UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

REGRESIÓN LOGÍSTICA

DR. JORGE HERMOSILLO VALADEZ

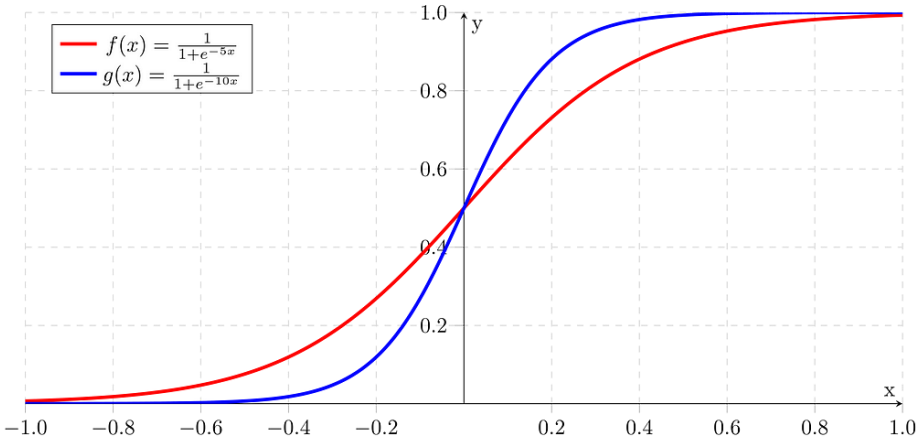
LABORATORIO DE SEMÁNTICA COMPUTACIONAL

¿QUÉ ES LA REGRESIÓN LOGÍSTICA?

REGRESIÓN LOGÍSTICA (NOCIÓN BÁSICA)

- Modelo estadístico utilizado para modelar una variable binaria usando la función logística (o sigmoide):

$$F(x) = \sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$



- Se le llama "regresión logística" por tratarse del tratamiento probabilista de una función de regresión lineal.

Regresión lineal:

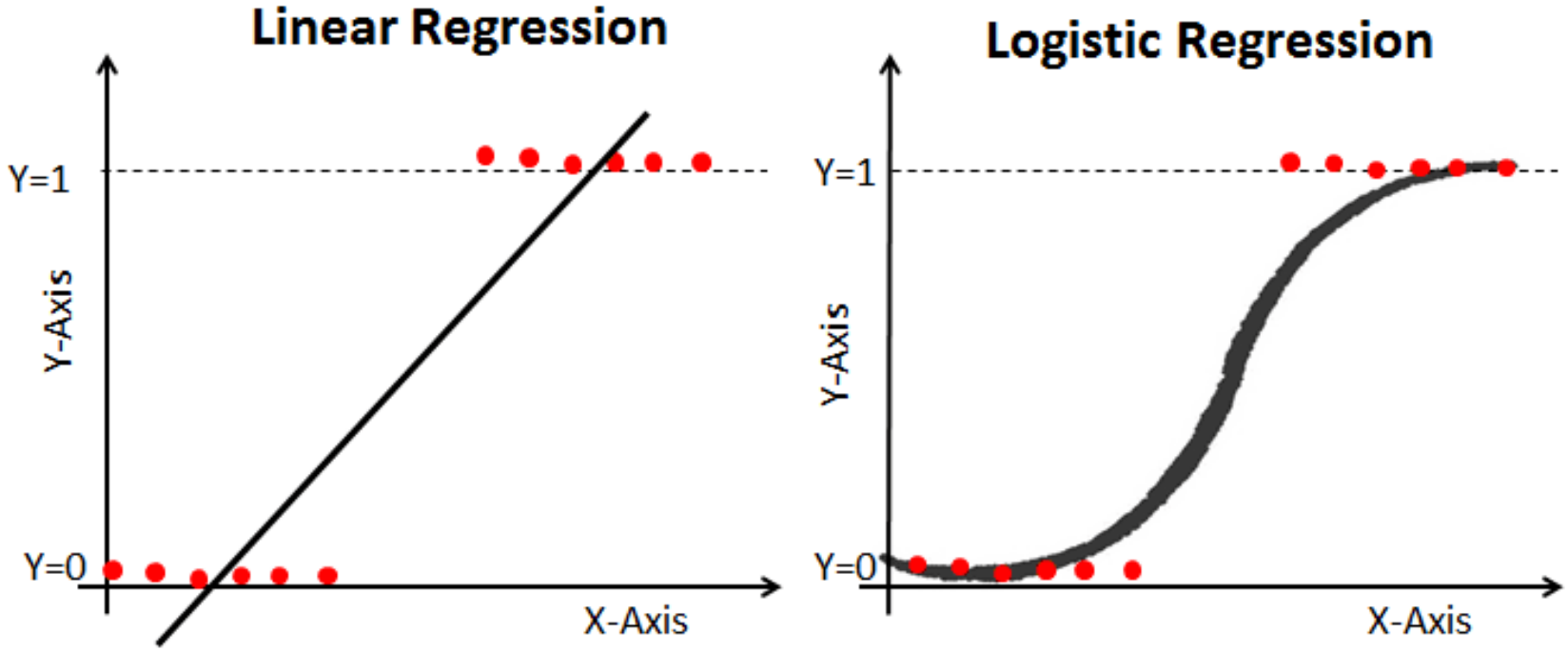
$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

