

Regularización

Es una técnica para prevenir el overfitting.

Las regresiones por sí solas tienen un problema de varianza. Frena la varianza de las regresiones. No tiene sentido hacer regresiones sin regularizar.

¿Qué hace?

- Restringe tamaño de coeficientes
- Reduce varianza (minimiza sobreajuste) pero PUEDE aumentar el sesgo de los estimadores.
- Reduce la variabilidad observada en situaciones de multicolinealidad (ridge).
- También controla el desbalance entre los números de observaciones y el número de variables.

¿Para qué?

- Reducir overfitting,
- Sirve para hacer selección de features (porque regularizamos tanto que se terminan yendo features) Cuando tengo variables con colinealidad, me dice que me quede con tales. O “dale menos importancia, no los borres”.

La **multicolinealidad** genera más overfitting (dos variables que dicen casi lo mismo) y problemas de dimensionalidad (por ejemplo, un feature temperatura celsius y otro temperatura fahrenheit). La regularización te dice cuán importante es cada variable.

Para aplicar las técnicas, es necesario **estandarizar las variables**: llevarlas a todas a la misma escala. No es la magnitud lo que le da correlación a lo que quiero predecir, sino el valor relativo que tienen.

Para estandarizar, a cada variable restamos media y dividimos por desvío estándar

Esta estandarización hace que mi regresión al calcular los pesos para cada feature no se vea sesgada por la **magnitud** de cada clase, sino por la importancia que tiene cada feature a la hora de predecir.

¿Por qué sucede el overfitting en las regresiones?

- **features irrelevantes**, que no están relacionados a Y. Si pongo features irrelevantes, las regresiones le van a dar un peso a eso. La regresión le asigna un coeficiente a todos los atributos independientemente de si es ruido o info.
- **Cuando el número de features aproxima o es igual al número de observaciones**, Puede recordar para cada feature un valor de cada instancia.
- **Features correlacionados** (multicolinealidad). Las estimaciones cuando hago mínimos cuadrados, ya tienen el supuesto de que no hay multicolinealidad.
- **Coeficientes muy grandes**. Da una mayor varianza.

Entonces queremos que los coeficientes, los betas, sean más chicos ya que tendríamos un modelo más simple. Entonces “pisamos”/**regularizamos** los betas para evitar que crezcan

Idea de regularización

- Concretamente, “restringe” o “regulariza” el tamaño de los coeficientes. Va a hacer que el valor de los coeficientes que sean demasiado grandes tengan valores normales.
- Reduce la varianza del modelo y por lo tanto minimiza el sobreajuste.
- Si el modelo es menos complejo, tiende a tener mejor varianza y generalizar mejor.
- Penaliza los valores altos de B.

Regulariza todos menos el B0

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

¿Qué pasa si todos los betas son 0? tenemos un modelo muy simple, muy bobo. Siempre contesta lo mismo sin importarle que está pasando en los datos (Modelo constante). Eso sería un modelo demasiado regularizado que quedó demasiado sesgado.

Regla del pulgar: valores altos llevan al sobreajuste.

¿Cómo mejorabamos una regresion antes? **minimizamos una funcion de perdida**, como el error residual. Ahora le agregamos un segundo término.

$$RSS = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{j=1}^M |\hat{\beta}_j|^q$$

q=2: “ridge”
q=1: “lasso”

λ=0: sin regularización
λ→∞: modelo que tiende a β₀

A la función de pérdida le sumo los Betas. Al agregar esto, además de minimizar la función va a tener que bajar los valores de los B.

Lambda me dice cuanto quiero regularizar.

Lambda = 0 : no regulariza.

Lambda → inf : el modelo tiende a B0. Penaliza mucho el valor de los coeficientes (modelo constante)

RIDGE (Los betas están elevados al cuadrado)

- Sesgo creciente y varianza decreciente a medida que crece lambda
- Siempre hay **un** valor de lambda tal que el MSE de ridge es estrictamente menor que si no lo uso. o sea siempre es mejor usar ridge
- obtenemos lambda con cross validation
- El lambda es uno en específico.

LASSO (Los betas están elevados a 1)

- Puede llevar los coeficientes a 0 anulandolos por completo.
- Es más poderoso y tiene un efecto fuerte sobre los betas. Ridge es tipo vamos viendo, por ahí lo disminuimos. Esto es por que si $0 < B < 1$, si lo elevo al cuadrado se achica el número entonces en lasso tienen más peso.
- Puede usarse como metodo de seleccion de variables. Se usa más para esto. Tengo que probar con lambdas más chicos que con Ridge.

Limitaciones

- Lasso va a seleccionar como máximo n variables.
- Lasso no hace una selección agrupada. Tiende a seleccionar una variable e ignorar a las demás.


ELASTIC NET

¿Qué es? Es una combinación lineal de Ridge y Lasso.

¿Cómo lo hace? Agrega otro parámetro alfa y cambia también la fórmula de RSS: pone un poco de penalización con ridge y un poco con lasso.

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \alpha \lambda \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \lambda \sum_{j=1}^p \beta_j^2$$

Parámetro de equilibrio



Si $\alpha = 1 \rightarrow$ Lasso.

Si $\alpha = 0 \rightarrow$ Ridge.

La parte de Lasso de la penalización genera un modelo que me elimina a los features que no sirven.

La parte de Ridge fomenta ver las cosas de forma más agrupada (deja una parte de cada una), vuelve a Lasso menos agresivo.