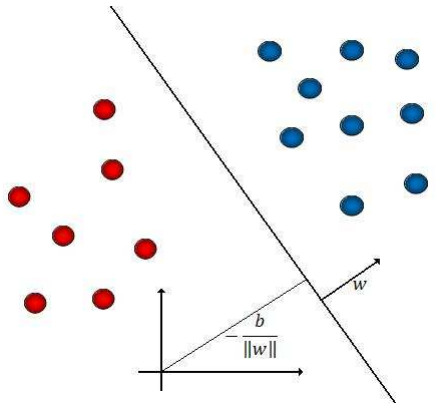
Trimestre Enero-Marzo 2018
Departamento de Cómputo Científico y Estadística
Optimización no lineal I - CO5412
Proyecto: Máquinas de Aprendizaje con Mínimos Cuadrados no Lineales

Máquinas de Aprendizaje:

Suponga que tenemos una máquina que debe 'aprender' a diferenciar entre dos clases de objetos, patrones, datos, estímulos, etc, para emitir como respuesta la clase a la que pertenece el dato suministrado. Este es uno de los problemas centrales de estudio en Máquinas de Aprendizaje, que es abordado usando Redes Neuronales, Reconocimiento de Patrones, etc.

Para darle un significado a 'una máquina que aprende' veamos un ejemplo concreto. Suponga que la máquina es una báscula electrónica cuyas funciones básicas es medir el peso y estatura de una persona (dato o patrón), queremos programar una función adicional para que la báscula indique si la persona es hombre o mujer (clases). Claramente es un ejemplo muy idealista porque necesitaría muchos más que dos atributos para diferenciar entre hombre y mujer, pero es sólo para ilustrar. Para enseñarle a la báscula a diferenciar entre hombre y mujer usando las medidas obtenidas, necesitamos una muestra de la población, es decir, dos grupos de datos asociados a hombre/mujer con medidas de peso y altura (datos de entrenamiento). Estamos partiendo desde la suposición de que los atributos que se miden en los dos grupos de datos permiten diferenciar una clase de la otra. En el ejemplo, el peso promedio del hombre es $75Kg$ y la altura promedio es de $175cm$, mientras que el peso promedio de la mujer es de $55Kg$ *coff* y altura promedio es de $160cm$ [1]. Si la muestra es representativa de la población entera y los atributos permiten diferenciar una clase de la otra, todos los datos que medirá la máquina estarán en dos cluster alrededor de estos promedios, como se muestra en el gráfico.



Al suponer que estos dos cluster pueden diferenciarse uno del otro, entonces es posible construir una línea recta (en dimensión superior un hiperplano) que separe la mayor cantidad

[1] La información estadística suministrada en el presente proyecto es ficticia, por lo que ninguna mujer resultó herida sentimentalmente en la realización del mismo. Cualquier parecido con la realidad es pura coincidencia.

de datos. Entonces, la función de clasificación para la máquina consiste en obtener el peso y altura de una persona y todo lo que debe hacer es establecer de qué lado de la línea de separación se encuentra este dato para establecer su clase. Construir esa línea en función a la muestra suministrada es el proceso de aprendizaje de la máquina, visto desde un punto de vista geométrico bastante simplificado.

Ahora vamos a plantear un modelo matemático de esta función de la máquina y del problema de aprendizaje para plantear lo anterior como un problmea matemático que podamos resolver.

Problema de Mínimos Cuadrados:

Un dato o patrón lo representamos con $x \in \mathbb{R}^m$, donde cada componente de este vector es un atributo cualitativo/cuantitativo (peso, altura, talla, índice de masa corporal, ancho de espalda, etc.). La muestra que usaremos para el entrenamiento la denotamos por $x^{(i)} \in \mathbb{R}^m$, para $i = 1, 2, \ldots, n$. Entonces, m es la cantidad de atributos y n es la cantidad de datos para el entrenamiento. Estos estos datos pertenenecen a dos clases distintas, a las que representamos con $+1$ o -1 (lados opuestos del hiperplano de separación). Las clases de los datos de entrenamiento las representamos con $d \in \mathbb{R}^n$, donde $d_i \in \{-1, +1\}$ es la clase a la que pertenece el dato $x^{(i)}$. La componente de la máquina que se encarga de clasificar un dato la representamos con una función $f : \mathbb{R}^m \mapsto \{-1, +1\}$, como esta función lo que hace es evaluar en qué lado del plano de separación está el dato x, la podemos modelar como

