Regularización

Carlos Hernani Morales

20 de Diciembre, 2019

Abstract

La presencia del sobreaprendizaje en los modelos de aprendizaje automático reducen drásticamente la generalización y aplicabilidad de estos sobre datos nuevos, lo que los hace prácticamente inútiles para predecir y estimar en entornos reales donde nuevos datos aparecen constantemente. La regularización es una forma de evitar este sobreaprendizaje y así poder tener modelos que generalizen mejor. En el presente trabajo explicaré los diversos procedimientos aplicándolos sobre un banco de datos de la NBA.

Introducción

El banco de datos *NBA* presente en Kaggle contiene 835 observaciones con 20 columnas, donde nuestra variable respuesta de interés es **PTS**, los puntos marcados por un equipo en una temporada.

Table 1: Tabla de variables y estadísticos.

Variable	Percentil 0	Percentil 25	Percentil 50	Percentil 75	Percentil 100
Team	NA	NA	NA	NA	NA
SeasonEnd	1980	1989.0	1996	2005.0	2011
Playoffs	0	0.0	1	1.0	1
W	11	31.0	42	50.5	72
PTS	6901	7934.0	8312	8784.5	10371
oppPTS	6909	7934.0	8365	8768.5	10723
FG	2565	2974.0	3150	3434.5	3980
FGA	5972	6563.5	6831	7157.0	8868
X2P	1981	2510.0	2718	3296.0	3954
X2PA	4153	5269.0	5706	6753.5	7873
X3P	10	131.5	329	481.5	841
X3PA	75	413.0	942	1347.5	2284
FT	1189	1502.5	1628	1781.0	2388
FTA	1475	2008.0	2176	2352.0	3051
ORB	639	953.5	1055	1167.0	1520
DRB	2044	2346.5	2433	2516.5	2753
AST	1423	1735.0	1899	2077.5	2575
STL	455	599.0	658	729.0	1053
BLK	204	359.0	410	469.5	716
TOV	931	1192.0	1289	1395.5	1873

Hay que tener en cuenta ciertas consideraciones sobre el problema para evitar problemas a la hora de aplicar nuestro modelo de regresión lineal. En esta tarea nos ayuda enormemente el conocimiento previo del problema que nos permite eliminar ciertas variables predictoras como:

- **SeasonEnd,Playoffs,W,oppPTS** que se corresponden con el año de cierre de temporada, si jugaron o no *playoffs*, el número de partidos que ganaron y los puntos que han recibido de oponentes. Está justificado descartar estas variables puesto que no existe relación causal entre estas y los puntos marcados.
- **FG** y **FGA** son los *field goals* y *field goals attempts* que son la suma de los puntos dobles y triples, sin contar tiros libres. Si los tuviéramos en cuenta existirían problemas de colinealidad con las variables asociadas a dobles y triples.
- Además como nuestro interés es predecir los puntos que marcarán un equipo nos interesa saberlo en función de los intentos o *attempts*, ya que estos incluyen los puntos marcados al rival.

De modo que realizaremos el análisis sobre un conjunto inicial de variables que será:

Variables	Descripción
PTS	Variable respuesta
X2PA	Intentos tiro de campo 2 pts
X3PA	Intentos tiro de campo 3 pts
FTA	Intentos tiro libre
ORB	Rebotes ofensivos
DRB	Rebotes defensivos
AST	Asistencias
STL	Robos de balón
BLK	Bloqueos
TOV	Pérdida del balón

Regularización

Los modelos de autoaprendizaje, y en este caso concreto, de regresiónl lineal sufren en ocasiones de un fenómeno llamado sobreajuste o *overfitting*. Al ir añadiendo variables predictoras, i.e. aumentando la complejidad, nos acercamos al modelo óptimo pero pero puede ser que nos pasemos de complejidad y el modelo de autoaprendizaje aprenda todas las características de nuestra muestra haciendo que al comparar el resultado con una muestra de validación no sea capaz de generalizar correctamente. A esto se le conoce como el dilema sesgo/varianza ilustrado en la siguiente imagen.

