# Análisis Inteligente de Datos: Tarea 2

Iván González, Diego Salazar

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

24 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 ($\beta_0$) cuando se realice el ajuste en base a los datos. De hecho, el intercepto es definido manualmente en la línea 37 del código, con un valor igual a 1.

   El proceso anterior es conveniente por varias razones. Por ejemplo, si se realiza un ajuste automático del intercepto, puede que se obtenga un valor 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 $\alpha$ 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.

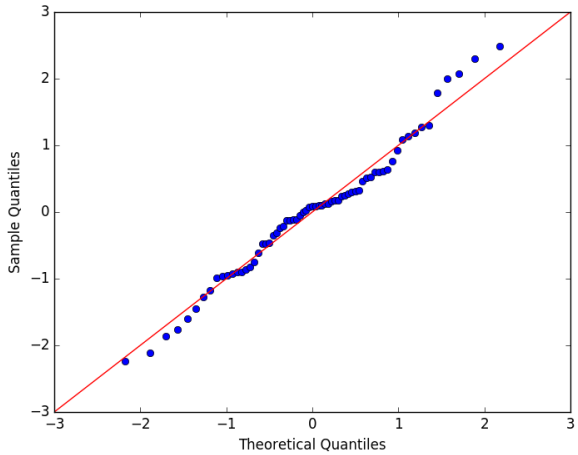


**Figura 1:** QQ plot de los residuos para los datos de entrenamiento.

## 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 4 atributos (**intercept**, **lcavol**, **lweight**, **svi**), para luego comenzar a aumentar a medida que el modelo se hace más complejo.

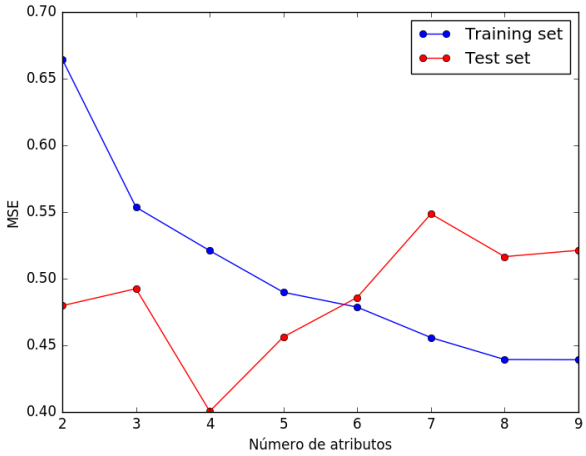


**Figura 2:** Error cuadrático medio de los sets de entrenamiento y pruebas, en función del número de atributos utilizados (Forward stepwise Selection).

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 cuatro atributos predictores.

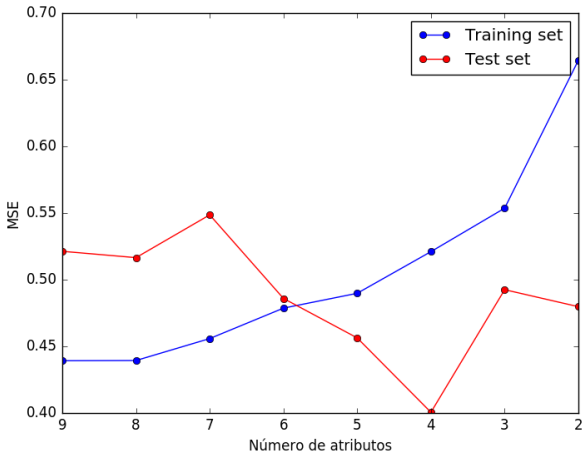


**Figura 3:** Error cuadrático medio de los sets de entrenamiento y pruebas, en función del número de atributos utilizados (Backward stepwise Selection).

## 3. Regularización

a. Se ajustó un modelo lineal del dataset, utilizando regresión de Ridge, variando el parámetro $\lambda$ de regularización en el rango $[10^4, 10^{-1}]$. En la figura 4, se puede ver el efecto de la variación de $\lambda$ en los coeficientes de los atributos predictores. Ahí se observa que $\lambda$ 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.

