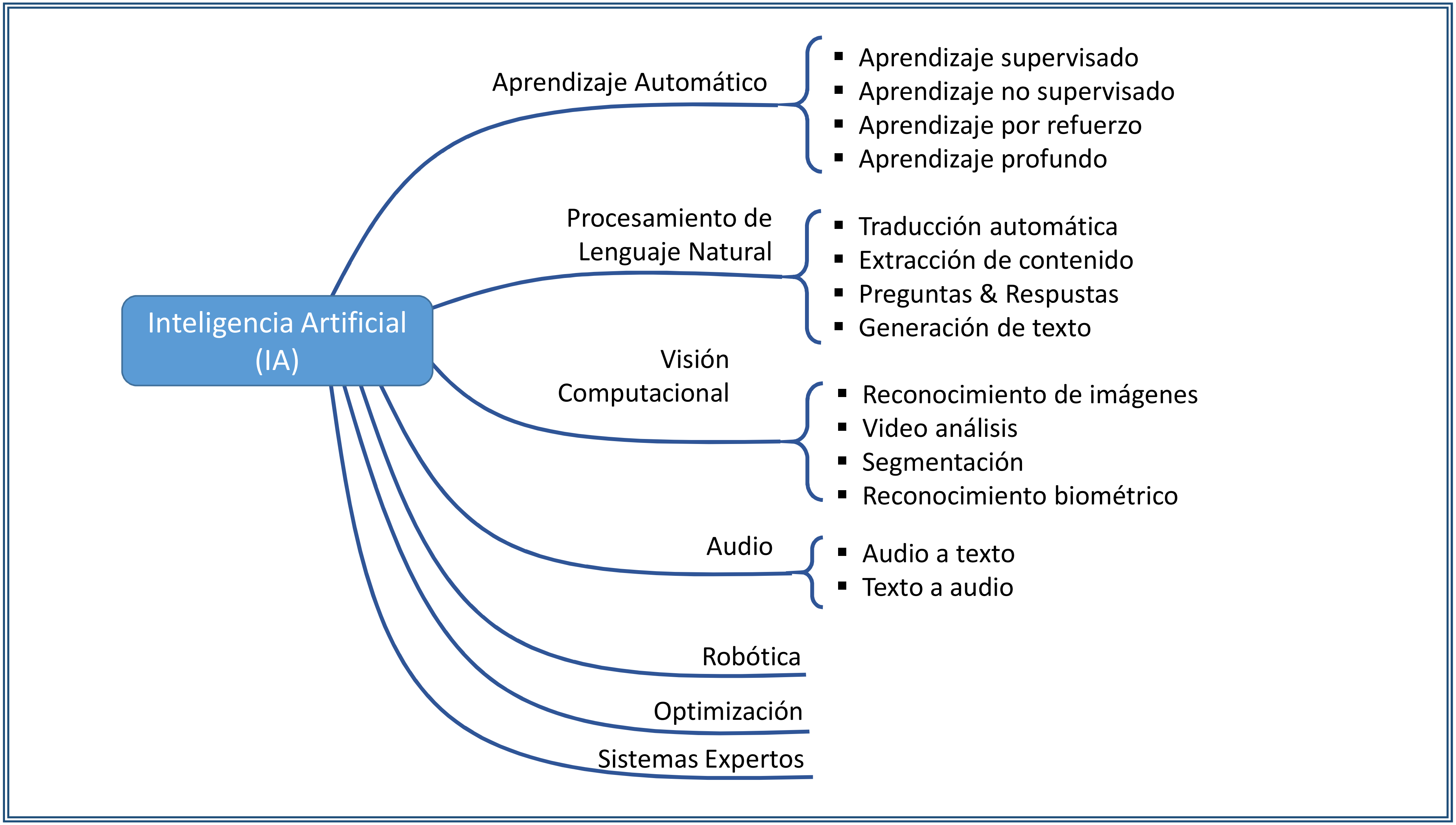
Maestría en Inteligencia Artificial Aplicada