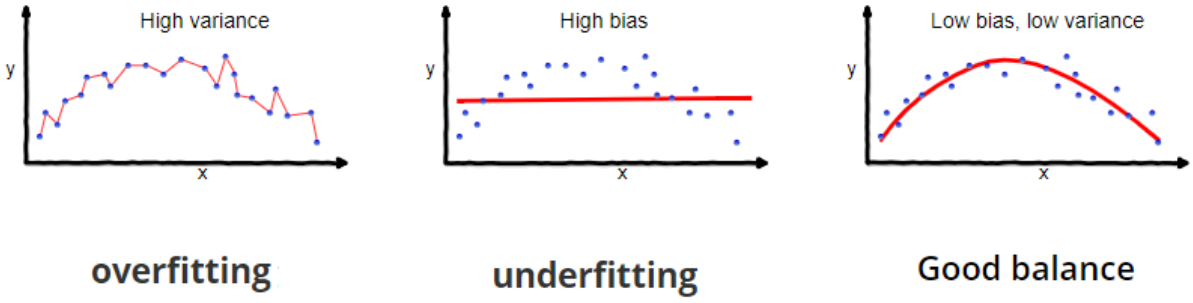


Figure 1: Dilema Sesgo/Varianza

Los modelos de regresión lineal utilizan el método de mínimos cuadrados para calcular el error entre los valores reales y estimados. El objetivo es minimizar el RSS. que es la suma de los residuos al cuadrado. Para evitar el *overfitting* podemos penalizar los modelos haciendo que los coeficientes se acerquen a cero y por tanto reduciendo la varianza de los coeficientes estimados. Esto es lo que se conoce como **regularización**.

Si tenemos un modelo de regresión $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ existen dos tipos de penalización que se pueden incluir en el RSS, es decir dos tipos de regularización:

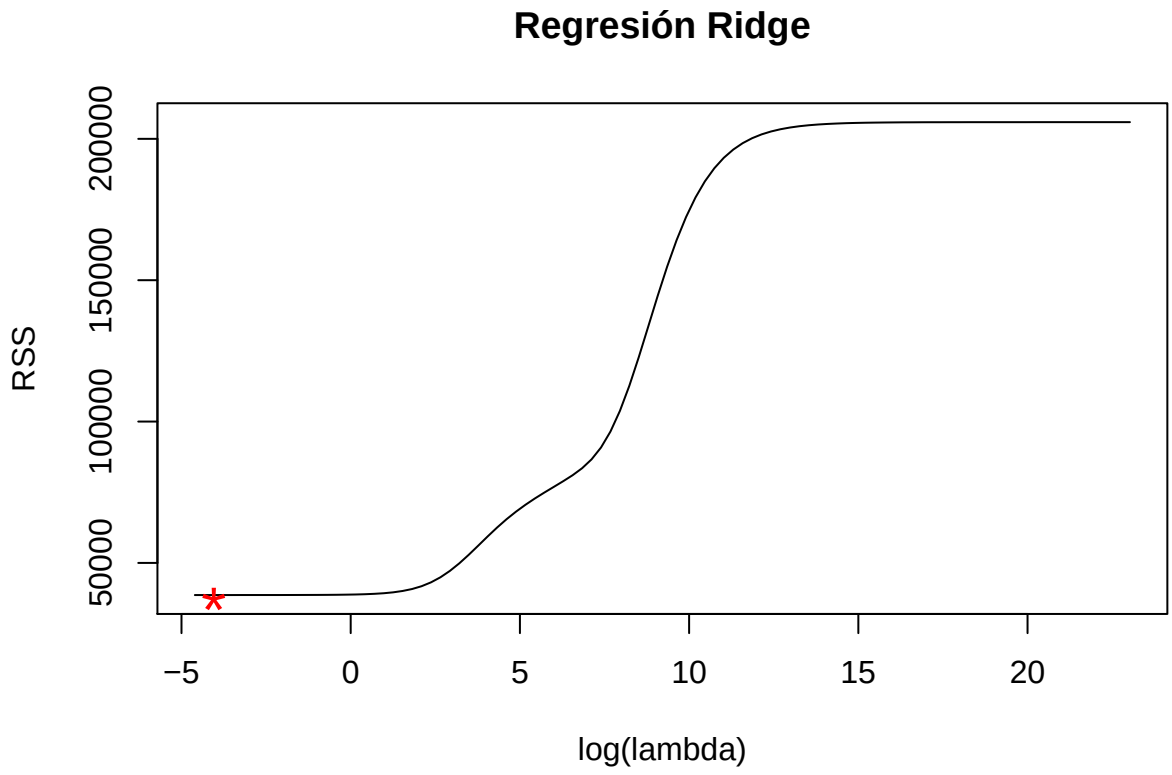
- Regresión Ridge (L2): $\lambda \sum_{j=1}^{p} \beta_j^2$

- Regresión Lasso (L1): $\lambda \sum_{j=1}^{p} |\beta_j|$

Donde λ es el factor de penalización, para encontrar el factor óptimo plantearemos un grid de modelos con diferentes valores y eligiremos el mejor.

Ambas opciones están implementadas en R en la librería *glmnet*.

