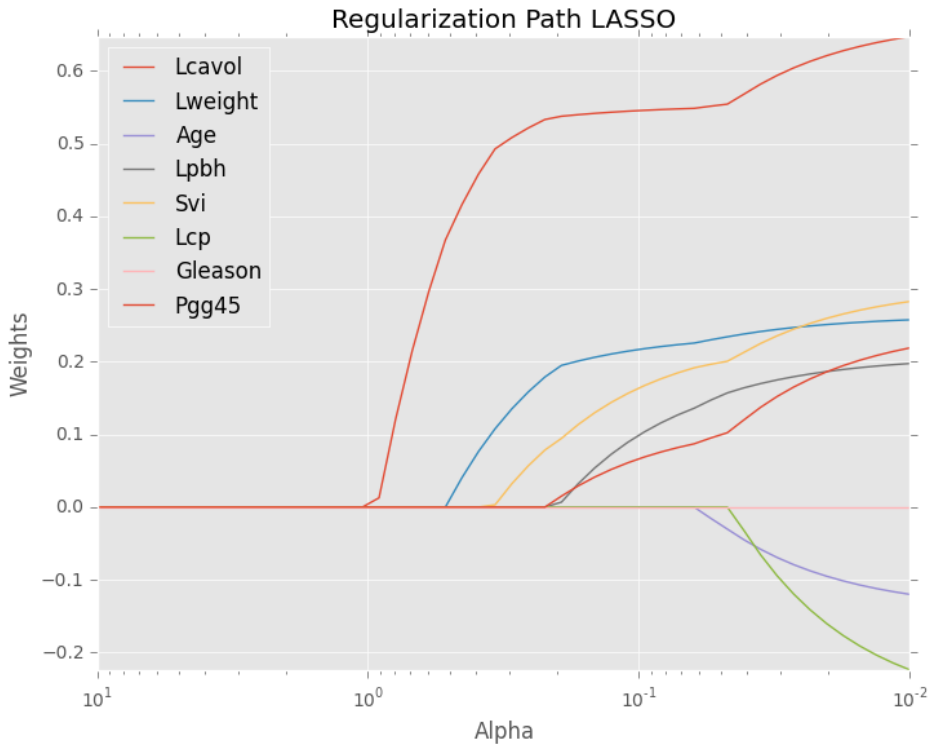


Figura 5: Coeficientes de predicción en función de parámetro de regularización, Lasso Regression.

c Con el gráfico de la Figura 6, se pueden confirmar varios de los puntos extraídos del análisis de los pesos de las variables en función del parámetro de regularización, siendo el principal el que valores grandes de este parámetro sobreregularizan y hacen alejarse a la regresión de lo que realmente dice el modelo, y en la contra parte . Se observa en primer lugar como alphas mayores a 10 aumentan considerablemente el error cuadrático medio, pues sobreregularizan, y para alphas menores a 1 el error comienza a ser constante, y de hecho, cada vez más cercano al MSE obtenido en la pregunta 1, con un valor que fluctua cerca de los 0,53 para el set de test, y entre 0,44 para el set de entrenamiento. **El menor error se observa en alpha cercano a 40 para el set de test, además el error de test es siempre menor al de entrenamiento antes de este valor, luego de este alpha el error comienza a aumentar de nuevo para test, a diferencia del set de entrenamiento, que solo disminuye su error a menor alpha.**

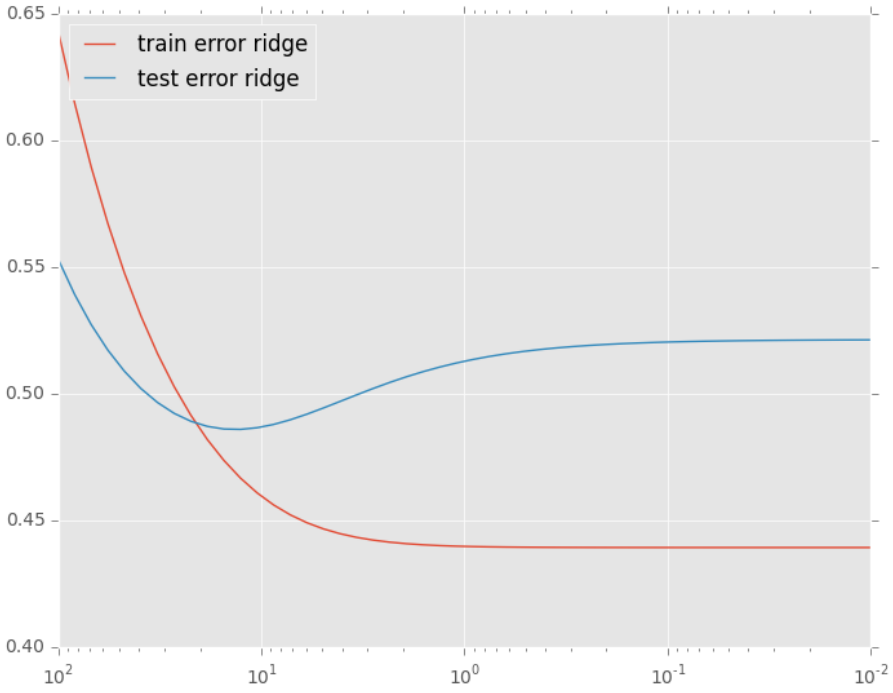


Figura 6: Error de entrenamiento y error de prueba en función de parámetros de regularización, Ridge Regression.

