



Aprendizaje de Máquina

ITAM



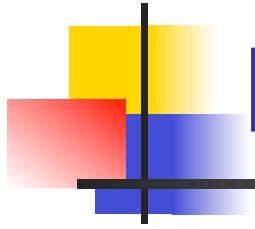
Menú

- Métodos Lineales
 - Regularización



Ajuste de Conjuntos No-lineales

- Hablamos de cómo utilizar la regresión lineal para ajustar conjuntos de datos que no son lineales mediante la adición de atributos
- Qué atributos agregar?
 - Por ahora no resolveremos esto
- Al agregar demasiados atributos podemos sobre ajustar nuestro modelo
- Ahora hablaremos de una manera en la que podemos disminuir “automáticamente” la influencia de atributos irrelevantes



Regularización

- Cuando tenemos demasiados atributos que agregan poca información
 - Atributos poco correlacionados con el valor de la función objetivo
 - Atributos muy correlacionados entre si (como en el caso de agregar x^2 x^3 x^4)
 - Cuando se usan la eq. Normales esto ocasiona que el inverso no exista
- La regularización es un técnica que nos ayuda mantener los valores de los coeficientes (w 's) bajos y a reducir el valor de los que poco aportan



Regularización: Ridge y Lasso

- Agregamos un término a nuestra función costo (error) de manera que penalice valores de w altos
- Tenemos entonces que minimizar

- Ridge

$$Costo(W) = \sum_{i=1}^N \left(y_i - w_0 - \sum_{j=1}^p x_{i,j} w_j \right)^2 + \sum_{i=1}^p \lambda w_i^2$$

- Lasso

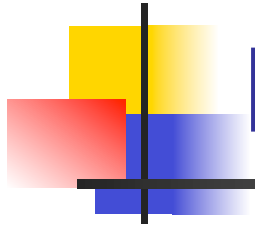
$$Costo(W) = \sum_{i=1}^N \left(y_i - w_0 - \sum_{j=1}^p x_{i,j} w_j \right)^2 + \sum_{i=1}^p \lambda |w_i|$$



Regularización

- Diferencias

- Ridge: el término es $\sum_{i=1}^p \lambda w_i^2$
 - La función a minimizar sigue siendo convexa y por tanto fácil de encontrar el óptimo global
- Lasso: el término es $\sum_{i=1}^p \lambda |w_i|$
 - No hay algoritmos muy eficientes (hay uno reciente...)
 - Obliga que algunos de los coeficientes se vuelvan cero. Esto es deseable



Regularización: Intuición

- Si Lambda es muy grande entonces el término $\sum_{i=1}^p \lambda w_i^2$ o $\sum_{i=1}^p \lambda |w_i|$ es muy grande y lo que sucederá es que las W tenderán a ser cero y el modelo ignora los datos (bajo-ajuste o underfit)
- Si Lambda es demasiado chica entonces es como si no regularizaramos (sobre-ajuste o overfit)
- Lambda controla la complejidad del modelos

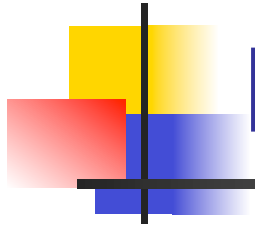


Algoritmo de Entrenamiento iterativo

(Gradient Descent regularización de Ridge)

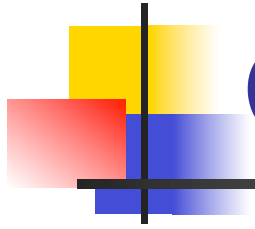
$$w_0 < -w_0 + \eta \left(y^i - \hat{V}_{ent}^i \right)$$

$$w_j < -w_j + \eta \left[\left(y^i - \hat{V}_{ent}^i \right) x_j^i \right] - \lambda w_j$$



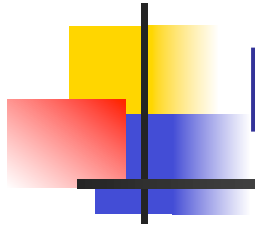
Regularización: Intuición

- El valor justo de Lambda es aquel que ayuda a distinguir entre los valores (o combinaciones) que si aportan en realidad y los que no. Una lambda que es chica en relación a los atributos de importancia y grande en relación a los irrelevantes



Comentarios

- Si los pesos se inicializan en algo grande usar regularización los hace pequeños. Si comienzan chicos no se nota el efecto de regularizar
- Si hay muchos datos con respecto al número de atributos es menos necesario regularizar
- Si el modelo no es lo suficientemente expresivo es importante regularizar



Regularización: Uso

- Ahora tenemos un parámetro más a aprender. La Lambda.
 - Esto implica que tenemos que seleccionar lambda por separado de las w 's
 - Usando bootstrapping
 - Usando validación cruzada
 - Eso lo revisaremos más adelante