$$f(x) = sig(h(x))$$

donde $sig : \mathbb{R} \mapsto \{-1, +1\}$ es la función signo

$$sig(u) = \begin{cases} +1, & u \geq 0 \\ -1, & u < 0. \end{cases}$$

y $h : \mathbb{R}^m \mapsto \mathbb{R}$ es el hiperplano de separación

$$h(x) = w^T x + b,$$

con $w \in \mathbb{R}^m$ y $b \in \mathbb{R}$ parámetros desconocidos. El proceso de aprendizaje o entrenamiento consiste en determinar estos parámetros de modo que los datos de muestra esten bien clasificados en su mayoría.

Supongamos que fijamos w y b y resultan datos mal clasificados, entonces el error cometido en cada dato lo podemos medir como

$$e_i = \frac{1}{2}|f(x^{(i)}) - d_i|.$$

Note que e_i toma valores en $\{0, 1\}$, entonces el total del datos mal clasificados puede medirse como

$$ET = \sum_{i=1}^{n} e_i.$$

Como queremos que la cantidad de datos mal clasificados sea la menor posible entones para determinar w y b, debemos resolver el problema de optimización

$$\min_{(w,b)\in\mathbb{R}^m\times\mathbb{R}} ET(w,b) = \frac{1}{2}\sum_{i=1}^{n} |sig(w^T x^{(i)} + b) - d_i|$$

Claramente esta función objetivo no es suave, debido a la no suavidad de la función valor absoluto y por la discontinuidad de la función signo. Si queremos aplicar métodos de optimización suave para resolver el problema de clasificación, debemos reformular la forma de medir el error de clasificación para que sea por lo menos continuamente deferenciable. Podemos aplicar una estrategia similar al problema de mínimos cuadrados para reemplazar el valor absoluto por otra forma de medir distancias, como el error cuadrático medio

$$e_i = \frac{1}{2}(sig(w^T x^{(i)} + b) - d_i)^2.$$

La función signo debemos reemplazarla por una versión 'suave' de la función signo. Una función conocida $s : \mathbb{R} \mapsto (-1, 1)$ con caracteristicas similares a la función signo es

$$s(u) = tanh(\rho u),$$

donde $\rho > 0$ es un parámetro fijo. Mientras más grande ρ, mejor aproxima $s(u)$ a $sig(u)$, entonces la medida del error de calsificación queda

$$\epsilon_i = \frac{1}{2}(s(w^T x^{(i)} + b) - d_i)^2.$$

Note que e_i toma valores en $(0, 1)$, este error es cercano a 0 si el dato está bien clasificado y cercano a 1 si el dato está mal clasificado. Entonces el error total lo medimos con

$$ECMT = \sum_{i=1}^{n} e_i.$$

y el problema de optimización suave queda

$$\min_{(w,b)\in\mathbb{R}^m\times\mathbb{R}} ECMT(w,b) = \frac{1}{2}\sum_{i=1}^{n} \left(tanh(\rho(w^T x^{(i)} + b)) - d_i\right)^2$$

Este es un problema de ajuste por mínimos cuadrados no lineales.

Proyecto:

El objetivo de este proyecto es resolver y analizar este problema de optimización teórica y experimentalmente, sin perder de vista el origen del mismo. El proyecto en su mayoría debe ser realizado utilizando las técnicas de optimización no lineal irrestricta vistas en el curso, centradas en la resolución del problema de optimización como objetivo principal. Como objetivo secundario, debe analizar el problema de aprendizaje teórica y experimentalmente, para ello ud tiene libertad de combinar las técnicas de optimización utilizadas con conocimiento de otras áreas de las matematicas/computación como estadśticas, modelaje matemático, redes neuronales, etc.