Análisis Inteligente de Datos: Tarea 2

Iván González, Diego Salazar

{ivan.gonzalezlo,diego.salazarb}@alumnos.usm.cl

26 de junio de 2016

1. Regresión Lineal Ordinaria

a. El código correspondiente a esta sección es `regresion.py`. La línea de código 10 elimina la columna de ids que vienen incluídos de forma original en el dataset. Tal columna no tiene nombre y no representa un atributo predictivo. La línea 11 guarda una copia de la columna *train*, la cual indica si la fila en cuestión se utilizó como parte del conjunto de pruebas (T) o del conjunto de entrenamiento (F). A partir de esa copia, se crean las listas booleanas `istrain` e `istest` (líneas 12-13), que indican si la fila en cuestión se utilizó como parte del set de entrenamiento o de pruebas. Finalmente se elimina la columna *train* del dataset (línea 14).

b. El dataset se compone de 97 filas o muestras, cada una de estas con 9 columnas. De acuerdo al sitio web[1] donde se describe el dataset, las columnas corresponden a ocho atributos predictores (columnas 1-8) y de una columna de salida **lpsa** (columna 9) que mide el nivel de antígeno prostático específico (variable de punto flotante, tipo de intervalo). A continuación, se describirán los atributos predictores:

   - **lcavol**: Logaritmo del volumen de cáncer. Variable de punto flotante, tipo de dato intervalo.
   - **lweight**: Logaritmo del peso de la próstata. Variable de punto flotante, tipo de dato intervalo.
   - **age**: Edad del paciente. Variable entera, tipo de dato intervalo.
   - **lbph**: Logaritmo de la cantidad de hiperplasia prostática benigna. Variable de punto flotante, tipo de dato intervalo.
   - **svi**: Invasión vesículo seminal. Variable entera, tipo de dato intervalo.
   - **lcp**: Logaritmo de la penetración capsular. Variable de punto flotante, tipo de dato intervalo.
   - **gleason**: Calificación de Gleason. Variable entera, tipo de dato intervalo.
   - **pgg45**: Porcentaje de Gleason 4 ó 5. Variable entera, tipo de dato intervalo.

   No existen filas con columnas o datos faltantes.

c. Al estandarizar los datos, los predictores tienen varianza igual a 1. Este proceso es útil cuando se quiere que los coeficientes de la regresión sean comparables entre sí, especialmente cuando las variables predictoras son cantidades físicas distintas o cuando los valores numéricos tienen diferente escala o magnitud (por ejemplo: **lcavol** y **age**).

d. En este punto, al utilizar la función `LinearRegression`, se setea el parámetro `fit_intercept` a `False` con el fin de que tal función no calcule el intercepto del modelo (β_0) cuando se realice el ajuste en base a los datos. Si no se agrega la columna de unos (línea 4), el intercepto del modelo pasaría por el origen (igual a cero). Agregando esa columna, se desplaza ese intercepto. Además, un intercepto igual a cero puede que no tenga relación ni sentido con el dominio del problema en cuestión. En este caso, **lpsa** mide (en [ng/mL]) el nivel de antígeno prostático específico en la sangre[2]. Los valores que arrojan los tests son mayores que cero. Un intercepto nulo o negativo no tendría significado alguno.

e. La tabla con los pesos y z_core de cada predictor, se encuentran en la tabla 1. Ahora, se utiliza la distribución de probabilidad *t-student*, con 67 - 9 = 58 grados de libertad (67 datos del dataset de entrenamiento y 9 predictores) y un α del 5 %. Esto da como resultado $t_{58} = \pm 1{,}672$.

   Aquellas variables cuyo z_score se encuentre dentro del intervalo $[-1{,}672, 1{,}672]$, no existirá suficiente evidencia que demuestre su relación con la respuesta. En este caso, las variables **pgg45**, **gleason** y **age** no presentan relación con la respuesta **lpsa** utilizando una significancia del 5 %.

