Regularización Lasso L1, Ridge L2 y ElasticNet

Jose Martinez Heras https://www.iartificial.net/regularizacion-lasso-l1-ridge-l2-y-elasticnet/

En muchas técnicas de aprendizaje automático, el aprendizaje consiste en encontrar los coeficientes que minimizan una función de coste. La regularización consiste en añadir una penalización a la función de coste. Esta

penalización produce modelos más simples que generalizan mejor. En este artículo vamos a hablar de las regularizaciones más usadas en machine learning: Lasso (también conocida como L1), Ridge (conocida también como L2) y ElasticNet que combina tanto Lasso como Ridge.

Índice

¿Cómo funciona la regularización?

Cuando vimos el gradiente descendiente, usamos el error cuadrático medio como función de coste *J*.

$$J = MSE$$

Cuando usamos regularización, añadimos un término que penaliza la complejidad del modelo. En el caso del MSE, tenemos:

$$J = MSE + \alpha \cdot C$$

C es la medida de complejidad del modelo. Dependiendo de cómo midamos la complejidad, tendremos distintos tipos de regularización. El hiperparámetro α

indica cómo de importante es para nosotros que el modelo sea simple en relación a cómo de importante es su rendimiento.

¿Por qué funciona la regularización?

Cuando usamos regularización minimizamos la complejidad del modelo a la vez que minimizamos la función de coste. Esto resulta en modelos más simples que tienden a generalizar mejor. Lo modelos que son excesivamente complejos tienden a sobreajustar. Es decir, a encontrar una solución que funciona muy bien para los datos de entrenamiento pero muy mal para datos nuevos. Nos interesan los modelos que además de aprender bien, también funcionen tengan un buen rendimiento con datos nuevos.

Regularización Lasso (L1)

En la regularización Lasso, también llamada L1, la complejidad *C* se mide como la media del valor absoluto de los coeficientes del modelo. Esto se puede aplicar a regresiones lineales, polinómicas, regresión logística, redes neuronales, máquinas de vectores de soporte, etc. Matemáticamente quedaría:

$$C = \frac{1}{N}\sum_{j=1}^{N}|w_i|$$

Para el caso del error cuadrático medio, este es el desarrollo completo para Lasso (L1):

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i\ -\ estimado_i)^2 + \alpha\frac{1}{N}\sum_{j=1}^{N}|w_i|$$

¿Cuándo es efectiva Lasso (L1)?

Lasso nos va a servir de ayuda cuando sospechemos que varios de los atributos de entrada (features) sean irrelevantes. Al usar Lasso, estamos fomentando que la solución sea poco densa. Es decir, favorecemos que algunos de los coeficientes acaben valiendo 0. Esto puede ser útil para descubrir cuáles de los atributos de entrada son relevantes y, en general, para obtener un modelo que generalice mejor. Lasso nos puede ayudar, en este sentido, a hacer la selección de atributos de entrada. Lasso funciona mejor cuando los atributos no están muy correlados entre ellos.

Regularización Ridge (L2)

En la regularización Ridge, también llamada L2, la complejidad *C* se mide como la media del cuadrado de los coeficientes del modelo. Al igual que ocurría en Lasso, la regularización Ridge se puede aplicar a varias técnicas de aprendizaje automático. Matemáticamente quedaría:

$$C = \frac{1}{2N}\sum_{j=1}^{N} w_i^2$$

Para el caso del error cuadrático medio, este es el desarrollo completo para Lasso (L1):

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i \ - \ estimado_i)^2 + \alpha\frac{1}{2N}\sum_{j=1}^{N} w_i^2$$

¿Cuándo es efectiva Ridge (L2)?

Ridge nos va a servir de ayuda cuando sospechemos que varios de los atributos de entrada (features) estén correlados entre ellos. Ridge hace que los coeficientes acaben siendo más pequeños. Esta disminución de los coeficientes minimiza el efecto de la correlación entre los atributos de entrada y hace que el

modelo generalice mejor. Ridge funciona mejor cuando la mayoría de los atributos son relevantes.

Regularización ElasticNet (L1 y L2)

ElasticNet combina las regularizaciones L1 y L2. Con el parámetro *r* podemos indicar que importancia relativa tienen Lasso y Ridge respectivamente. Matemáticamente:

$$C = r \cdot Lasso + (1 - r) \cdot Ridge$$

Si lo desarrollamos, para el caso del error cuadrático medio tenemos:

$$J = \frac{1}{M}\sum_{i=1}^{M}(real_i \ - \ estimado_i)^2 + r \cdot \alpha \frac{1}{N}\sum_{j=1}^{N}|w_i| + (1 - r) \cdot \alpha \frac{1}{2N}\sum_{j=1}^{N}w_i^2$$

¿Cuándo es efectiva ElasticNet?

Usaremos ElasticNet cuando tengamos un gran número de atributos. Algunos de ellos serán irrelevantes y otros estarán correlados entre ellos.