CLASE 03: Aprendizaje Supervisado

Existen muchos enfoques de aprendizaje automático, cada uno con sus particularidades. Lo que tienen en común entre ellos es que todos postulan modelos que aprenden reglas matemáticas y estadísticas gracias a que son expuestos a un conjunto de datos muestreados con el fin de realizar diferentes acciones. Estos modeles son de variada complejidad y requieren capacidad de cómputo para ser ejecutado.

En este curso, vamos a ver particularmente dos tipos de aprendizaje, los cuales son los más populares y prácticos para la mayoría de los problemas:

- Aprendizaje Supervisado
- Aprendizaje No supervisado

Importante: para que los modelos de machine learning puedan aprender es necesario exponerlos a un conjunto de instancias (**SAMPLES**) muestreadas de una distribución de probabilidad compleja desconocida. Cada instancia (SAMPLE) esta caracterizada por un conjunto de features/variables/dimensiones. **Cada SAMPLE puede verse como un vector de dimensión d.**

$$X_i = [x_{i1}, x_{i2}, x_{i3}, \ldots, x_{id}]$$

Aprendizaje supervisado: Clasificación

El enfoque de aprendizaje supervisado se basa en disponer datos ordenados en un dataset S en pares de instancias y etiquetas (samples "x" & labels "y"). Las instancias/muestras son vectores d-dimensionales de variables aleatorias independientes e idénticamente distribuidos.

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Las etiquetas (LABELS) se suponen variables dependientes que puede tomar valores discretos (clases) o continuos a partir de distintos valores de x mediante una función f(x) llamada **FUNCIÓN OBJETIVO.** Es decir que f(x) explica la relación entre el input "x" y el output "y". Como la realidad es compleja generalmente no conocemos la verdadera f(x), por lo que trataremos de aproximarla o aprenderla.

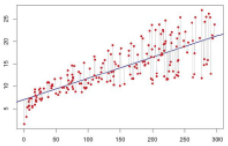
$$\mathcal{X} \in \mathbb{R}^d \qquad y \in \{-1, 1\} \qquad f(x) = y$$

Métodos de aprendizaje supervisado

Existen dos enfoques importantes en el aprendizaje supervisado:

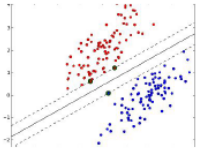
- **Clasificación**
- **Regresión**

Cuando las etiquetas toman valores categóricos hablamos de **clasificación**. Cuando las etiquetas toman valores continuos hablamos de **regresión.**

regression → 

Y is a real number

$$y \subseteq \mathbb{R}$$

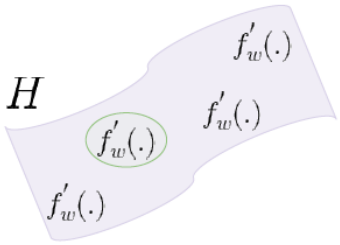
classification → 

Y is categorical

$$y \in \{-1, 1\}$$

Hipótesis

Para aproximarnos a la verdadera f(x) vamos a buscar alguna función f(.) dentro de un espacio de hipótesis que contiene muchas funciones f(.). De todas las funciones disponibles dentro del espacio de hipótesis H, vamos a intentar de encontrar alguna que explique lo mejor posible la relación entre el input y output de mi dataset. La función que vayamos a buscar estará caracterizada por parámetros (w) que puedan tomar distintos valores. Entonces existirá una combinación de parámetros que determinan una f'(.) que se aproxime a la verdad f(.) más que otras f'(.).



$$H = \left\{ f_w^1(.), f_w^2(.), \ldots, f'_w(.), \ldots \right\}$$

Suponiendo que tanto el dataset de samples-features y las etiquetas/labels están disponibles s=(x,y), vamos a aprender los parámetros que definen una función f'(x) que explique lo mejor posible la relación entre x e y. Es decir que aprenderemos una función que tomando como input las variables aleatorias "x" genere un output y' lo mas similar a las etiquetas "y" dadas por mi dataset.