[1] `http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/prostate.info.txt`
[2] `http://www.cancer.gov/types/prostate/psa-fact-sheet`

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

Cuadro 1: Peso y z_score de los predictores.

f. Se estimó el error de predicción usando *k-fold cross validation* con $k = 5, \ldots, 10$. Los resultados se pueden ver en la tabla 2. Ahí se puede ver que, aumentando el número de *folds*, el error cuadrático medio disminuye, obteniéndose su menor valor con $k = 10$. Lo anterior se explica por el hecho de que al usar un k más grande, se destina un mayor porcentaje datos para entrenamiento, por lo que se logra un mejor aprendizaje del modelo. Ahora bien, el mse obtenido al usar los datasets originales de entrenamiento (67 muestras) y de prueba (30 muestras) es igual a 0.521, valor menor que cualquiera a los obtenidos con *k-fold cross validation*. Nuevamente, se explica por la utilización de una mayor cantidad de datos de entrenamiento.

k	5	6	7	8	9	10
mse	0.957	0.957	0.895	0.880	0.819	0.757

Cuadro 2: Error cuadrático medio (mse) obtenido para cada proceso de cross-validation usando k subsets.

g. Para comprobar que la hipótesis de normalidad de los errores para cada dato de entrenamiento, se creó el qq-plot de la figura 1. Ahí se puede ver que al comparar los residuos con los percentiles de una distribución normal, estos describen una línea que se encuentra sobre la identidad. En ese sentido, se puede decir que el supuesto de normalidad de de los errores es correcto.

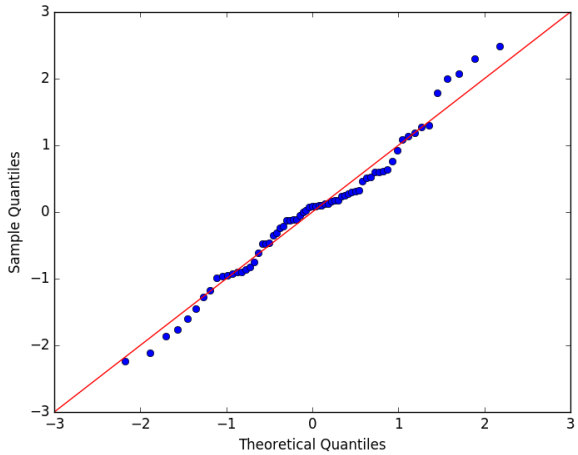

Figura 1: QQ plot de los residuos para los datos de entrenamiento.

2. Selección de atributos

a. El criterio de selección implementado en Forward stepwise selection (FSS), se basa en el cálculo del *z score* de cada atributo candidato. El que posea el *z score* con el valor absoluto más grande, será el seleccionado para ser

agregado al modelo. El modelo siempre comienza con el intercepto en él. En este caso, el orden de selección de los restantes atributos es: **lcavol**, **lweight**, **svi**, **lbph**, **pgg45**, **lcp**, **age** y finalmente **gleason**.

En la figura 2 se presenta el error cuadrático medio (mse) en función del número de atributos utilizados para construir el modelo de regresión lineal. Ahí se ve que para el conjunto de entrenamiento, el mse siempre disminuye a medida de que se agregan más predictores al modelo. Sin embargo, para el set de pruebas no es lo mismo. El mse desciende hasta el mínimo de 3 atributos (**lcavol**, **lweight**, **svi** más el **intercept**), para luego comenzar a aumentar a medida que el modelo se hace más complejo.

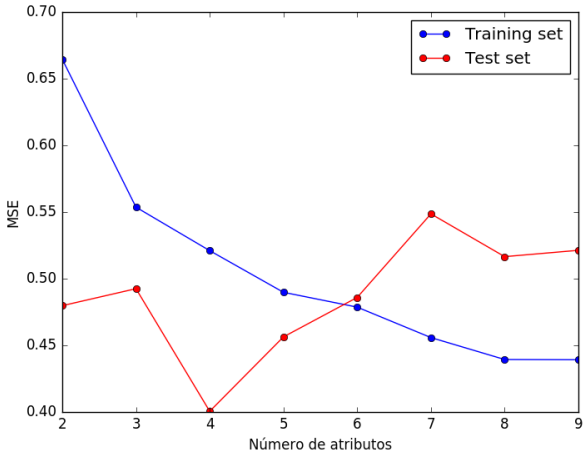

Figura 2: Error cuadrático medio de los sets de entrenamiento y pruebas, en función del número de atributos utilizados (Forward stepwise Selection).

b. Al utilizar Backward stepwise selection (bss), se comienza con un modelo considerando todos los predictores. Por cada iteración, se elimina el atributo que obtiene el menor *z score* en valor absoluto. Los resultados obtenidos con el algoritmo implementado, se muestran en la figura 3. Como se puede observar ahí, los resultados eran predecibles en el sentido de que se ocupa el mismo criterio que en fss, pero a la inversa. Para el set de entrenamiento, con más predictores se tiene un menor mse. Para el caso del set de pruebas, el mse mínimo se obtiene con 3 atributos predictores más el intercepto.