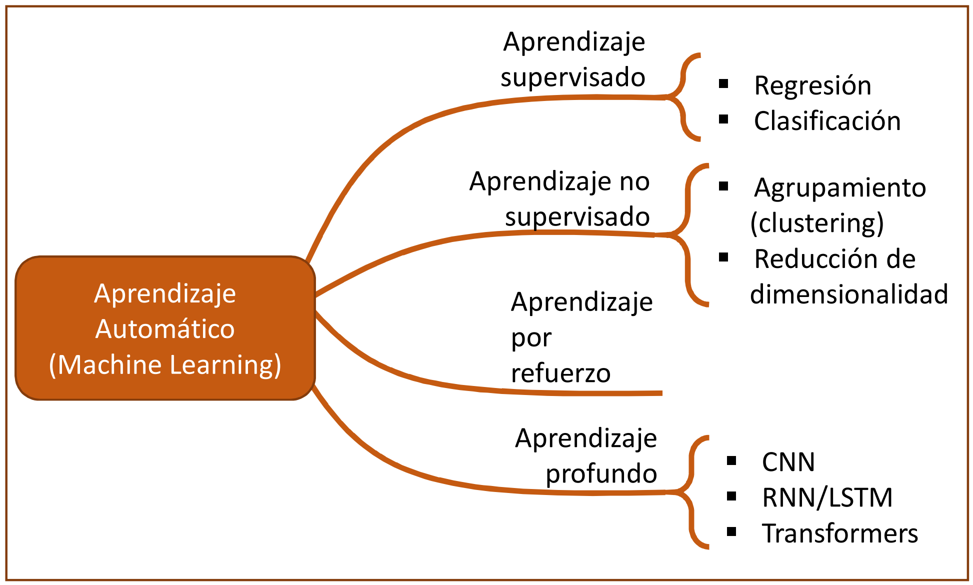
Regresión Logística

Inteligencia Artificial y Aprendizaje Automático

Tecnológico de Monterrey

Luis Eduardo Falcón Morales


Inteligencia Artificial (IA)
Aprendizaje Automático
Aprendizaje supervisado
Aprendizaje no supervisado
Aprendizaje por refuerzo
Aprendizaje profundo
Procesamiento de Lenguaje Natural
Traducción automática
Extracción de contenido
Preguntas & Respustas
Generación de texto
Visión Computacional
Reconocimiento de imágenes
Video análisis
Segmentación
Reconocimiento biométrico
Audio
Audio a texto
Texto a audio
Robótica
Optimización
Sistemas Expertos



Aprendizaje Automático (Machine Learning)
Aprendizaje supervisado
Regresión
Clasificación
Aprendizaje no supervisado
Agrupamiento (clustering)
Reducción de dimensionalidad
Aprendizaje por refuerzo
Aprendizaje profundo
CNN
RNN/LSTM
Transformers

Tipos de variable dependiente (o de salida)
para diferentes tipos de problemas:

- Reales (continuas):
  - Costo de una casa
  - Monto de crédito autorizado
  - Presión arterial

- Discretas/Categóricas:
  - Binarias:
    - Crédito aprobado: Sí/No
    - Comentario positivo: Sí/No
    - Calidad de un producto: Aprobado/No_Aprobado
  - Múltiples:
    - Productos más vendidos de un empresa en línea
    - Tipo de anuncio que más impacto a generado en redes sociales
    - Análisis de sentimiento de un servicio en 5 niveles

En esta sección estudiaremos el caso de la variable de salida binaria con distribución de probabilidad Bernoulli.

Función Logística

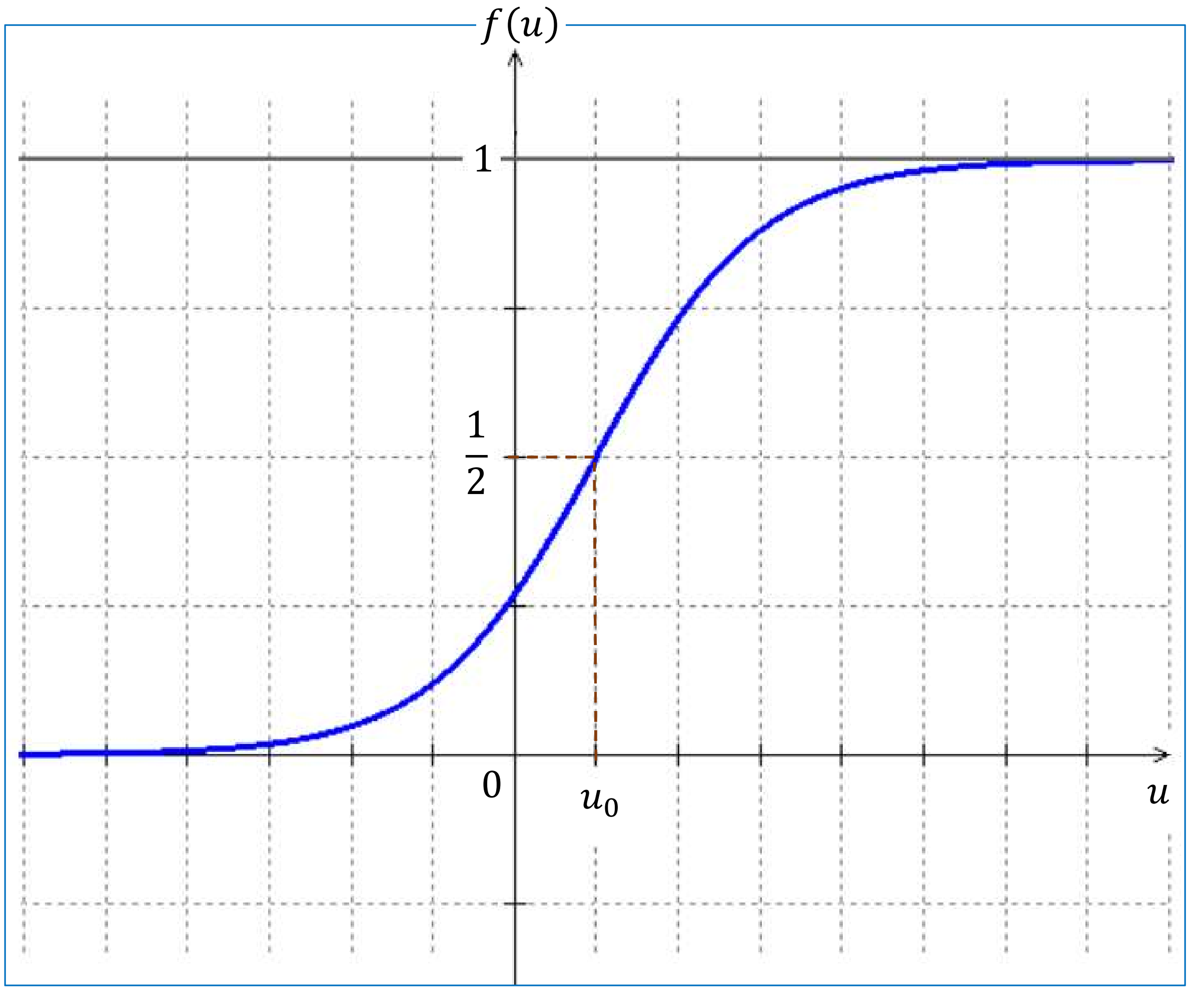
$$f(u) = \frac{1}{1 + e^{-\frac{u-u_0}{\delta}}}$$