d Al observar la Figura 7, se confirma lo concluído para Lasso anteriormente, pues luego de $alpha = 1$, el error es elevado y constante (desaparecen los pesos), y de hecho mayor para el set de entrenamiento. También se puede apreciar la gran significancia de una variación pequeña de alpha, pues los errores bajan en picada entre $1y0{,}1$, se aprecia un hecho interesante para **alpha mayor a 0.05, pues el error del set de test recién comienza a ser más grande (como debiera serlo) en comparación al error del set de entrenamiento**. Se observa que al igual que en el ítem anterior, **el error de test comienza siendo menor para alphas que sobreregularizan, llega a un punto más bajo y luego comienza a subir hasta empeorar más que el error de train, en cambio el error de train va siempre bajando a menor alpha**. Se considera interesante haber estudiado un gráfico un rango de mangitud mas baja para alpha, pues se podría haber dicho más de la magnitud del error, que solo a priori se observa parecido al MSE calculado normalmente para cada set, a medida que alpha baja su magnitud.

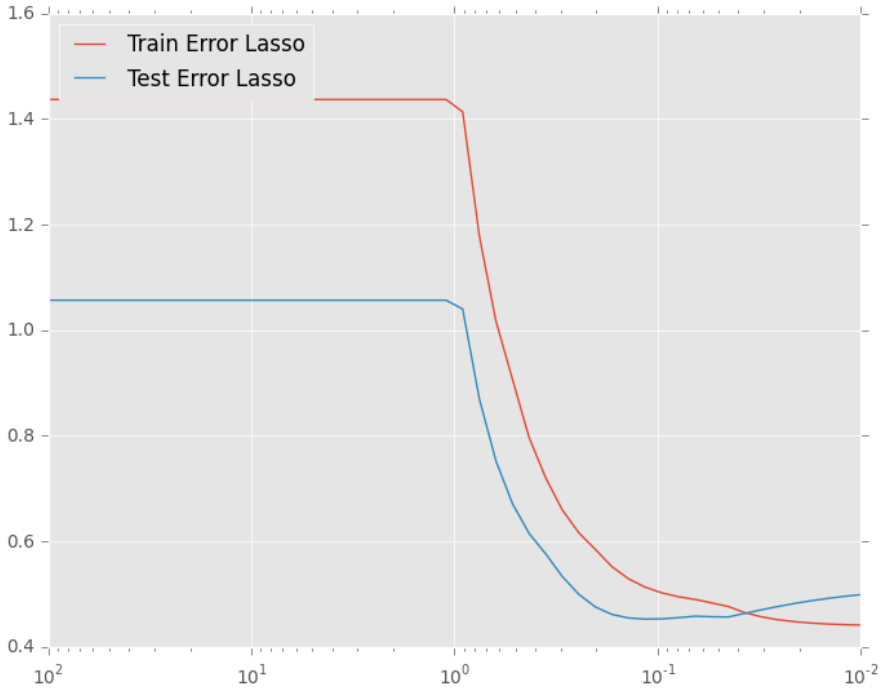


Figura 7: Error de entrenamiento y error de prueba en función de parámetros de regularización, Lasso Regression.

e Una vez implementada la validación cruzada con K=10, esta se puede aplicar a cualquiera de las dos regresiones utilizadas en este apartado, luego de aplicada a los rangos de los dos primeros ítems de este apartado, se observa en el Cuadro 2 la magnitud de parámetro que es mejor para cada regresión en el rango estudiado, pero se observa como el error cuadrático medio, en ningún caso logra superar al calculado normalmente para el set de datos, por lo que se subentiende debe aumentarse el valor de K para que logre aprender mejor del modelo, y de hecho, sería interesante haber estudiado rangos de valores de menor magnitud para alpha, pues se cree obtendrían mejores resultados en cuanto a lo que error respecta.

Regresión	**Rango**	**Best Parameter**	**MSE**
Ridge	$10^4, 10^{-1}$	2.121	0.752
Lasso	$10^1, 10^{-2}$	0.01	0.759

Cuadro 2: Tabla con mejor parámetro de regularización y MSE, utilizando regresión de Ridge y de Lasso.

5. Predicción de utilidades

a Por los resultados obtenidos de R2 al realizar una regresión lineal y su posterior regularización, se concluye que no es posible predecir las utilidades obtenidas, puesto que ninguno de los modelos utilizados se logró un resultado que funcionase como predictor. Los modelos

utilizados fueron la regresión lineal simple, y las regresiones regularizadas por el método de Lasso y el método de Ridge, además, para el método Ridge, se probó utilizar el comando fit-intercept=True, para ver si al realizar un intercepto con cero, se obtienen mejores valores de R2. Tanto para Ridge, como para Lasso, se probaron distintos valores de alfa, para encontrar aquel que permitía el mejor valor de R2, los resultados obtenidos se presentan a continuación.