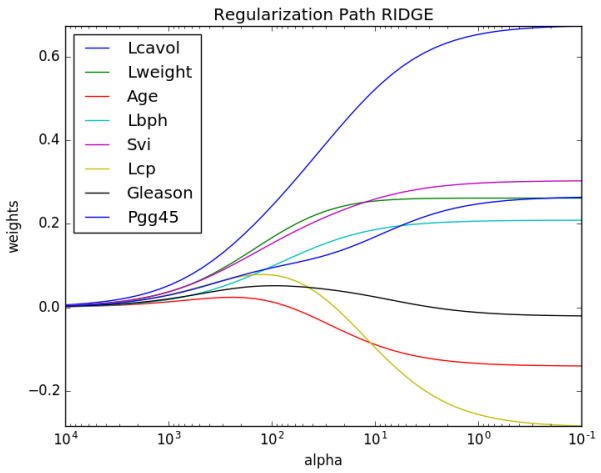


**Figura 4:** Coeficientes de los predictores en función del parámetro de regularización $\lambda$ usado en Ridge Regression.

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

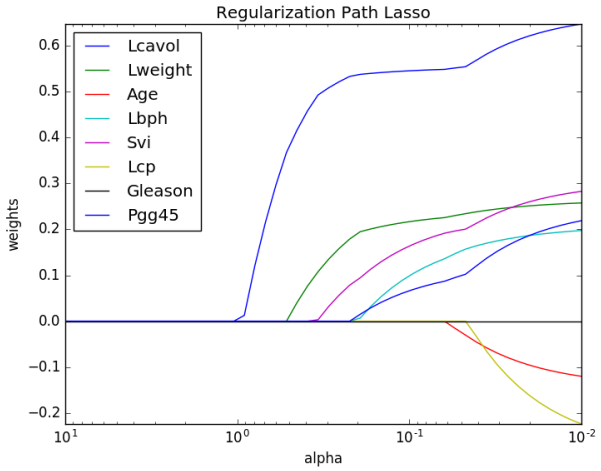


**Figura 5:** Coeficientes de los predictores en función del parámetro de regularización $\lambda$ usado en Lasso Regression.

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 $\lambda$, 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 $\lambda$ disminuye. A partir de $\lambda = 1$ los errores se estabilizan. El de test lo hace en $\approx 0{,}5$ y el de entrenamiento en $\approx 0{,}45$.

d. Utilizando regresión de Lasso para generar un modelo lineal, se evaluaron los errores de entrenamiento y de

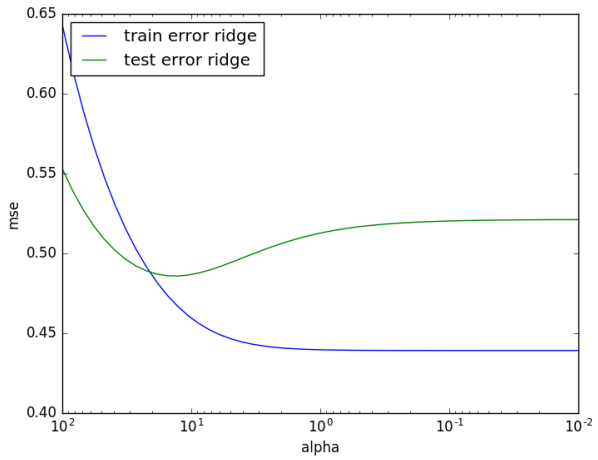


**Figura 6:** Errores de entrenamiento y de prueba en función del parámetro de regularización, usando Ridge Regression.

pruebas, en función del parámetro de regularización, variando el valor de este último en el rango $[10^1, 10^{-2}]$. El resultado se puede ver en la figura 7 (COMPLETAR).

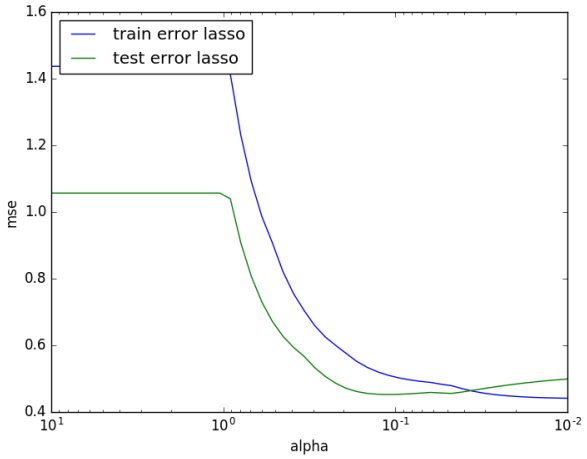


**Figura 7:** Errores de entrenamiento y de prueba en función del parámetro de regularización, usando Lasso Regression.

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.

| **Método** | **rango $\lambda$** | **Mejor parámetro** | **MSE** |
|---|---|---|---|
| Ridge | $[10^2, 10^{-2}]$ | 2.330 | 0.752 |
| Lasso | $[10^1, 10^{-2}]$ | 0.010 | 0.759 |

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

## 4. Predicción de utilidades de películas