Trabajo Práctico N°3 - Modelos de Clasificación

Aprendizaje Estadístico - FIUBA

1er Cuatrimestre - 2020

José F. González - 100063 - jfgonzalez@fi.uba.ar

Atento a: Ing. Jemina García Revisión 1.8

1. Introducción

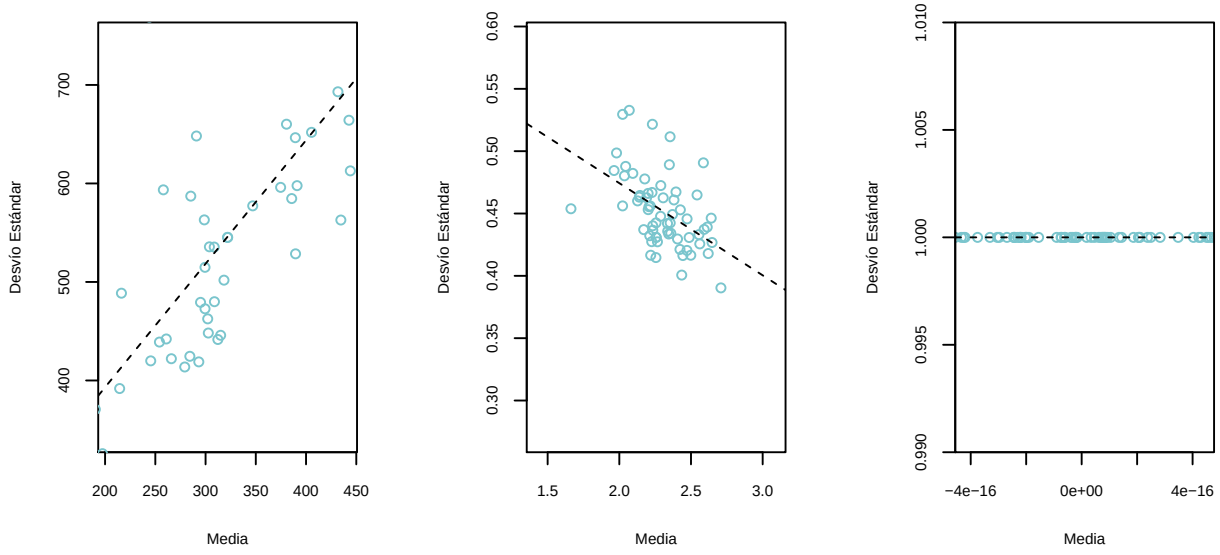
Nos interesa contruir un modelo de clasificación para detectar la presencia de tumores en muestras de tejido celular. Para ello disponemos el conjunto de datos `levels` con mediciones de los niveles de actividad de 2000 genes presentes en células de 62 tejidos distintos, donde 40 tejidos están etiquetados como tumores y 22 como normales. **Como criterio general buscamos un modelo con pocas variables explicativas que maximize la medida *accuracy* en una validación cruzada.**

2. Estandarización

Los niveles de actividad tienen un rango muy amplio lo que se traduce en una alta variabilidad, esto es evidente el panel izquierdo de la Figura 1 donde se muestran la media y el desvío estándar para los niveles de cada una de las muestras de tejido. Aplicando un logaritmo base diez a todos los datos se comprime el rango de datos y se obtiene un desvío estándar más representativo del conjunto como se ve en el panel central de la Figura 1, definimos así un nuevo conjunto `logLevels`.

Los métodos de reducción de variables que utilizaremos, para enfrentar la desproporción entre cantidad de predictoras y muestras, funcionan mejor con predictores estandarizados en media cero y desvío uno, por ejemplo para no afectar las penalizaciones de los métodos de reducción. Luego resultará conveniente definir un nuevo conjunto `stdLogLevels` en esta escala a cambio de perder un poco de interpretabilidad. En el panel derecho de la Figura 1 se ve el efecto de la nueva escala sobre la media y desvío de los genes de cada uno de los tejidos.

Figura 1: Variabilidad de la media y el desvío estandar en las 62 muestras de tejido. **Izquierda:** Desvío estandar y media para las 62 muestras de tejido del conjunto `levels`. **Centro:** Desvío estándar y media para las 62 muestras del conjunto `logLevels` luego de tomar logaritmo en base diez a los niveles de expresión de los genes. **Derecha:** Desvío estándar y media para las muestras de `stdLogLevels` luego de estandarizar los datos a media cero y desvío uno.