donde:

u_0 : media de la v.a. u

δ : proporcional a la desviación estándar

Llamada en estadística Distribución Logística ya que cumple con las característica de una función de distribución de probabilidad acumulada.



Dominio: $(-\infty, +\infty)$
Rango: $(0, +1)$

Función Sigmoide

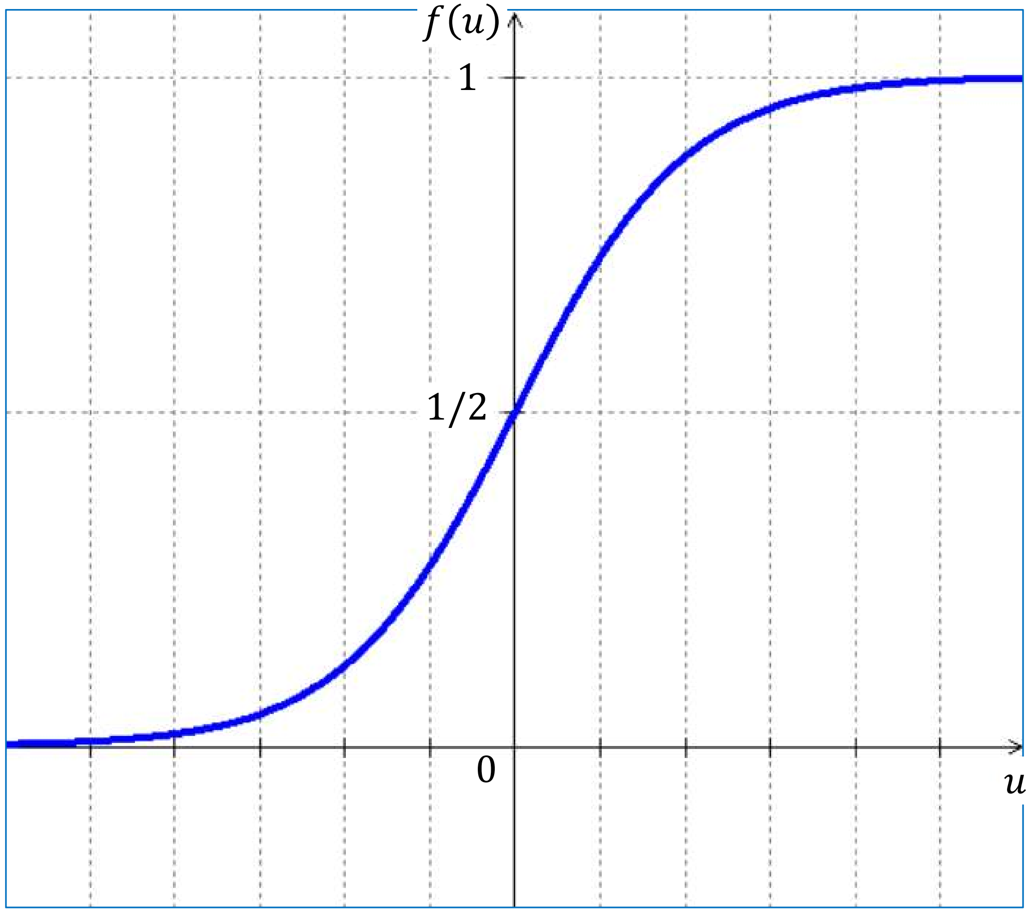
$$f(u) = \frac{1}{1 + e^{-u}}$$

O bien,

$$f(u) = \frac{e^{u}}{1 + e^{u}}$$

También llamada **función sigmoide exponencial**.