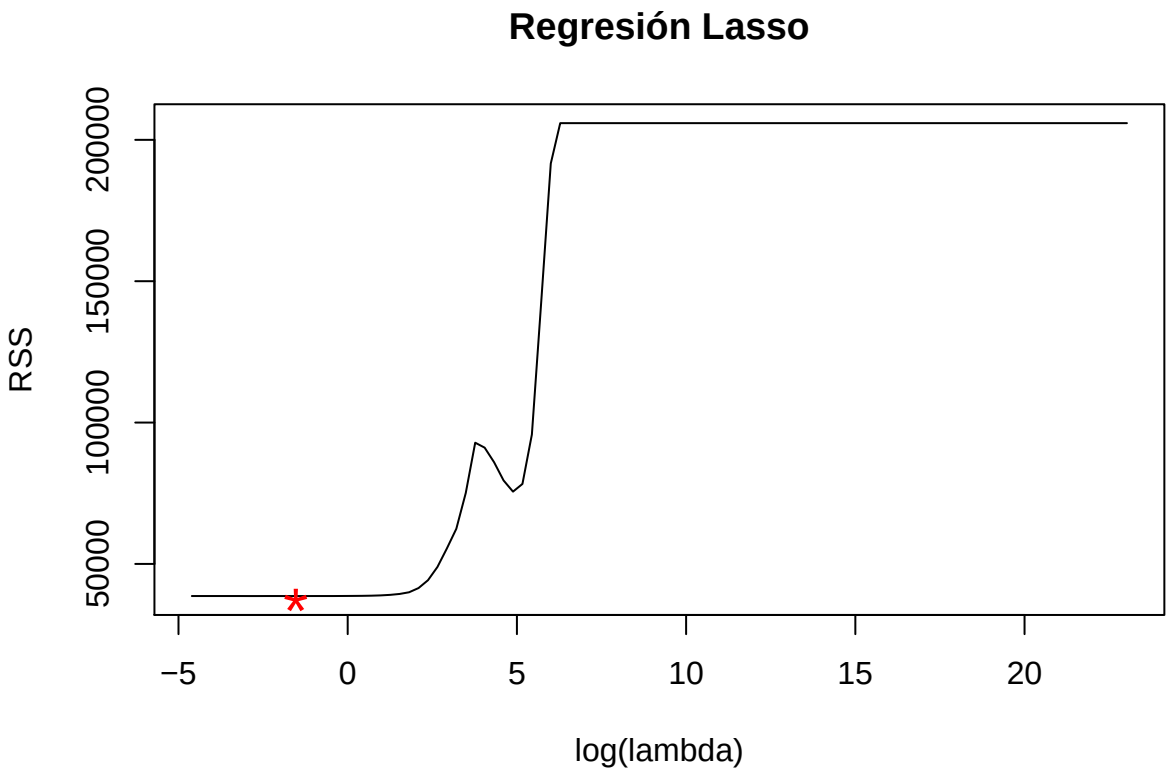
Aplicación de Regresión Ridge



El valor obtenido para λ es 0,01747528 y sus coeficientes asociados son:

	Coeficientes
(Intercept)	-2042.4692507
X2PA	1.0401433
X3PA	1.2556696
FTA	1.1273159
AST	0.8873147
ORB	-0.9506317
DRB	0.0404522
TOV	-0.0248536
STL	-0.1976625
BLK	-0.0573548

Aplicación de Regresión Lasso



El valor obtenido para λ es 0,2154435 y sus coeficientes asociados son:

	Coeficientes
(Intercept)	-2019.1277334
X2PA	1.0342094
X3PA	1.2484337
FTA	1.1242843
AST	0.8885266
ORB	-0.9402043
DRB	0.0435372
TOV	-0.0244242
STL	-0.1915474
BLK	-0.0579379

Comparación de métodos.

En general la regresión Ridge tiene un grave problema y es que mantiene todas las variables predictoras lo que puede ser un reto para modelos de muchas variables mientras que la regresión Lasso es capaz de dar valor nulo a ciertos coeficientes haciendo el modelo más explicable y sencillo de calcular.

En este caso concreto, ambos modelos no muestran resultados diferentes debido a que tenemos pocas variables y el método Lasso no ha eliminado variables.

Conclusiones

Si se nos plantea una situación donde el sobreajuste aparece, los métodos Ridge y Lasso son capaces de evitar esto penalizando el modelo. Esto se extiende al resto de modelos de autoaprendizaje como por ejemplo en redes neuronales donde además existe el método *dropout* que de forma aleatoria desconecta nodos de la red. Esto dificulta al modelo sobreaprender el ruido presenta en la muestra de entrenamiento.

Como ya he mencionado, con tan pocas variables ambos métodos han proporcionado resultados parecidos entre ellos.

Anexo: Código utilizado

Dataset: https://www.kaggle.com/amanajmera1/national-basketball-associationnba-dataset
Repositorio de GitHub: https://github.com/carhermo/TrabajoModelosLineales