3. Reducción de variables

El mapa de calor de de la Figura 2 muestra los logaritmos de niveles de actividad, en los distintos genes del eje vertical, como tonos rojos para genes con mayor expresión y tonos azules para genes con menor expresión. Los genes fueron agrupados en un cluster jerárquico según los patrones de su cadena de intesidad revelando estructuras similares en las expresiones de distintos genes. Esto se interpreta como procesos de regulación entre ciertos grupos de genes, lo que sugiere efectos de multicolinealidad fuertes. Analizando la matriz de correlación de `stdLogLevels` se cuenta un 9 % de las correlaciones entre genes son mayor a 0,7 y un 2 % son correlación mayor a 0,8. Los casos de alta correlación son generalmente interpretados como genes con regulación directa entre ellos.

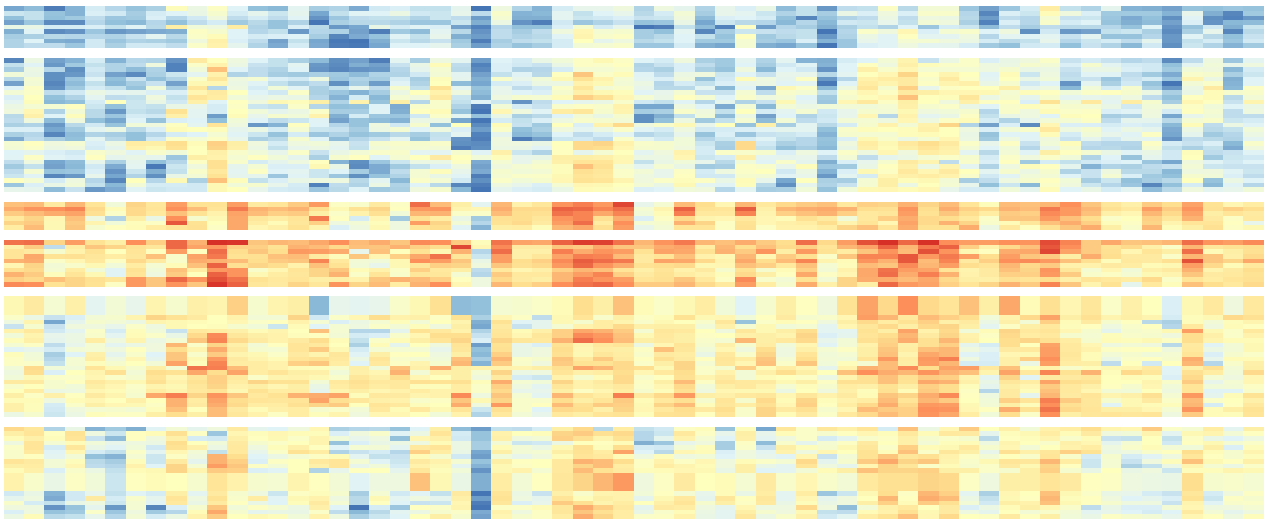
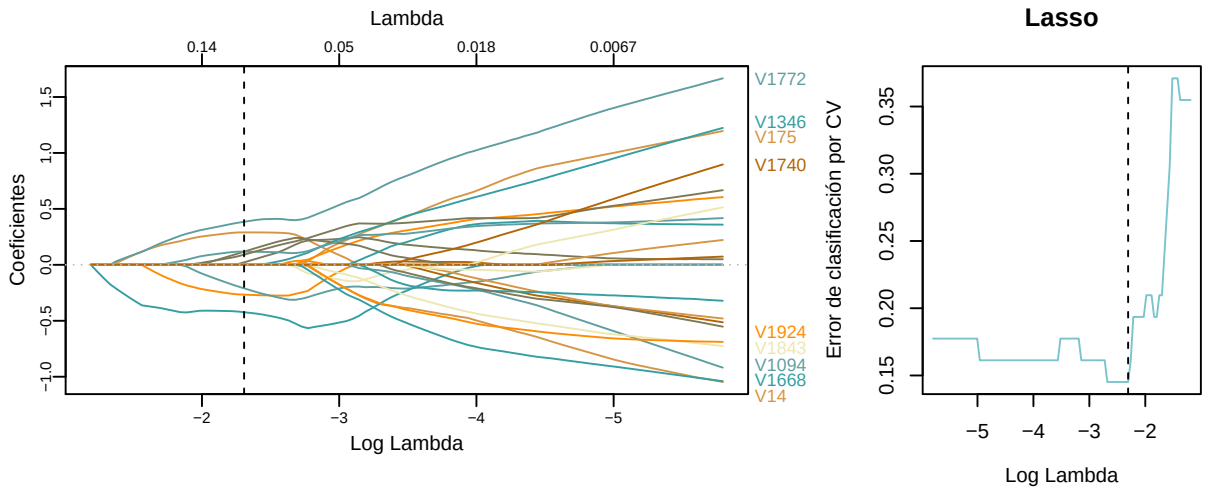