Figura 8: Gráfico de R2 en función de alfa, utilizando el método de Ridge.

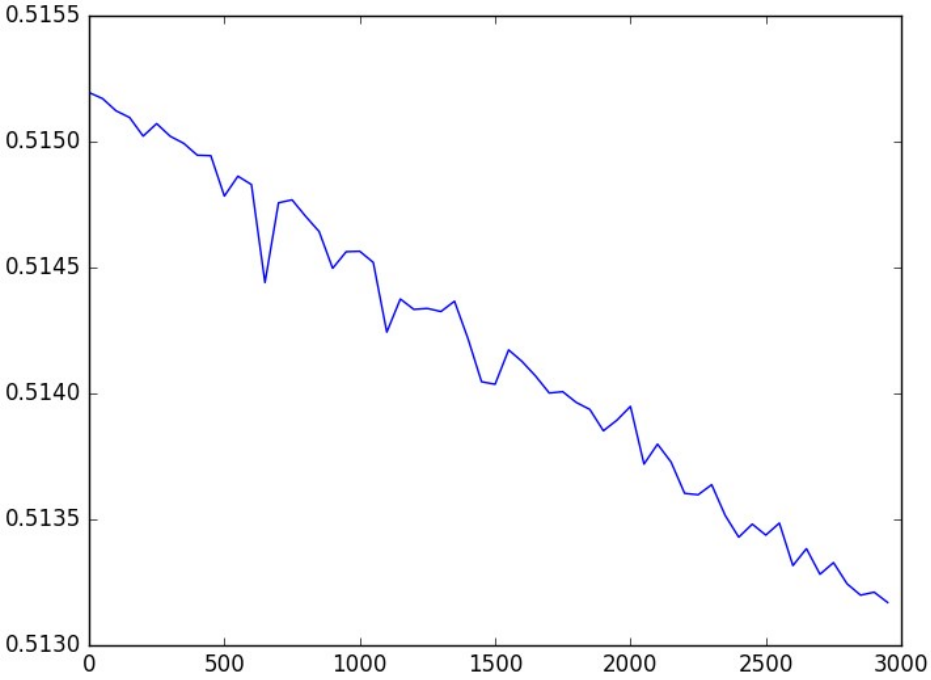


Figura 9: Gráfico de R2 en función de alfa, utilizando el método de Ridge, fiteando el intercepto.

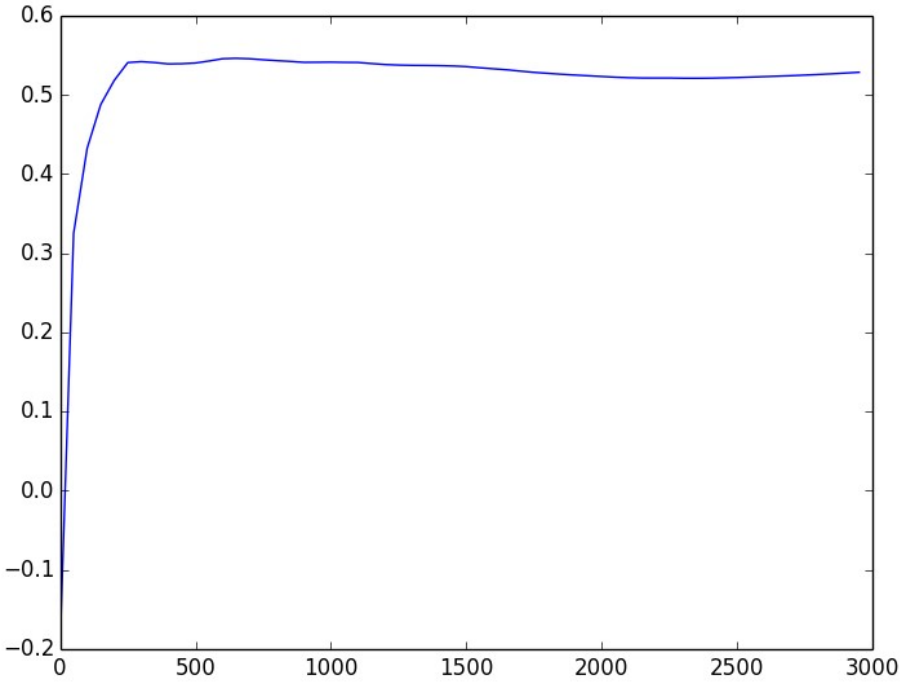


Figura 10: Gráfico de R2 en función de alfa, utilizando el método de Lasso.

Referencias

[1] Prostate Cancer Data description, Stamey, T.A., Kabalin, J.N., McNeal, J.E., Johnstone, I.M., Freiha, F., Redwine, E.A. and Yang, N. (1989) Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients, Journal of Urolog, http://www.biostat.jhsph.edu/ ririzarr/Teaching/649/prostate.html

[2] t-Student Table, http://http://cms.dm.uba.ar/academico/materias/1ercuat2015/probabilidadesyestadisticaC/tablatstudent.pdf