Tácticas de eliminación de overfitting

Martín Gutiérrez

August 7, 2022

Introducción

Se asustaron. Se asustaron mucho la clase pasada. Apareció un término bizarro mientras revisábamos el descenso por el gradiente para las SVM.

En ese momento les dije que no lo pescaran mucho. Hoy vamos a revisar de dónde proviene. La técnica se denomina regularización.

Esa técnica es una de las formas de mitigar overfitting.

Overfitting

Recordemos qué es el overfitting primero: viene de “over” (sobre) y “fit” (ajustar).

Se refiere a una falta excesiva de error que se asocia a una no generalización del concepto a aprender. Esto significa, en el contexto del aprendizaje supervisado, que la herramienta logra ajustarse muy bien a los ejemplos. Sin embargo, no es capaz de responder correctamente ante nuevos datos.

Esto es cierto también para aprendizaje no supervisado?

Cuáles son las razones posibles detrás del overfitting?

La razón principal viene del número de características que buscan modelar el concepto a aprender.

Esto por qué? Porque a mayor número de características involucradas, más compleja se puede hacer la herramienta que busca aprender. La consecuencia es que entonces se empieza a ajustar excesivamente a los ejemplos que le son mostrados... y ya sabemos qué ocurre en esos casos.

Lo contrario ocurre también?

Efectivamente, lo contrario también existe y se llama “underfitting”.

Dicha condición ocurre cuando no hay suficientes características para poder aprender el concepto razonablemente. En términos concretos, el error es demasiado grande, pero no puede ser reducido más.

Regularización

La regularización es una técnica que persigue ajustar la importancia de cada característica empleada en el modelamiento del concepto a aprender (en el caso nuestro, de los θ).

La solución a tal problema es obvia: se le asigna un ponderador que representa (inversamente) la importancia de la característica. Recordar $J(\theta)$ que es la función de costo y sobre la que se optimiza los valores de θ, por lo que a ese nivel es que se aplica regularización.

Ejemplo - regresión lineal (I)

Para regresión lineal (multivariada), recordemos $J(\theta)$:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Como se desea asignar una importancia a los valores de θ, entonces se agrega un término que:

1. Introduce un ponderador (λ) - Parámetro de input de la regularización
2. Lo asocia a una expresión dependiente de los parámetros (θ)

Una expresión que cumple con tales propiedades sería:

$$\lambda \sum_{j=1}^{n} \theta_j^2$$

Ejemplo - regresión lineal (II)

La función $J(\theta)$ regularizada queda entonces:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Esta nueva función se emplea de la misma forma que $J(\theta)$ original (con descenso por el gradiente). Al respecto, unas preguntas:

- Diferencia entre i y j? (alguna relación con $h_\theta(x)$?)
- Qué sucede al derivar?

Ejemplo 2 - regresión logística

Para regresión logística, la función $J(\theta)$ es:

$$-\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log\ h_\theta(x^{(i)}) + (1 - y^{(i)}) log\ (1 - h_\theta(x^{(i)}))]$$

La misma expresión nos sirve en este caso:

$$\lambda \sum_{j=1}^{n} \theta_j^2$$

Queda entonces la $J(\theta)$ regularizada como:

$$-\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log\ h_\theta(x^{(i)}) + (1 - y^{(i)}) log\ (1 - h_\theta(x^{(i)}))] + \lambda \sum_{j=1}^{n} \theta_j^2$$

Próxima clase

Clasificadores/predictores adicionales (árboles de decisión, random forest). Luego seguimos con ANNs.