Figura 2: Actividad de los distintos genes. El eje vertical corresponde a genes y el horizontal a los tejidos. Genes con mayor nivel de actividad están codificados con tonos rojos y los de menor actividad con azules. Los genes se encuentran ordenados por un *clustering jerarquíco* según similitudes en la cadena de intesidades para facilitar la distinción de grupos correlacionados.

3.1. Selección por *Lasso*

El primer enfoque para reducir dimensión y seleccionar variables simultaneamente es penalizar por *LASSO*. En la Figura **??** se muestra el λ que minimiza el error cuadrático medio por validación cruzada para Lasso con `nfolds=10`, para el valor obtenido sobreviven 9 genes de los 2000 iniciales.



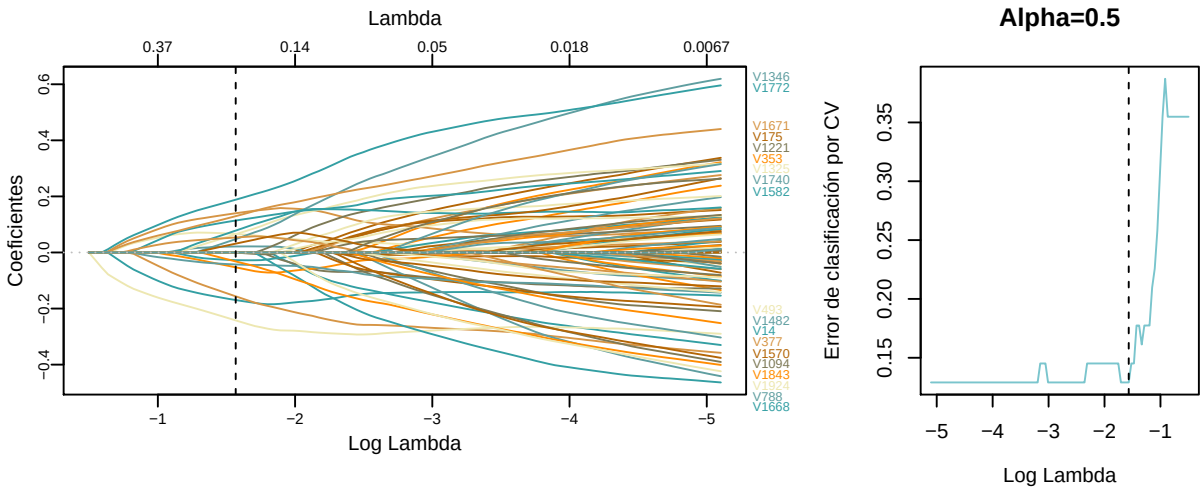


Figura 3: Comparación de *Lasso*, *Ridge* y *Elastic Net* en función de los errores de clasificación por validación cruzada de un modelo logístico construido con los coefficientes de cada método de reducción

3.2. Selección por Elastic Net

Bajo efectos de colinealidad *Lasso* elige arbitrariamente entre muchos posibles conjuntos colineales de predictoras omitiendo algunos genes que puedan estar relacionados con la salida. *Elastic Net* combina las penalizaciones de *Ridge* y *Lasso* obteniendo un método que es más inclusivo con las variables correlacionadas, por efector de la penalización *Ridge*, y da mejores resultado en este tipo de datos a costa de aumentar la cantidad de genes y tener que ajustar un nuevo parámetro α. En la Figura XX se muestran los resultados de Elastic Net para tres valores de α.