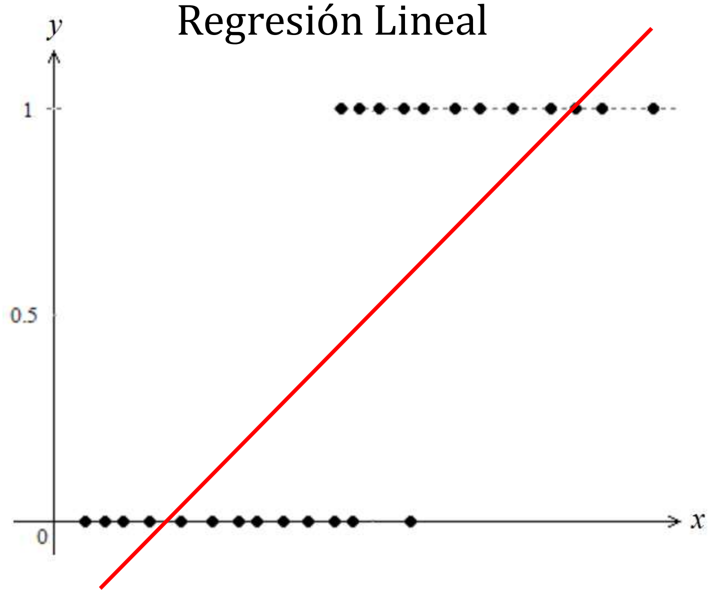
Observamos que la función sigmoide es un caso particular de la función logística, aunque muchos autores a esta función también la llaman función logística.



Dominio: $(-\infty, +\infty)$
Rango: $(0, +1)$

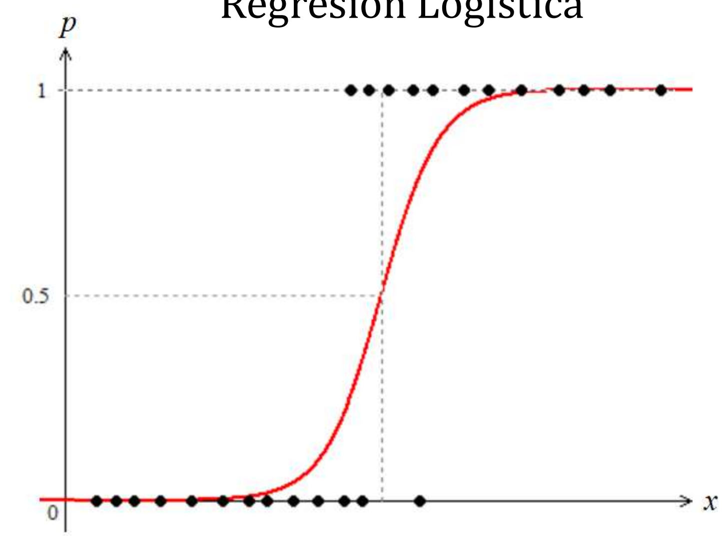
La función logística como un aproximador de la función escalón

Regresión Lineal



La función lineal Y generada para predecir el comportamiento de los datos mostrados. Rango en el intervalo $(-\infty, +\infty)$.

Regresión Logística

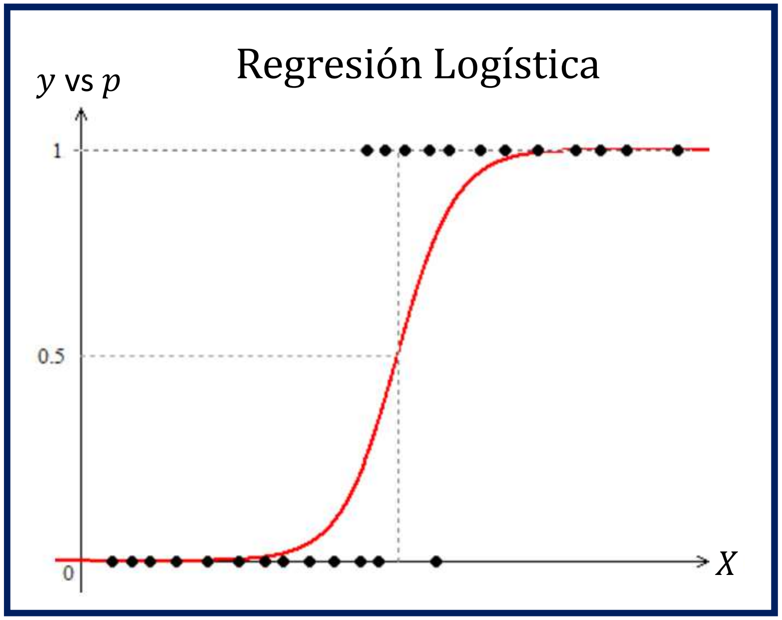


La función logística P generada para predecir el comportamiento de los datos mostrados. Rango "probabilístico" en el intervalo $(0, +1)$.

Aún cuando matemáticamente se puede utilizar el modelo de regresión lineal para encontrar la ecuación de la recta que aproxime la salida de un problema de clasificación, los resultados no serían muy satisfactorios.

La función logística permitirá aproximar con un sentido probabilístico los valores de predicción en un problema de clasificación.