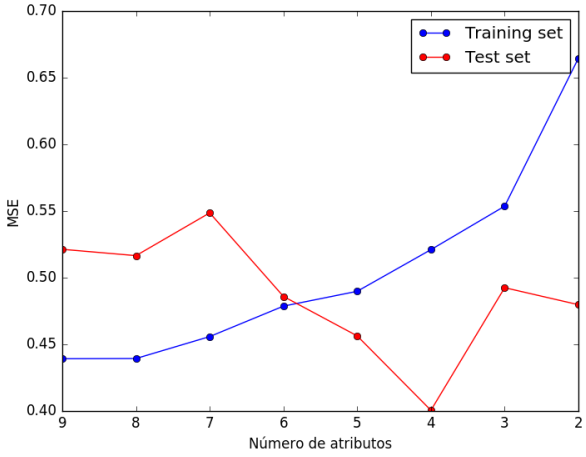

Figura 3: Error cuadrático medio de los sets de entrenamiento y pruebas, en función del número de atributos utilizados (Backward stepwise Selection).

3. Regularización

a. Se ajustó un modelo lineal del dataset, utilizando regresión de Ridge, variando el parámetro λ de regularización en el rango $[10^{-1}, 10^{4}]$. En la figura 4, se puede ver el efecto de la variación de λ en los coeficientes de los atributos predictores. Ahí se observa que λ tiene poco efecto hasta alcanza un valor entre 10^{1} y 10^{2}. A partir de un $\lambda = 10^{3}$, se puede ver un efecto significativo sobre los coeficientes, los cuales comienzan a acercarse a cero.

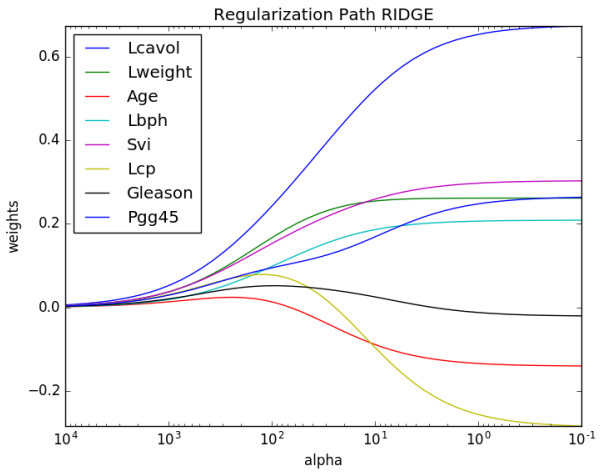

Figura 4: Coeficientes de los predictores en función del parámetro de regularización λ usado en Ridge Regression.

b. Ahora, se ajustó un modelo lineal del dataset utilizando regresión de Lasso, variando el parámetro λ de regularización en el rango $[10^{-2}, 10^{1}]$. En la figura 5, se puede ver el efecto de la variación de λ en los coeficientes de los atributos predictores. Ahí ser observa que con un λ un poco menor a $10^{-}1$, prácticamente borra del modelo a los predictores **lcp** y **age**. Con un λ entre 0.1 y 1, los atributos **lweight**, **svi**, **lbph** y **pgg45** se hacen igual a cero. Con un λ mayor a 1, todas los coeficientes de los atributos se hacen cero.

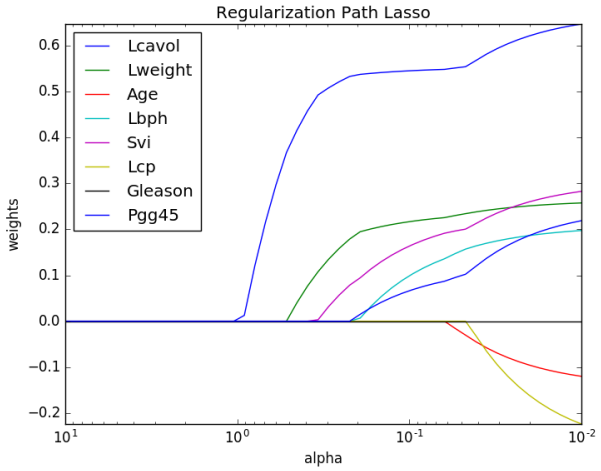

Figura 5: Coeficientes de los predictores en función del parámetro de regularización λ usado en Lasso Regression.

