$P \rightarrow \boxed{f_A(P)} \rightarrow a$

Aprendizaje Supervisado

$f_A$ es desconocida

$f_A : X \rightarrow Y$

típicamente $X = \mathbb{R}^n$ — $x$ es un vector

$x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}$

Si $Y = \mathbb{R}$ Regresión

Si $Y = \{-1, 1\}$ Clasificación binaria

Si $Y = \{C_1, C_2, \ldots, C_K\}$ clasificación en varias clases

$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ es una muestra de $X$

$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

donde $y^{(i)} = f_A(x^{(i)}) + \hat{e}$ V.A. dist. desconocida

El Problema encontrar una función $h^* : X \rightarrow Y$

tal que $\boxed{h^* \simeq f_A}$

$h^* \in H$ hipótesis Posibles

$H = \{h_\alpha \mid h_\alpha : \mathbb{R} \rightarrow \mathbb{R}$ donde $h_\alpha(x) = \alpha x, \forall \alpha \in \mathbb{R}\}$

$h : X \times \Theta \rightarrow Y, \ \Theta = \mathbb{R}^p$

$h(x^{(i)}, \Theta) = \hat{y}^{(i)}$

Si $\theta$ es un vector de Parámetros físicos

$h_\theta : X \rightarrow Y$

$h_w(x) = \sum_{j=1}^{T} w_j x^j + w_0, \ x^j \in \mathbb{R}$

$h_w : \mathbb{R} \rightarrow \mathbb{R}$

$w = (w_0, w_1, \ldots, w_T) \in \mathbb{R}^{T+1}$

$\hat{y} = w_1 x + w_0 \quad \hat{y} = w_0 x^2 + w_1 x + b$

## Hipótesis:

1: $f: \mathcal{X} \to \mathcal{Y}$ existe y es desconocida

2: Tengo $\{x^{(1)}, \dots, x^{(m)}\} \subseteq \mathcal{X}$ una muestra de $m$ datos con distribución desconocida.

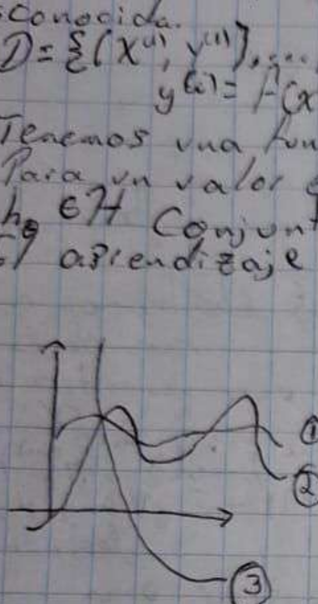
3: $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$, donde
$y^{(i)} = f(x^{(i)}) + e^{(i)}$ donde $e^{(i)}$ son VA

4: Tenemos una función "Parametrizada" $h: \mathcal{X} \times \Theta \to \mathcal{Y}$, típicamente $\Theta = \mathbb{R}^p$ (Vector de parámetros)

5: Para un valor específico de $\theta$, $h_\theta: \mathcal{X} \to \mathcal{Y}$

6: $h_\theta \in \mathcal{H}$ Conjunto de hipótesis

7: El aprendizaje supervisado consiste en encontrar $h^* \in \mathcal{H}$ tal que
$h^* \simeq f$




$L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ función de Pérdida

$$L(y, \hat{y}) = L(f(x), h^*(x))$$

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$$

$$= |y - \hat{y}| \quad y \in \mathbb{R}$$

$$= \frac{|y - \hat{y}|}{y}$$

$$L(y - \hat{y}) = \begin{cases} 0 & \text{si } y = \hat{y} \\ 1 & \text{si } y \neq \hat{y} \end{cases} = \mathbb{1}\{y \neq \hat{y}\} \quad y \in \{-1, 1\}$$

Error fuera de muestra

$$E_{out}(h^*) = \mathop{E}_{x \in \mathcal{X}}\left[L(f(x), h^*(x))\right]$$

$f \simeq h^*$ si y solo si $E_{out}(h^*) \simeq 0$

$$E_{in} = \frac{1}{m}\sum^{m} L(y^{(i)}, h(x^{(i)}))$$

Error en muestra

$f \simeq h^*$ si $\to E_{in}(h^*) \simeq 0$
$\qquad\qquad\to E_{in}(h^*) \simeq E_{out}(h^*)$