Método de Optimización en un problema de Clasificación mediante una Función de Costo



En el caso del modelo de regresión logística para clasificación, se busca la función logística $p = f(X)$ que mejor aproxime los puntos dados $(x_k\ y_k)$.

Consideremos el conjunto de **datos de entrada**:

$$\{(x_{k1}, x_{k2}, \ldots, x_{km}, y_k)\}_{k=1}^{N}$$

de un problema de clasificación binario (logístico), donde las etiquetas $y_k \in \{0, 1\}$. El total de factores es m y la cantidad de registros es N. Entonces, se desea encontrar los parámetros $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_m)$ de un modelo que denotamos

$$\hat{y} = h_\beta(X)$$

donde $X_{N\times(m+1)} = [1, x_1, x_2, \ldots, x_m]$ y que minimice el valor de una **función de costo** J.

Simbólicamente lo podemos expresar este problema de optimización como:

$$\min_{\beta} J(\beta)$$

Función sigmoide: $$f(u) = \frac{1}{1+e^{-u}} \quad \begin{cases} \text{Dominio: } (-\infty, +\infty) \\ \text{Rango: } (0, 1) \end{cases}$$

Mediante un poco de manipulación algebraica podemos obtener la inversa de la función sigmoide:

$$p = \frac{1}{1+e^{-u}}$$

$$1 + e^{-u} = \frac{1}{p}$$

$$e^{-u} = \frac{1}{p} - 1$$

$$e^{-u} = \frac{1-p}{p}$$

$$\ln(e^{-u}) = \ln\left(\frac{1-p}{p}\right)$$

$$-u = \ln\left(\frac{1-p}{p}\right)$$

$$u = -\ln\left(\frac{1-p}{p}\right)$$

$$u = \ln\left(\frac{p}{1-p}\right)$$

Función Logit:
función inversa
de la sigmoide.

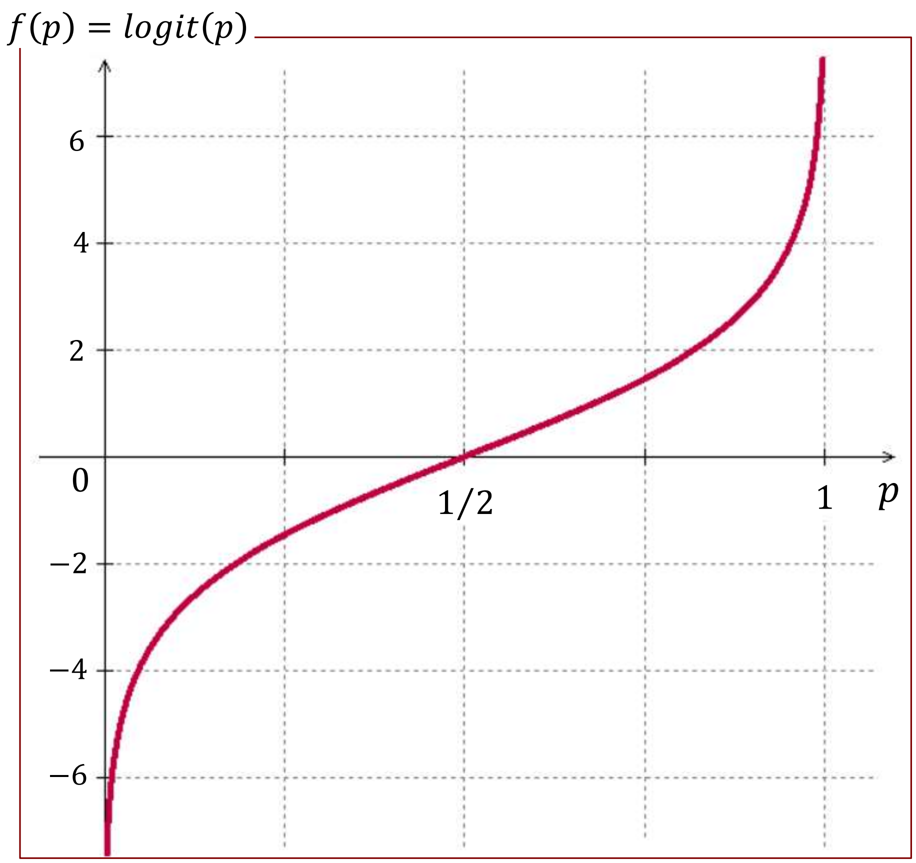
Dominio: $(0, 1)$

Rango: $(-\infty, +\infty)$

Función logit

$$logit(p) \equiv ln\left(\frac{p}{1-p}\right)$$

donde $0 < p < 1$



En la literatura Estadística la variable p está asociada a una probabilidad.
Al cociente

$$\frac{p}{1-p}$$

se le llama razón de oportunidades (*odds ratio*).

Observamos que la función *logit* es la función inversa de la función sigmoide.

Modelo de regresión logístico multivariable para la variable biclase Y:

$$logit(p) \equiv ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

Se considera a la variable dependiente Y con distribución Bernoulli, de manera que:

$$Prob[Y = y \mid x_1, x_2, \dots, x_m] = \begin{cases} p & si \quad y = 1 \\ 1-p & si \quad y = 0 \end{cases}$$

Formalmente: $logit(\mathbb{E}[Y \mid x_1, x_2, \dots, x_m]) = logit(p) = ln\big(p/(1-p)\big)$.