c. Utilizando regresión de Ridge para generar un modelo lineal, se evaluaron los errores de entrenamiento y de pruebas, en función del parámetro de regularización λ, variando el valor de este último en el rango $[10^{-2}, 10^{2}]$. El resultado se puede ver en la figura 6. Se observa que ambos mse disminuyen a medida que λ disminuye. A partir de $\lambda = 1$ los errores se estabilizan. El de test lo hace en $\approx 0{,}5$ y el de entrenamiento en $\approx 0{,}45$. El mse mínimo de test es igual 0.486.

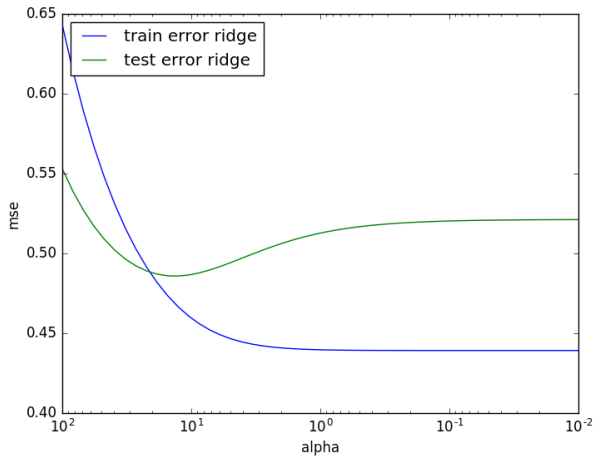

Figura 6: Errores de entrenamiento y de prueba en función del parámetro de regularización, usando Ridge Regression.

d. Utilizando regresión de Lasso para generar un modelo lineal, se evaluaron los errores de entrenamiento y de pruebas, en función del parámetro de regularización, variando el valor de este último en el rango $[0{,}5, 10^2]$. Sin embargo, con este rango no se logró apreciar ningún tipo de variación del error. Por lo tanto, tal rango se cambió $[10^{-3}, 10^0]$. El resultado se puede ver en la figura 8. Con esto, se observa que con un $\lambda = 1$, ambos mse disminuyen bruscamente, hasta estabilizarse en torno a 0,5de error. El mse mínimo de test es igual 0.453, para luego ascender y mantenerse en torno a los 0,5.

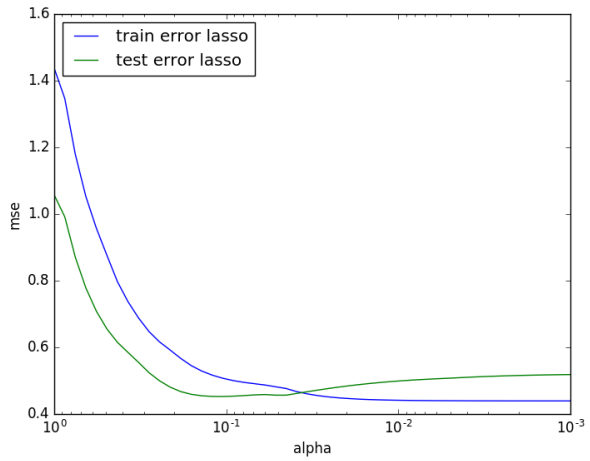

Figura 7: Errores de entrenamiento y de prueba en función del parámetro de regularización, usando Lasso Regression.

e. Usando validación cruzada, se determinó el mejor valor del parámetro de regularización para los métodos Ridge y Lasso Regression. Los resultados pueden verse en la tabla 3. Si se compara con los valores mínimos obtenidos en la sección de Regularización (Ridge: 0.486 y Lasso: 0.453), se obtienen mse mayores cuando utilizando validación cruzada en vez de utilizar un set de pruebas separado para realizar los tests. Lo anterior podría explicarse por el hecho de que al usar validación cruzada, se utiliza una menor cantidad de datos para entrenar el modelo que cuando se usa el método ordinario con el set completo de entrenamiento. Esto radica en un aprendizaje más pobre del problema por parte del modelo de regresión.