## Desigualdad de Hoeffding (PAC - Learning)

$$\Pr(|E_{out}(h^*) - E_{in}(h^*)| > \varepsilon) \leq 2e^{-2\varepsilon^2 M}$$

donde $M$ es el tamaño de la muestra

dimensión

$d_{vc}(\mathcal{H}) \approx$ # Parámetros INDEPENDIENTES

EL APRENDIZAJE ES POSIBLE SI:

$10 \cdot d_{vc}(\mathcal{H}) << M \leftrightarrow E_{in}(h^*) \approx E_{out}(h^*)$

---

$$\begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(m)T} \end{bmatrix} \begin{matrix} \to \\ \to \\ \\ \to \end{matrix} \begin{array}{cccc} a_1 & a_2 & \cdots & a_n \\ \end{array} \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \quad \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Instancias $\qquad X \ (m, n) \qquad Y \ (m, 1)$

$x^{(i)} \in \mathbb{R}^n$
$y^{(i)} \in \mathbb{R}$

$\to D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$
Conjunto de aprendizaje

$h_\theta(x) = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b = \sum_{i=1}^{n} w_i x_i + b = w^T x + b$
$= x^T w + b$

Si $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$
$w = (w_1, \ldots, w_n) \in \mathbb{R}^n$

$\theta = (w_1, \ldots, w_n, b) \in \mathbb{R}^{n+1}$

Aprendizaje:
$E_{in}(h^*) \approx E_{out}(h^*)$
$E_{in}(h^*) \approx 0$

$h^* = \arg\min E_{in}(h) \leftrightarrow \theta^* = w^*, b^* = \arg\min_{\substack{w \in \mathbb{R}^n \\ b \in \mathbb{R}}} E_{in}(h_{w,b})$

$w^*, b^* = \arg\min_{\substack{w \in \mathbb{R}^n \\ b \in \mathbb{R}^n}} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (y^{(i)} - h_{w,b}(x^{(i)}))^2$ MSE
Mean Square Error

$w^*, b^* = \arg\min_{\substack{w \in \mathbb{R}^n \\ b \in \mathbb{R}}} \frac{1}{M} \sum_{i=1}^{m} \frac{1}{2} (y^{(i)} - \underbrace{w^T x^{(i)} - b}_{= x^{(i)T} w - b})^2$

$= \arg\min_{\substack{w \in \mathbb{R}^n \\ b \in \mathbb{R}}} \frac{1}{2M} [(y^{(1)} - x^{(1)T} w - b) \ldots (y^{(m)} - x^{(m)T} w - b)] \begin{bmatrix} y^{(1)} - x^{(1)T} w - b \\ y^{(2)} - x^{(2)T} - b \\ \vdots \\ y^{(m)} - x^{(m)T} w - b \end{bmatrix}$

$= \arg\min_{w \in \mathbb{R}^n} \frac{1}{2M} [Y - [X, \vec{1}]\theta]^T [Y - [X, \vec{1}]\theta]$

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} - \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(m)T} \end{bmatrix} w - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} b$$

Para encontrar un mínimo de una función: se deriva y se iguala a cero

En este caso como tenemos dos variables tenemos que derivar parcialmente para encontrar el gradiente.

Ejemplo: $f(x_1, \dots, x_n)$
$f(y)$ $\quad x \in \mathbb{R}^n,\ f: \mathbb{R}^n \to \mathbb{R}$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

Regresamos al Problema...

$$[Y - X_e\theta]^T[Y - X_e\theta]$$

$$\frac{\partial}{\partial M}[Y - X_e\theta]_{m,1}$$

$$\partial(y-Kx)^{-2} = -2(y-Kx)K$$

$$\frac{1}{M}X_e^T[Y - X_e\theta] = \vec{0}$$

$$X_e^T y - X_e^T X_e\theta = \vec{0}$$

$$X_e^T X_e\theta = X_e^T Y$$

$$\boxed{\theta = \underbrace{[X_e^T X_e]^{-1} X_e^T}_{\mathrm{Pinv}(X_e)} Y}$$ ← Este es el metodo mínimo de cuadrados