Para poder medir cuan similares son las etiquetas generadas por la función aprendida f'(x) utilizaremos una función **L(y,y') de Costo o Pérdida** que tomará valores altos cuando "y" sea muy distinto a "y'". Por el contrario, cuando "y" sea muy parecido a "y'", la función de costo tomara valores bajos. Por esta razón, buscamos minimizar la función de costo. En otras palabras, **el aprendizaje supervisado puede plantearse como un problema de optimización.**

$$f(x) = y \qquad \hat{f}(x) = \hat{y} \qquad L(y, \hat{y})$$

desconocida

$$\min_w L(y, \hat{y}) = \min_w L(y, \hat{f}(x))$$

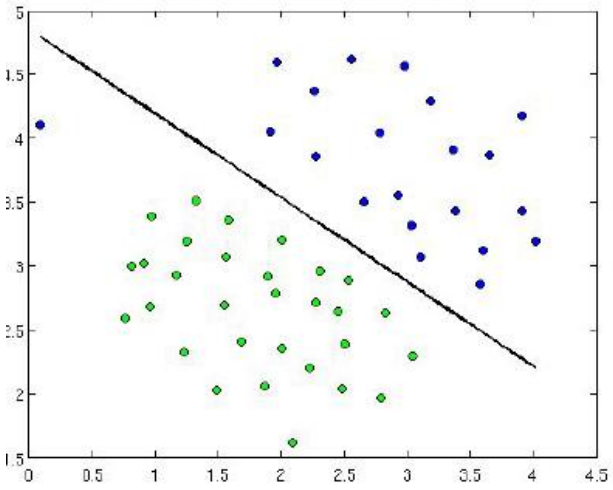
Existen muchos tipos de funciones de decisión. La familia de funciones mas conocida es la de las funciones lineales. Estas funciones son hiper-planos caracterizados por parámetros "w" que determinarán como se posiciona la frontera de decisión en el hiper-espacio de dimensión d. En la clasificación binaria, la función de decisión asignará un valor de y=1 o y=-1 según de que lado del hiper-plano se posicionen las muestras "x".

Hiper-plano separador $$f(x) = w^T x + b = 0$$

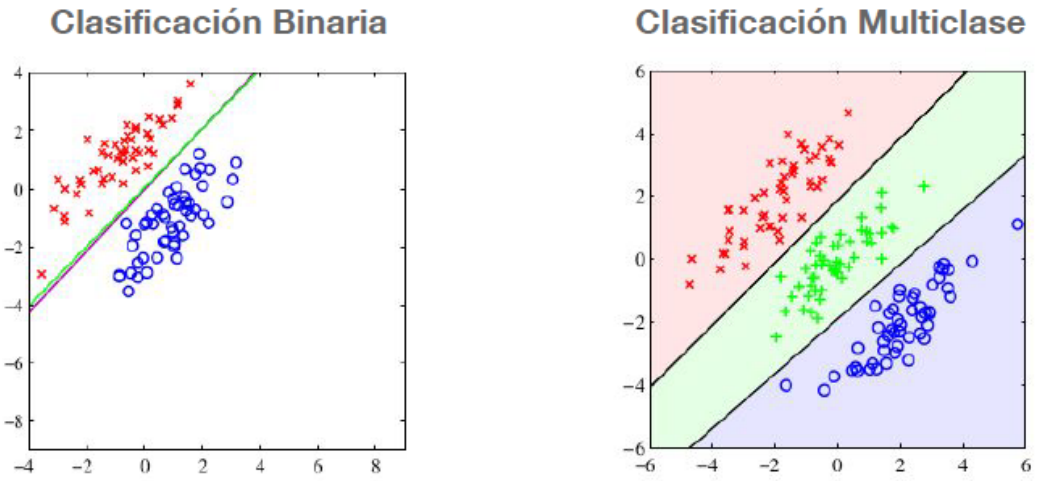
Funcion de decision $$D(x) = \text{sign}[w^T x + b]$$

Una función de decisión toma un vector input “X” (sample) con “d” features, y le asigna una de las “K”, llamada Ck.

- Cuando Ck=2: BINARIA
- Cuando Ck>2: MULTICLASE



Para un mismo set de datos etiquetados, distintos modelos pueden generar distintas funciones de decisión. Algunos modelos generarán funciones de decisión mas sencillas y otras aprenderán funciones mas complejas. La complejidad de la función a aprender dependerá de la complejidad de las muestras de entrenamiento.



En clasificación binaria aprenderemos una sola función (clasificador) mientras que en multiclase será k o k-1 funciones según el caso.