Ridge Regularization as Data Augmentation

by Walter Sosa Escudero[1]

Regularized regression plays an important role in the statistical/machine learning toolkit. This pedagogical note highlights the interpretation of regularization as a process of proper data augmentation.

Consider the following loss funtion:

$$\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\hat{\beta}^2$$

where x_i is a scalar standardized explanatory variable (zero mean), $\hat{\beta}$ is an unknown parameter and $\lambda \geq 0$. Estimamtion is based on data (y_i, x_i), $i = 1\ldots, n$. This loss function corresponds to the *ridge* regularization. Its first term penalizes lack of fit and the second, departures away from $\hat{\beta} = 0$. When $\lambda = 0$ it reduces to the OLS penality function, and when $\lambda = \infty$, minimizaton requires $\hat{\beta} = 0$. Then λ controls the relative importance of the regularization term.

When $\lambda > 0$, an illustrative (an approximate) way of presenting the ridge loss function is the following

$$\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \lambda\hat{\beta}^2 = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \sum_{s=1}^{\lambda}(0 - \hat{\beta})^2,$$

or, alternatively,

$$\sum_{i=1}^{n+\lambda}(y_i - x_i\hat{\beta})^2,$$

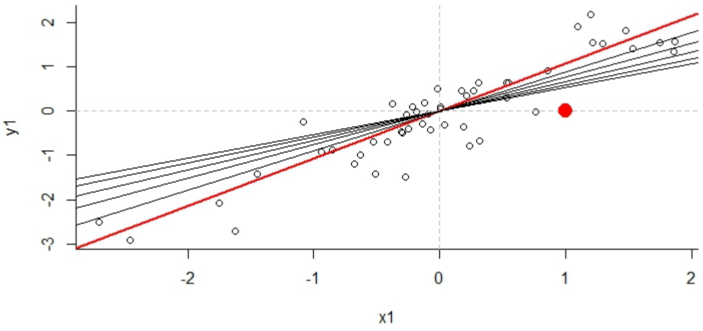
where $(y_i, x_i) = (0, 1)$ for $i = n+1, n+2, \ldots, n+\lambda$. Then the ridge estimate can be seen as arising from fitting a line to the original n points plus λ 'artificial' points that lie on the horizontal axis. More artificial points result in a more horizontal fit. Of course this is an approximation since in the original problem λ can be any real positive number and not just a natural one.

This idea of adding artificial points is compatible with the bayesian view of ridge. In such perspective the ridge loss funtion arises from adding up a normal loss function (that leads to standard OLS) and a normal prior

[1] Universidad de San Andres and CONICET. Email: wsosa at udesa.edu.ar. Very preliminary version. March 2019. Nota inspirada en la interesante charla de Pablo Zivic en el seminario Machin Lenin, en el Instituto de Calculo de la Facultad de Ciencias Exactas y Naturales de la UBA. Perdon el ingles, pero me cuesta mucho pensar estas cosas en castellano. Valen los descargos habituales de responsabilidad.

centered at $\beta = 0$. From this perspective the artificial points $(0, 1)$ tend to favor the prior model with zero slope, so resulting estimates are 'shrunk' towards it.

The following figure illustrates this interpretation:



The black points are the original data, and the red line is the OLS estimation ($\lambda = 0$). Next λ artificial points $(0, 1)$ (represented in red) are added, and the resulting fit are the black lines. As more artificial points are added, the resulting line becomes flatter, representing the idea that more points favor the zero slope model.

```
#Codigo R para el ejemplo
#Ridge as data augmented regression

x1<-rnorm(50)
x1<-(x1-mean(x1))/sqrt(var(x1))
y1<-x1+rnorm(50)*0.5

plot(x1,y1, bty="l")
abline(h=0,lty=2, col="grey")
abline(v=0, ,lty=2, col="grey")
abline(lm(y1~-1+ x1), lwd=2, col='red')
for(i in 1:5){
  yr<-c(y1,rep(0,i*10))
  xr<-c(x1,rep(1,i*10))
  abline(lm(yr~-1+xr))
}
points(1,0,col="red", pch=19, cex=2)
```