Al usar el logaritmo estamos extendiendo el Rango del intervalo $(0, 1)$ al intervalo $(-\infty, +\infty)$.

La manera de encontrar los coeficientes de regresión β_k es mediante alguno de los algoritmos numéricos, usualmente de la familia quasi-Newton.

Para variables cuantitativas, al aumentar una unidad en el factor x_j cuando los demás factores permanecen constantes, aumenta el logaritmo del cociente (odds) $\frac{p}{1-p}$ en β_j unidades.

Distribución de Bernoulli y Binomial

Una variable aleatoria discreta X se dice que sigue un modelo dicotómico o de Bernoulli, si al realizar un único experimento aleatorio sus posibles resultados son 1 o 0, con probabilidades p y $1-p$, respectivamente.

La función probabilidad de Bernoulli está dada como:

$$f(x) = \begin{cases} p & si\ x = 1 \\ 1-p & si\ x = 0 \\ 0 & en\ otro\ caso \end{cases}$$

La distribución Binomial es el caso en el cual se tienen n experimentos Bernoulli independientes entre sí.

Nota: Bernoulli es el caso particular de la distribución Binomial con $n = 1$.

Función de Costo
para el modelo de
Regresión Logística

Loss function:
Binary Cross-Entropy Loss/ Log-Loss

$$Loss = y\log(\hat{y}) + (1-y)\log(1-\hat{y})$$

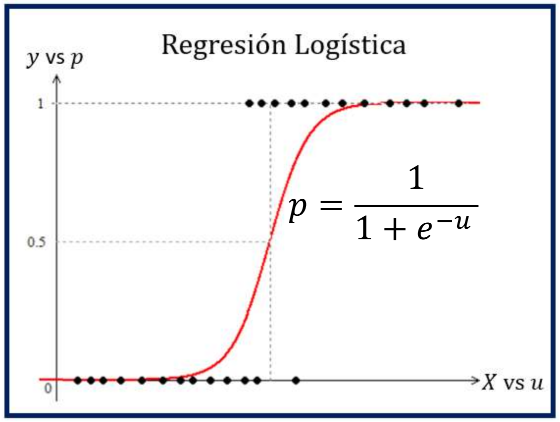
Tecnológico de Monterrey

Luis Eduardo Falcón Morales

Modelo de Regresión Logística

Consideremos el conjunto de **datos de entrada**: $\{(x_{k1}, x_{k2}, \dots, x_{km}, y_k)\}_{k=1}^{n}$, donde $y_k \in \{0, 1\}$.

Deseamos ajustar dichos datos al **modelo de regresión logística** para un problema biclase dada por la **hipótesis** $y = h_\beta(x)$:



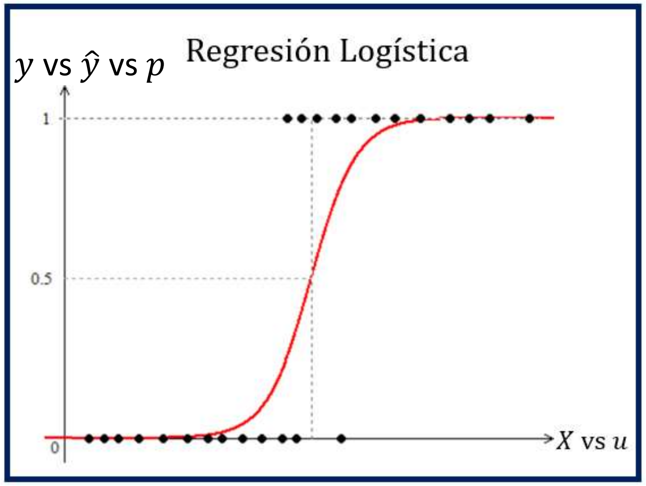
$$h_\beta(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

Se considera a la variable dependiente Y con distribución Bernoulli, de manera que:

$$Prob[Y = y \mid x_1, x_2, \dots, x_m] = \begin{cases} p & si \quad y = 1 \\ 1 - p & si \quad y = 0 \end{cases}$$

El problema de modelo Bernoulli

$$Prob[Y = y \mid x\,] = \begin{cases} p & si \quad y = 1 \\ 1 - p & si \quad y = 0 \end{cases}$$



se desea aproximar mediante la función logística multivariable

$$\hat{y} = p = h_\beta(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

Podemos sintetizar la representación de las igualdades anteriores como sigue:

$$\boxed{Prob[Y = y \mid x\,] = \hat{y}^y (1 - \hat{y})^{1-y} \text{ , donde } y \in \{0, 1\}}$$

y : es el valor real
$\hat{y}$: la aproximación/predicción

Podemos considerar esta expresión como una aproximación o predicción del valor real: $\hat{y} \longmapsto y$.

