Machine Learning para Inteligencia Artificial

Aprendizaje estadístico

Universidad ORT Uruguay

2 de Abril, 2025

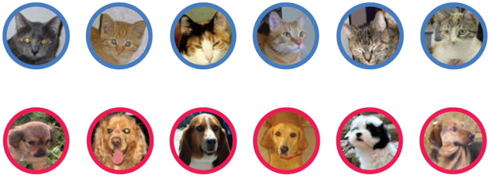
Ingredientes: Atributos y Etiquetas

Problema: Clasificación binaria

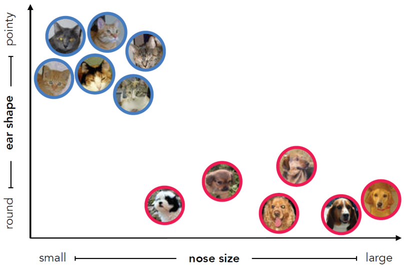
$$\mathcal{Y} = \text{ Espacio de etiquetas } = \{\text{Cat}, \text{Dog}\}, \quad y = \begin{cases} \text{Cat} \\ \text{Dog} \end{cases}$$

$$\mathcal{X} = \text{ Espacio de atributos } \subset \mathbb{R}^D$$

Espacio de etiquetas $\mathcal{Y}$



Espacio de atributos $\mathcal{X}$



Ingredientes: Datos

Datos: conjunto de instancias (muestra) de perros y gatos

$$N = 12 \text{ observaciones } \textcolor{red}{\text{etiquetadas}} \text{ de la forma } (\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$$

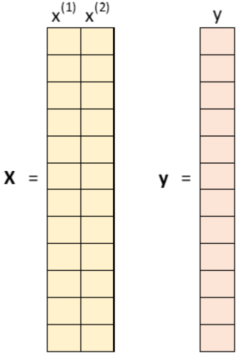
Los atributos de cada instancia son un vector

$$\boldsymbol{x} = \left(x^{(1)}, x^{(2)}\right) = (\text{nose size}, \text{ear shape}) \in \mathbb{R}^D, \;\; D = 2$$

En este caso están representados en forma tabular

$$\boldsymbol{X} = \;\; \text{matriz de diseño} \;\; \in \mathbb{R}^{(N,D)}$$

$$\boldsymbol{y} = \;\; \text{vector de etiquetas} \;\; \in \mathcal{Y}^N$$

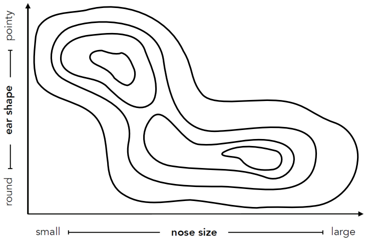


Ingredientes: Distribución

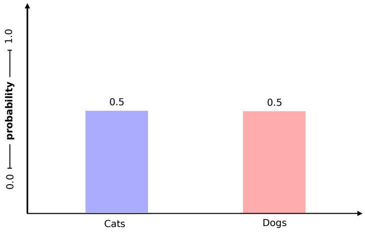
Datos: $T = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ muestra **i.i.d.** de distribución desconocida $\mathcal{D}$ en $\mathcal{X} \times \mathcal{Y}$

- $\mathcal{D}$ representa la distribución conjunta del par $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$

Atributos $p(x)$



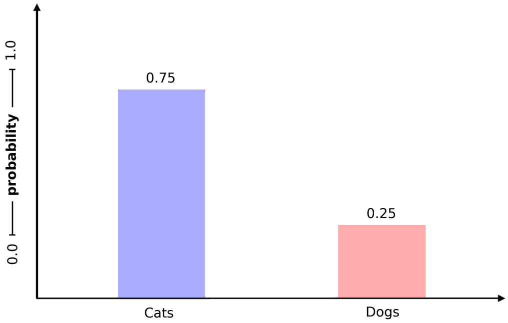
Etiquetas $p(y)$



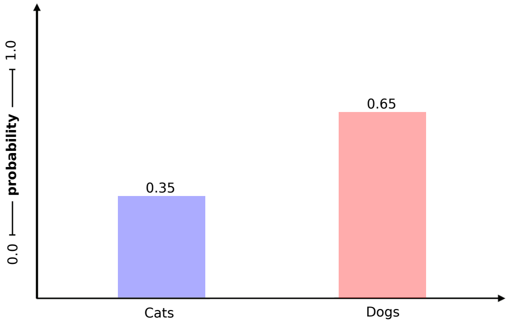
Ingredientes: Distribución

La relación estocástica entre $\boldsymbol{x}$ e y viene dada por las condicionales