Método	rango λ	Mejor parámetro	MSE
Ridge	$[10^{-2}, 10^{2}]$	2.330	0.752
Lasso	$[10^{-3}, 10^{0}]$	0.004	0.756

Cuadro 3: Mejores parámetros de regularización para los métodos Ridge y Lasso.

4. Predicción de utilidades de películas

a. La ventaja de trabajar con formatos especiales para matrices dispersas (*sparse matrices*) y no con una librería de matrices densas es la optimización en el uso de la memoria. Si cargáramos una matriz dispersa en una estructura convencional, tendríamos una matriz densa llena de ceros ocupando el espacio de un *integer* . De este modo, dada la naturaleza del conjunto de datos a cargar, se decide trabajar con la librería para matrices *sparse* de *scipy*.

b. Se ponen a prueba los distintos modelos lineales aprendidos en clase. El método general se resumen en:

   i. Usar los datos de entrenamiento para generar una serie de regresiones lineales (ordinarias y regularizadas).

   ii. Usar los datos de validación para calcular el error cuadrático medio (MSE) en cada uno de los modelos. Se selecciona el modelo con menor error.

   iii. Poner a prueba el modelo con los datos de test, calculando el R^2 final.

   Cabe destacar que durante este proceso no se asumió la centralidad de los datos, a la vez de que no se hizo ningún tipo de pre-procesamiento. Dado este motivo, al ajustar los modelos de regresión se asumió intercepto no-nulo (*fit_ intercept=True*) , lo que aumentó considerablemente el tiempo de cómputo. Ésto último nos llevó a evitar el uso del método LASSO por problemas de convergencia y lentitud.

   Se decide comenzar con el set de datos volumen de utilidades por lugar de proyección, al contener mayor cantidad de datos. En la siguiente tabla se indica

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

Cuadro 4: Tests sobre datos por lugar de proyección

   El desempeño de estos algoritmos está muy por debajo de lo esperado, por lo que finalmente se opta cambiar el *data set* . Al repetir nuevamente el procedimiento más eficaz de la ronda anterior se obtiene:

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

Cuadro 5: Tests sobre datos totales

   Lo que es un resultado, si bien deficiente, mucho más razonable que lo obtenido.

   En razón de lo obtenido, se trata de observar bajo qué condiciones se podría cumplir que R^2 sea mayor o igual a 0,75 . Por temas de tiempo de procesamiento, se omite el paso de validación y únicamente se hacen múltiples ajustes de entrenamientos y tests para distintos α sobre el método de Ridge. A partir de los tests realizados, se genera un gráfico que permite visualizar cómo se requeriría de un alfa negativo muy alejado de 0 para llegar a

un R^2 deseado. En efecto, dada la pendiente con la que descienden los datos, es posible estimar que $R^2 = 52\,\%$ sólo ocurrirá cerca de $\alpha = -6000$, lo que claramente podría provocar *overfitting*.

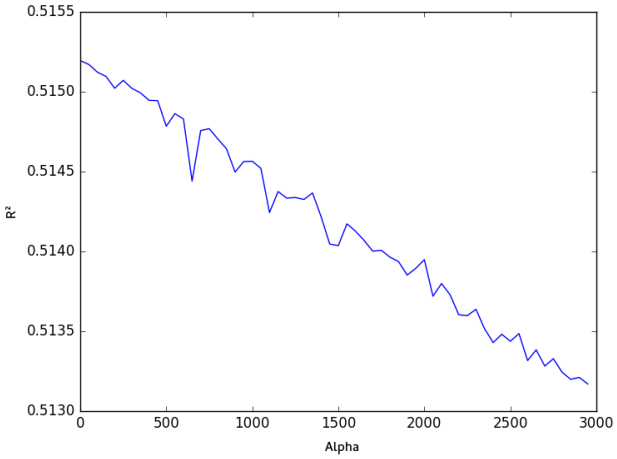

Figura 8: Relación entre parámetro α y R^2

Finalmente, y tras revisar el *paper* asociado a los datos (*Movie Reviews and Revenues: An Experiment inText Regression*) se llega a la conclusión de que lo que los investigadores utiliaron para llegar a resultados más óptimos fueron métodos de redes elásticas, lo que escapa del ámbito del curso.