4. Selección por *Nearest Shrunken Centroids*

Nearest Shrunken Centroids es un método desarrollado en el contexto de identificar clases de genes y selección automática de genes que caracterizan la clase. Es una modificación del nearest centroids aplicando una contracción de los centroides de cada clase hacía cero, eliminando los que alcanzan el cero y clasificando sobre la nueva disposición por proximidad a los centroides que sobreviven.

En la Figura ?? se muestran los resultados del error de clasificación por validación cruzada según el parámetro de umbral Δ. Como clasificador el método no es notablemente mejor que las regresiones logísticas con reducción por *Lasso* y Elastic Net, pero puede ser un buen selector de variables. Si costruimos un modelo logístico utilizando los genes elegidos por *NSC* obtenemos ...

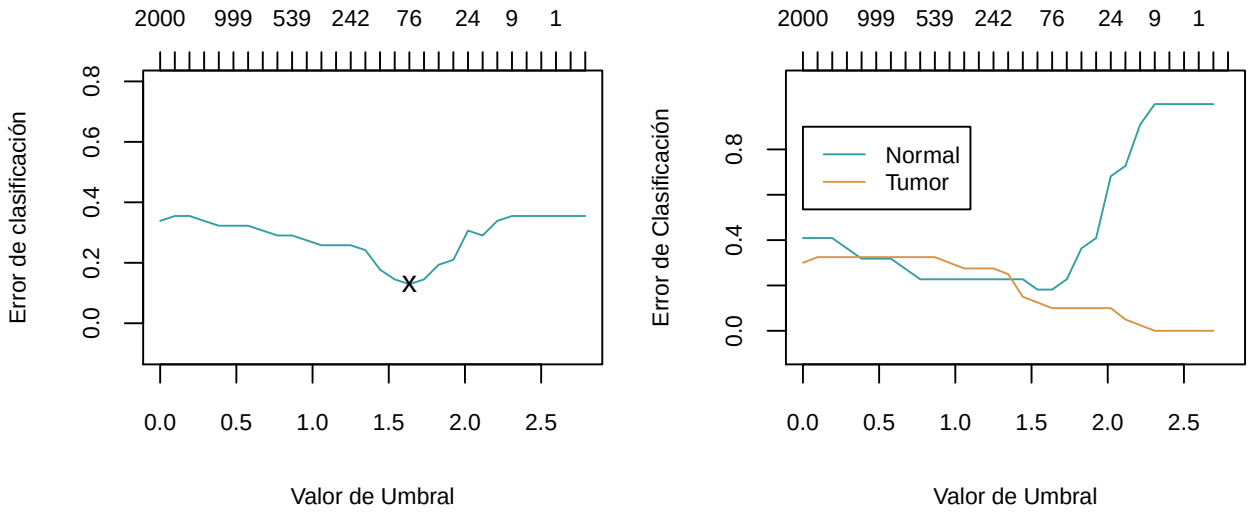


Figura 4: Comparación de *Lasso*, *Ridge* y *Elastic Net* en función de los errores de clasificación por validación cruzada de un modelo logístico contruido con los coefficientes de cada método de reducción

4.1. Resultados

5. Clasificación

5.1. SVM

```
## Error in library(e1071): there is no package called 'e1071'
## Error in svm(train.x, train.y, kernel = "linear"): could not find function "svm"
## Error in summary(svm.trained): object 'svm.trained' not found
## Error in svm(train.x, train.y, kernel = "linear", cross = 10): could not find function "svm"
## Error in summary(svm.trained.cv): object 'svm.trained.cv' not found
```

```
## Error in library(caret): there is no package called 'caret'
## Error in createFolds(labels, k, list = TRUE, returnTrain = FALSE): could not find function
"createFolds"
```

6. Bibliografía

1. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America. 1999.
2. https://stats.stackexchange.com/questions/184029/what-is-elastic-net-regularization-and-how-does-it-solve-the-drawbacks-of-ridge

7. Anexo

Gene Expression es un proceso en que la información contenida en algún gen de una célula se dispara para generar un producto, por ejemplo una proteina. *Gene Expression Level* es una medida de esta actividad de los genes computada a partir de la concentracion de *mRNA - messenger Ribonucleioc acid*, que cuando es medida puede no representar la verdadera actividad del gen pues el *mRNA* puede estar siendo regulado por otros procesos.