Función sigmoide:

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

- Se utiliza para la clasificación binaria, pero puede usarse también en la clasificación multiclase.

REGRESIÓN LOGÍSTICA (INTUICIÓN)



La regresión lineal explica la variable y en función de x:

$$y = \alpha_0 + \alpha_1 x_1 = z$$

La relación logística predice la probabilidad de un nuevo dato x de pertenecer a la clase $y = 1$. Entonces, queremos convertir los valores predichos z en una probabilidad: $p = \frac{1}{1+e^{-z}}$

REGRESIÓN LOGÍSTICA VS REGRESIÓN LINEAL

- En la Reg.Log. la variable dependiente sigue una ley de distribución de Bernoulli. (puede ser multinomial u ordinal). La salida de la Reg,Lin. es un valor continuo.

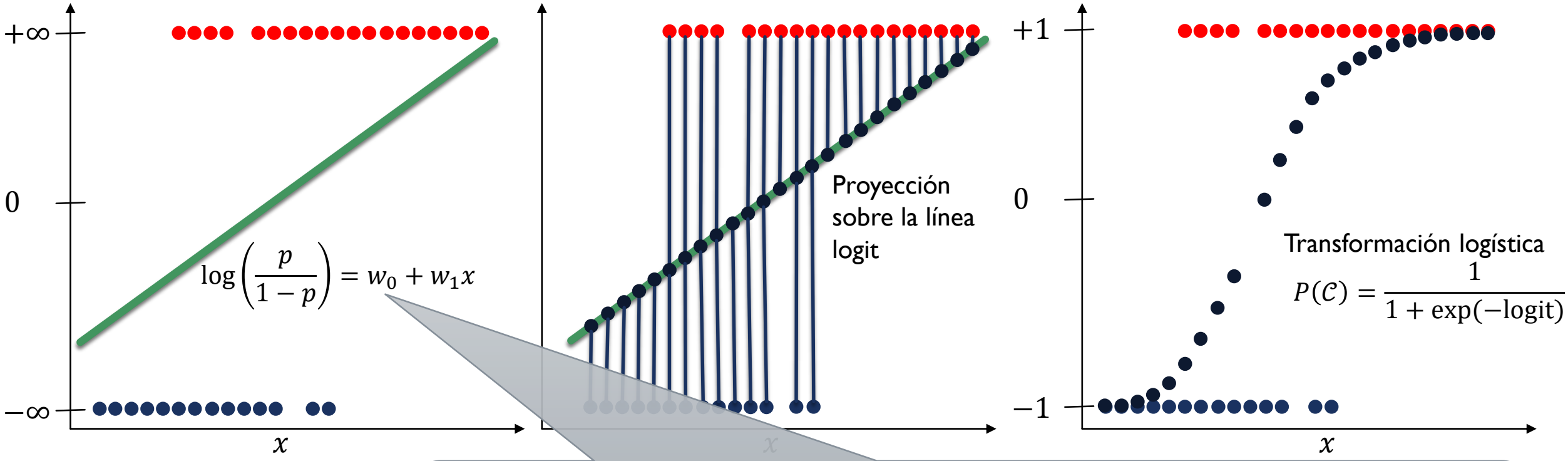
- La Reg.Log requiere observaciones independientes unas de otras, supone linealidad de las variables independientes (log lineales) y ligera correlación entre ellas. Requiere de una muestra de datos grande para obtener alta exactitud.

- Los parámetros en Reg.Lin. se estiman usando Mínimos Cuadrados (OLS), en Reg.Log se estiman usando el Máximo de Verosimilitud (*Maximum Likelihood* MLE).