Condicional $p(y \mid \boldsymbol{x} = v_1)$

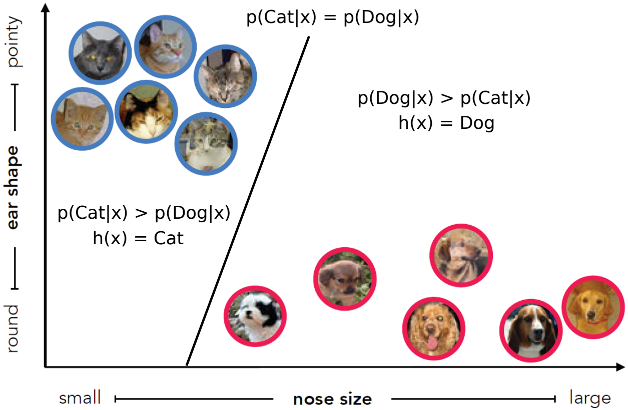


Condicional $p(y \mid \boldsymbol{x} = v_2)$



Ingredientes: Hipótesis o Modelo

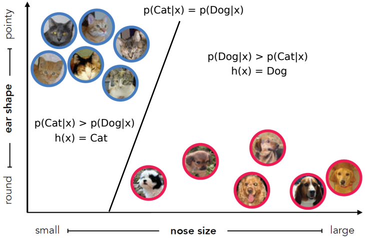
Aprender: inferir una hipótesis/modelo h a partir de T en $\mathcal{H}$ espacio de hipótesis.



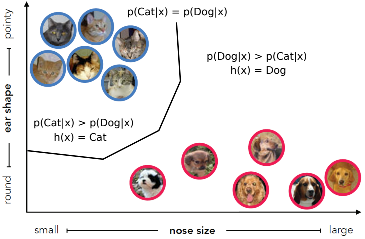
Ingredientes: Sesgo Inductivo

- $\mathcal{H}$ se llama sesgo inductivo. Por ejemplo:

$\mathcal{H} =$ **Rectas**



$\mathcal{H} =$ **Poligonales**



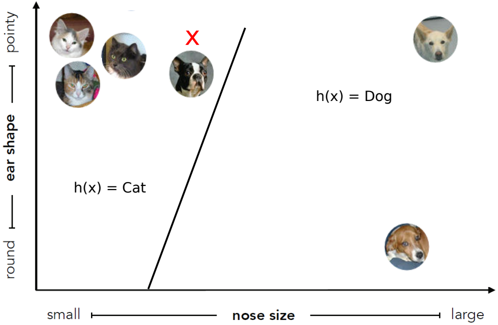
- Una hipótesis es una función $h : \mathcal{X} \to \mathcal{Y}$ perteneciente a $\mathcal{H}$
- Suele ser de la forma $h(\boldsymbol{x}) = \begin{cases} \text{Cat} & \text{si } p(Cat \mid \boldsymbol{x}) \geq 1/2 \\ \text{Dog} & \text{si } p(Cat \mid \boldsymbol{x}) < 1/2 \end{cases}$

Ingredientes: Predicción

Error de generalización

Un modelo debe desempeñarse bien en datos no vistos (validación).

Datos nuevos



Ingredientes: Función de Pérdida (Loss) y Costo

- **Pérdida de una predicción** $\widehat{y} = h(x)$ con respecto a la *verdad* y:

$$\text{Loss}(\text{Predicción}, \text{Verdad}) = L(\widehat{y}, y)$$

- Ejemplo (0-1 loss): $L(\widehat{y}, y) = \text{Loss}(\widehat{y}, y) = \mathbb{1}_{\{\widehat{y} \neq y\}} = \begin{cases} 1 & \text{si } \widehat{y} \neq y \\ 0 & \text{si } \widehat{y} = y \end{cases}$

- **Costo (o riesgo, o error) verdadero de una hipótesis** h respecto a $\mathcal{D}$:

$$J_{\mathcal{D}}(h) = \text{Cost}_{\mathcal{D}}(h) = \underset{(\boldsymbol{x}, y) \sim \mathcal{D}}{\boldsymbol{E}} \left[\text{Loss}\left(h(\boldsymbol{x}), y\right) \right]$$

- Ejemplo (0-1 loss): $J_{\mathcal{D}}(h)$ es igual a $\text{Prob}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[h(\boldsymbol{x}) \neq y \right]$

Ingredientes: Minimización del costo verdadero

- **Objetivo**: Idealmente construir la hipótesis que minimiza el costo verdadero

$$\boxed{h_{\text{opt}} = \arg\min_{h \in \mathcal{H}} J_{\mathcal{D}}(h)}$$

- Si conociéramos $\mathcal{D}$ estaríamos frente a un problema de optimización clásico.