$Prob[Y = y \mid x] = \hat{y}^y (1-\hat{y})^{1-y}$

Estimador de Máxima Verosimilitud
(*Maximum Likelihood Estimation*)

Supongamos que tenemos un problema biclase y deseamos obtener un modelo de predicción $\hat{y} = h_\beta(X) \in (0,1)$ a partir de un conjunto de datos de entrenamiento $\{\ (\vec{x}_k, y_k)\ \}_{k=1}^{N} = \{\ (x_{k1}, x_{k2}, \dots, x_{km}, y_k)\ \}_{k=1}^{N}$, donde $y_k \in \{0,1\}$.

La **función de máxima verosimilitud** o **estimador de máxima verosimilitud**, busca los parámetros β del modelo $\hat{y} = h_\beta(X) \in (0,1)$ que maximice la probabilidad de predicción de la clase correcta para cada dato de entrenamiento, es decir:

$$\max_{\beta} \left\{ \prod_{k=1}^{N} \hat{y}_k^{y_k} \cdot (1-\hat{y}_k)^{1-y_k} \right\}$$

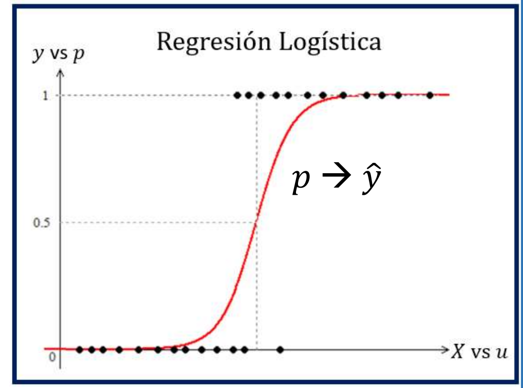
En los problemas de clasificación es común considerar como modelo de predicción la función sigmoide:

$$\hat{y}_k = \hat{h}_\beta(\vec{x}_k) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{k1}+\cdots+\beta_m x_{km})}}$$

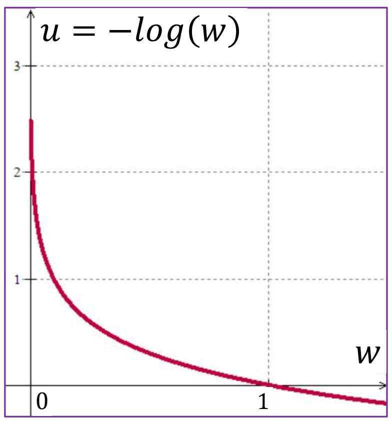
Función de costo de la Entropía Cruzada Binaria (cost function Binary Cross-Entroy / Log-loss)

En Estadística es usual considerar el negativo del logaritmo del estimador de verosimilitud.
Es decir, a partir de la expresión:

$$Prob[Y = y \mid x\,] = \prod_{k=1}^{N} \hat{y}_k^{\,y_k} \cdot (1 - \hat{y}_k)^{1-y_k}$$



aplicamos logaritmos en ambos lados de la igualdad:



$$\log\{P[Y = y \mid x\,]\} = \log\left\{\prod_{k=1}^{N} \hat{y}_k^{\,y_k} \cdot (1 - \hat{y}_k)^{1-y_k}\right\}$$

$$= \sum_{k=1}^{N} \{y_k \log(\hat{y}_k) + (1 - y_k)\log(1 - \hat{y}_k)\}$$

$$\equiv -\mathcal{L}oss$$

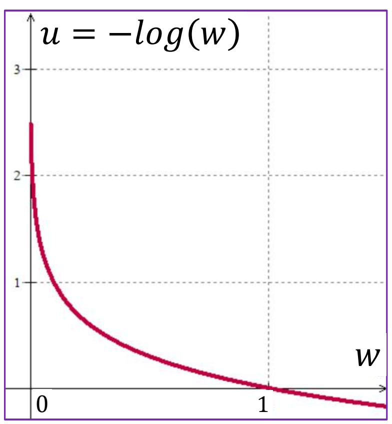
Como se tratará de un problema de minimización, conviene ponerle un signo negativo.

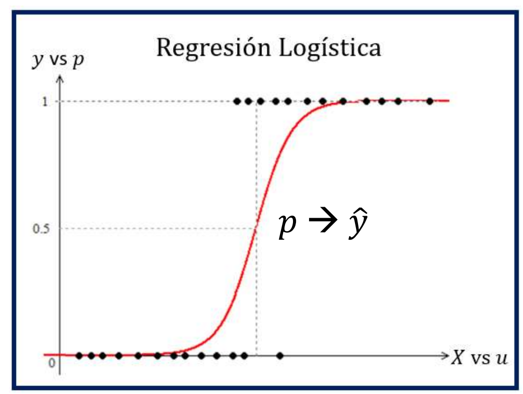
Observa que esta expresión ya se puede considerar como una medida del error, la cual podemos promediar y considerar como una medida sobre todos los datos de entrenamiento $x_1, x_2, \ldots, x_N$:

Así, la expresión anterior (el negativo del logaritmo del estimador de verosimilitud) la podemos promediar y considerarla como una medida del error sobre todos los datos de entrenamiento $x_1, x_2, \ldots, x_N$:

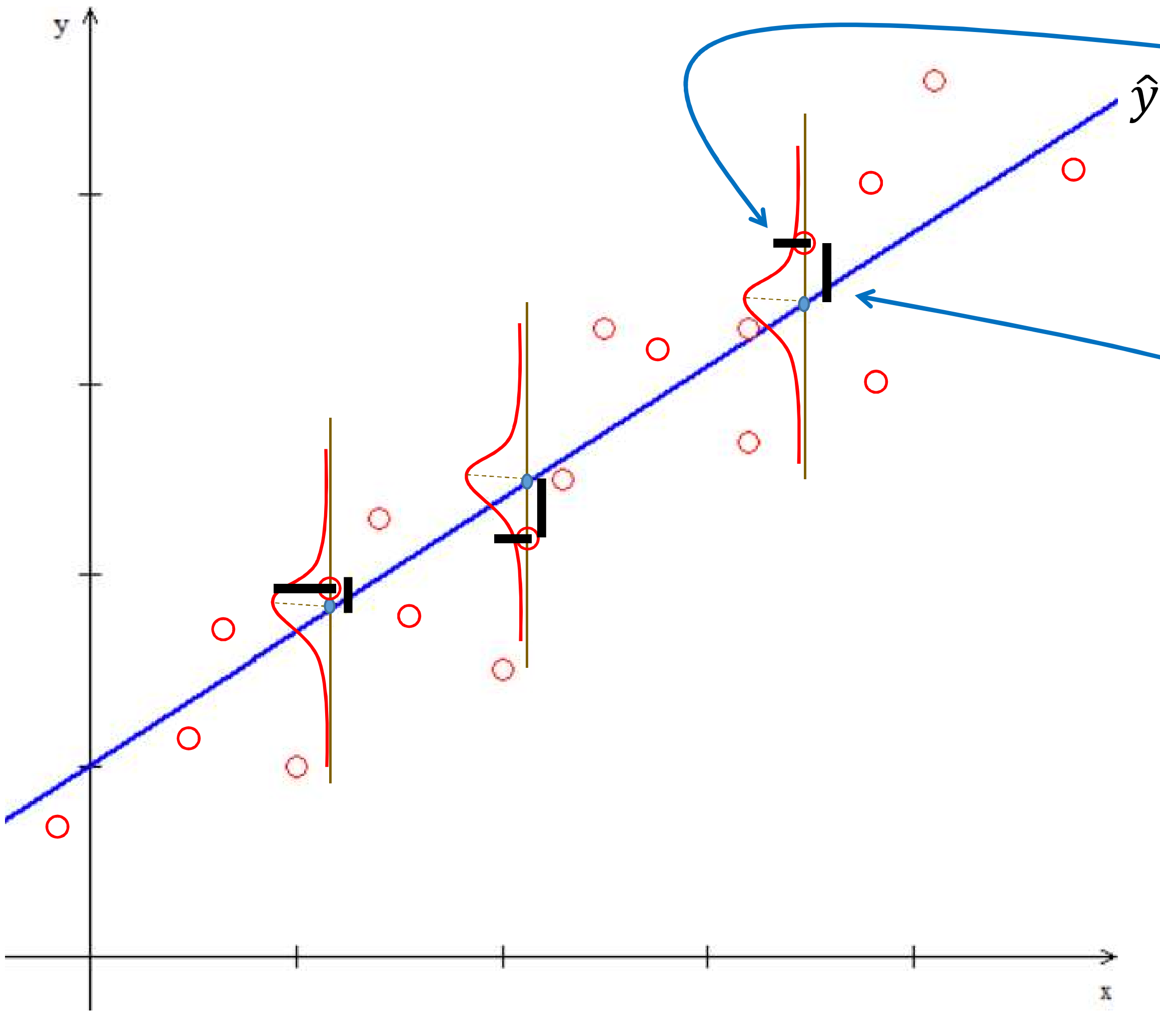
A esta función se le llama **Función de Costo de Entropía Cruzada para el caso Binario (cost function Binary Cross-Entropy / Log-loss)**.

$$Loss = J = -\frac{1}{N}\sum_{k=1}^{N}\{y_k log(\hat{y}_k) + (1 - y_k) log(1 - \hat{y}_k)\}$$





Maximizar el estimador de verosimilitud vs **Minimizar** la distancia de suma de cuadrados



- **Maximizar la probabilidad (verosimilitud)** de que cada punto esté dentro de la función de distribución gaussiana centrada en $\hat{y}$ con varianza σ^2.

- **Minimizar la distancia (errores/residuos)** de cada punto a $\hat{y}$.

$\hat{y}$: modelo de predicción a partir del conjunto de datos dado.

Modelo de Regresión Lineal Generalizado

Generalized Linear Model (*GLM*):

Supuestos sobre la variable dependiente de salida:

Distribución gaussiana de los errores:
Modelo de Regresión Lineal Estándar

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + error$$

Distribución de los errores diferente a la gaussiana:
Modelo de Regresión Lineal Generalizado

En particular al usar la función Logística:
Modelo de Regresión Logística