REGRESIÓN LOGÍSTICA: FUNDAMENTOS I

1. La Razón de Probabilidad (*RP – Odds*) es la relación entre la probabilidad de que algo ocurra y la probabilidad de que no ocurra. También es una métrica que representa la probabilidad de que ocurra el suceso (versosimilitud o likelihood). En el caso binario:
   - $p(y = 1|\mathbf{x}) = p, p(y = 0|\mathbf{x}) = 1 - p$
   - $RP = \frac{p}{1-p}$ si $RP > 1$ la decision es $y = 1$; si $0 < RP < 1$ la decision es $y = 0$.
2. La función ***logit*** es el **logaritmo de la razón de probabilidad** (log-odds). Toma valores $[0,1]$ y devuelve valores $(-\infty, +\infty)$
   - $\text{logit}(RP) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p)$
   - Se puede demostrar que $\text{logit}(x) = \sigma^{-1}(x)$. Toma y $= \text{logit}(x) = \log\frac{x}{1-x}$ y calcula e^y.
3. La Reg.Log supone que $\text{logit}(RP) = \text{logit}(odds) = \log\left(\frac{p}{1-p}\right) = w_0 + w_1x_1 + \cdots + w_nx_n$
4. La probablidad de RP sería: la evaluación de la regresión lineal en la sigmoide:
   - $p = \frac{1}{1+e^{-(w_0+w_1x_1+\cdots+w_nx_n)}}$

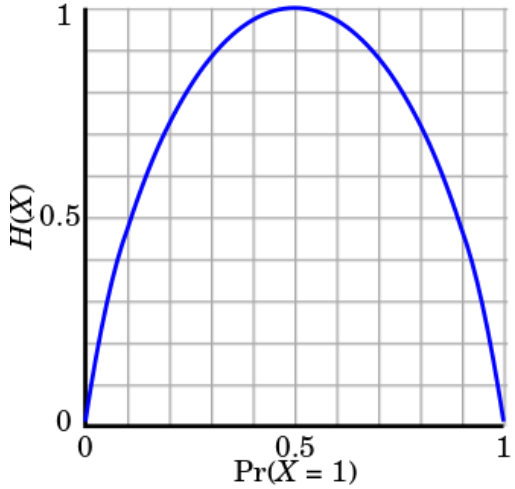
REGRESIÓN LOGÍSTICA: RESUMEN GRÁFICO



El problema consiste en estimar los pesos usando MLE

ENTROPÍA

1. La Entropía de una Variable Aleatoria (VA) es su incertidumbre promedio:
   - $H(X) = -\sum_{x \in X} p(x) \log p(x)$
   - $Mide\ la\ cantidad\ de\ información\ de\ una\ VA.$



2. La Entropía conjunta de un par de VA's es la cantidad de información que se necesita en promedio para determinar sus valores conjuntos:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

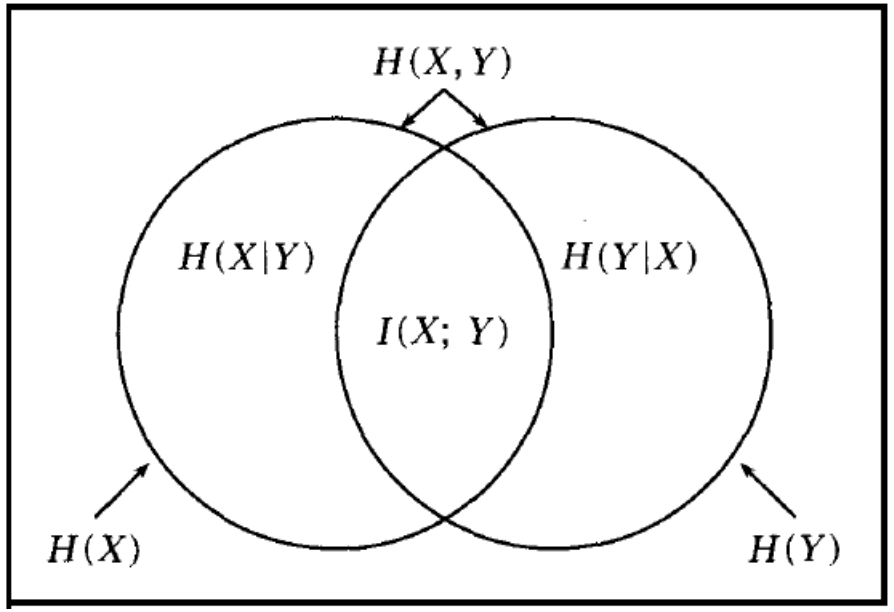
3. La Entropía condicional de una VA Y dada otra VA X, para X,Y ~p(x,y) es la cantidad de información que se necesita en promedio para determinar Y dado que se conoce X:

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x)$$

INFORMACIÓN MUTUA

1. Por la regla de la cadena para la entropía
$$H(X,Y) = H(X) + H(Y|X) = H(X) + H(X|Y)$$
2. Por lo tanto:
$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Esta diferencia es la Información Mutua (I) entre X y Y. Es la reducción en incertidumbre de una VA conociendo la otra; la cantidad de información que una VA contiene acerca la otra.