Ejemplo: si conociéramos $\mathcal{D}$

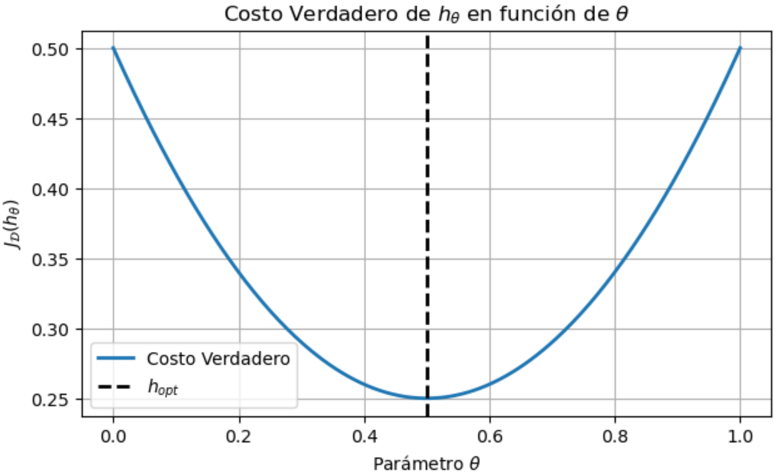
- $\mathcal{X} = [0,1]$, $\mathcal{Y} = \{\text{rojo}, \text{azul}\}$, y $\mathcal{D}$ la distribución $\begin{cases} p(x) \text{ uniforme en } [0,1] \\ p(y = \text{azul} \mid x) = x \end{cases}$

- $L(\widehat{y}, y) = \mathbb{1}_{\{\widehat{y} \neq y\}}$ la 0 - 1 loss

- $\mathcal{H} = \{h_\theta : \theta \in [0,1]\}$ donde $h_\theta(x) = \begin{cases} \text{azul} & \text{si } x > \theta \\ \text{rojo} & \text{si } x \leq \theta \end{cases}$

- Se puede ver que $J_{\mathcal{D}}(h_\theta) = \frac{1}{2}\theta^2 + \frac{1}{2}(1-\theta)^2 = \theta^2 - \theta + \frac{1}{2}$

Ejemplo: si conociéramos $\mathcal{D}$



Ingredientes: Minimización del costo empírico

Como NO conocemos $\mathcal{D}$ lo hacemos con la muestra disponible T

- **Costo (o riesgo, o error) empírico de una hipótesis** h respecto a T

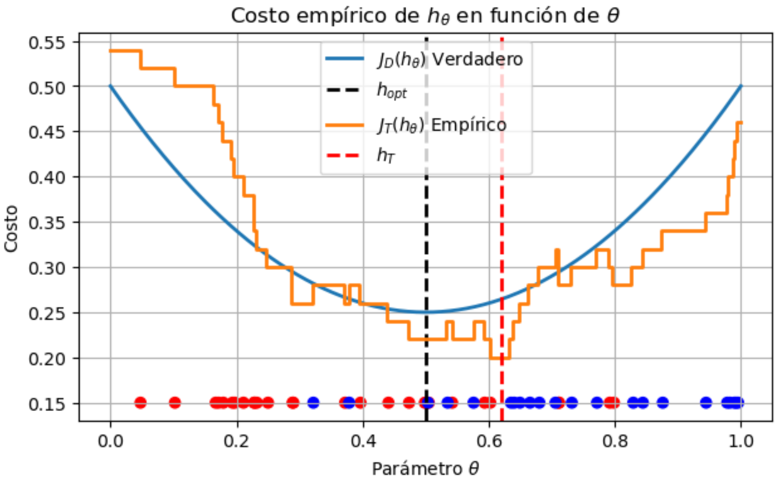
$$J_T(h) = \text{Cost}_T(h) = \underset{(\boldsymbol{x},y)\in T}{\boldsymbol{E}}\left[\text{Loss}\left(h(\boldsymbol{x}), y\right)\right] = \frac{1}{|T|}\sum_{(\boldsymbol{x},y)\in T} \text{Loss}\left(h(\boldsymbol{x}), y\right)$$

- **Minimizar el costo empírico (ERM)**: encontrar una hipótesis h_T tal que

$$h_T = \arg\min_{h\in\mathcal{H}} \text{Cost}_T(h)$$

- Pero si no tenemos cuidado, este enfoque puede conducir a sobreajuste ... más sobre esto en las próximas clases

Ejemplo: como NO conocemos $\mathcal{D}$



En resumen

Algoritmo de Machine Learning

Representa un procedimiento que a partir de datos genera una hipótesis:

$$\left.\begin{array}{l}\text{Datos}\\ \text{Loss}\\ \text{Sesgo inductivo}\end{array}\right\} \xrightarrow[\text{(.fit)}]{\text{ERM}} \text{Hipótesis o Modelo} = \left\{\begin{array}{l}\text{Parámetros}\\ \text{Procedimiento de Predicción}\\ \qquad\qquad\text{(.predict)}\end{array}\right.$$

En el ejemplo

$$\left\{\begin{array}{l}\text{Parámetros: } \theta \in [0,1]\\ \text{Predicción: azul si } x > \theta\text{; rojo si } x \leq \theta\end{array}\right.$$

Paramétrico vs No paramétrico

$$\left\{\begin{array}{l}\text{Paramétrico: } \dim\theta \text{ no depende de } N\\ \text{No paramétrico: } \dim\theta \text{ crece con } N\end{array}\right.$$

Bibliografía

- An introduction to statistical learning with applications in Python. Cap 2.
- Machine Learning - A First Course for Engineers and Scientists. Cap 2.
- Machine Learning Refined: Foundations, Algorithms, and Applications. Cap 1.
- Understanding Machine Learning: From Theory to Algorithms. Cap 2.