$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Es conveniente ver a la Información Mutua como una medida de independencia entre dos VA's porque:

1. Es 0 sólo cuando ambas VA's son independientes.
2. Para dos VA's dependientes I crece no sólo con el grado de dependencia, sino también acorde con la entropía de las variables.

ENTROPÍA RELATIVA O DIVERGENCIA KL

Dadas dos funciones de distribución de probabilidad p(x), q(x), su entropía relativa está dada por:

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

También conocida como la divergencia Kullback-Leibler, es una medida de cuán diferentes son dos distribuciones de probabilidad (sobre el mismo espacio de eventos).

$$D(p||q) = E_p \left(\log \frac{p(x)}{q(x)} \right)$$

La divergencia KL entre p y q es el número de bits promedio que se desperdician codificando eventos de una distribución p mediante un código basado en una no-tan-correcta distribución q.

Esta cantidad es siempre no-negativa, y $D(p||q) = 0$ ssi $p = q$. Sin embargo, no es una distancia propiamente dicha, porque no cumple con la desigualdad del triángulo.

La Información Mutua es una medida de que tan lejos está una distribución conjunta de la independencia:

$$I(X;Y) = D(p(x,y)||p(x)p(y))$$

ENTROPÍA CRUZADA

La entropía cruzada entre una VA X con distribución de probabilidad real p(x) y otra función de densidad de probabilidad q (usualmente un modelo de p) está dada por:

$$H(X,q) = H(X) + D(p||q) = -\sum_{x \in X} p(x) \log q(x)$$

$$= E_p\left(\log \frac{1}{q(x)}\right)$$

Para el problema de la regresión logística, la idea es usar la entropía cruzada binaria como una función de error, que permite saber cuán diferentes son las distribuciones de las etiquetas reales ($y = 1$) de las predichas por el modelo ($p(y|\mathbf{x})$).

$$\text{Costo}^{\text{i}} = -\sum_{y \in Y} y^i \log p^i = -[y^i \log p^i + (1 - y^i) \log(1 - p^i)]$$

ENTROPÍA CRUZADA Y MLE

- Por cada punto de entrenamiento $\mathbf{x}$, la clase a predecir es y, y $p(y = 1|\mathbf{x}) = p, p(y = 0|\mathbf{x}) = 1 - p$
- Supongamos n puntos y $d + 1$ dimensiones; es decir, cada punto es $(1, \mathbf{x})$ con $\dim(\mathbf{x}) = d$. La función de verosimilitud (likelihood) se escribe:

$$L(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

$$J(\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i) \qquad \text{Entropía Cruzada}$$

$$= \sum_{i=1}^{n} \log(1 - p_i) + \sum_{i=1}^{n} y_i \log \frac{p_i}{1 - p_i}$$

- $\log\left(\frac{p}{1-p}\right) = \mathbf{w} \cdot \mathbf{x}$, y $p = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathbf{x},)}}$

MLE

$$\frac{\partial J(\mathbf{x},\mathbf{w})}{\partial \mathbf{w}} = \sum_{i=1}^{n} -\frac{\partial}{\partial \mathbf{w}} \log(e^{\mathbf{w}\cdot\mathbf{x}_i}+1) + \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{w}} y_i(\mathbf{w}\cdot\mathbf{x_i})$$

$$= \sum_{i=1}^{n} -\mathbf{x_i} p + \sum_{i=1}^{n} y_i \mathbf{x}_i = \sum_{i=1}^{n} (y_i - p)\mathbf{x}_i$$

Buscamos ahora $-\frac{\partial J(\mathbf{x},\mathbf{w})}{\partial \mathbf{w}} = 0$ por descenso de gradiente.

REFERENCIAS

- Flach, Peter (2012). _Machine Learning: The Art and Science of Algorithms that Make Sense of Data_. Cambridge University Press.

- Bishop, Christopher M. (2006). _Pattern recognition and machine learning_. New York. Springer.

- Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.