Tecnológico de Monterrey

Aprendizaje automático

Regresión logística

Dr. Omar Mendoza Montoya

07/08/2024

Regresión logística es un modelo de clasificación probabilístico, en el sentido que intenta modelar $P(L = 0|X = x)$ y $P(L = 1|X = x)$.

La regla de decisión se define por:

$$P(L = 1|X = x) > P(L = 0|X = x)$$

Lo cual es equivalente a la siguiente relación:

$$\frac{P(L = 1|X = x)}{P(L = 0|X = x)} > 1$$

Regresión logística

- Aun cuando regresión logística es un clasificador probabilístico, su derivación viene de la regresión lineal generalizada.

- En regresión lineal múltiple, para una observación (x, y), la variable de respuesta y se modela como:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p = x\beta$$

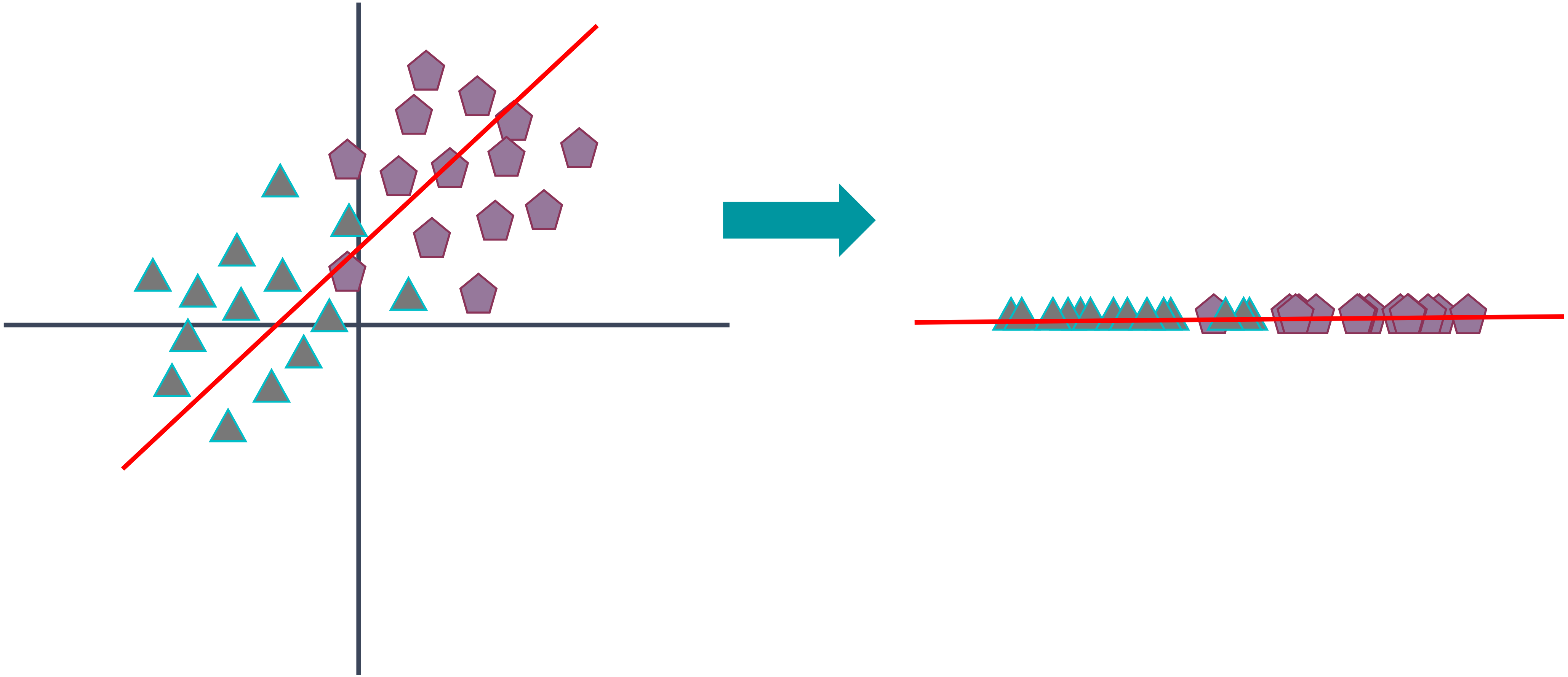
- En regresión logística, para una observación (x, l), se modela una variable de respuesta $P(L = l|X = x)$ está dada por:

$$P(L = l|X = x) = \begin{cases} \dfrac{1}{1 + e^{x\beta}} & \text{si } l = 0 \\ \dfrac{1}{1 + e^{-x\beta}} & \text{si } l = 1 \end{cases}$$

¿Cómo funciona regresión logística?

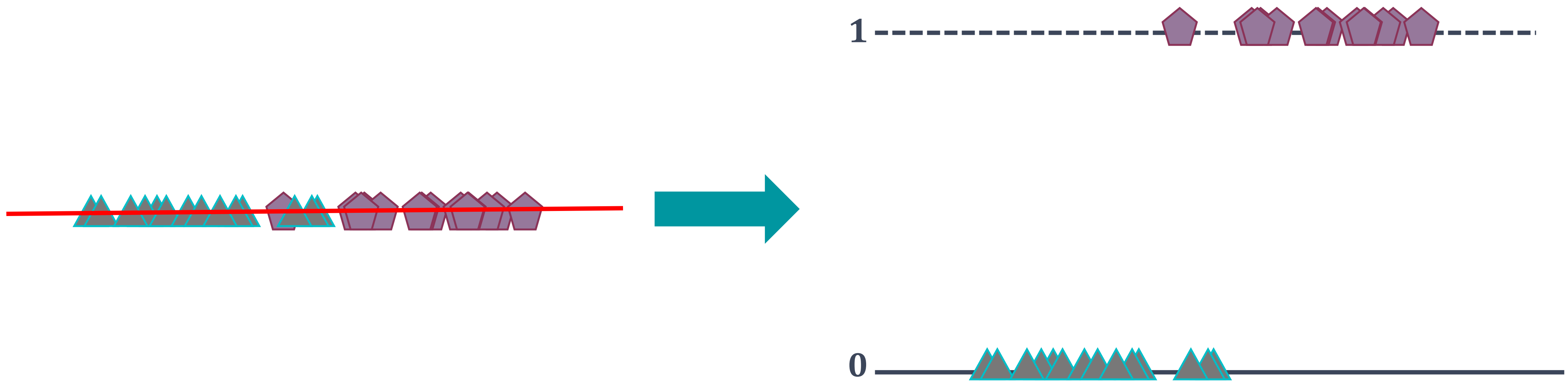
Proyección lineal de los datos

- La parte lineal del modelo ($x\beta$) tiene la función de proyectar los datos a una línea.



Separación en clases

- Los datos pasan a un plano donde las observaciones de una clase tienen un valor en el eje vertical de 0, y las de la otra clase un valor de 1.

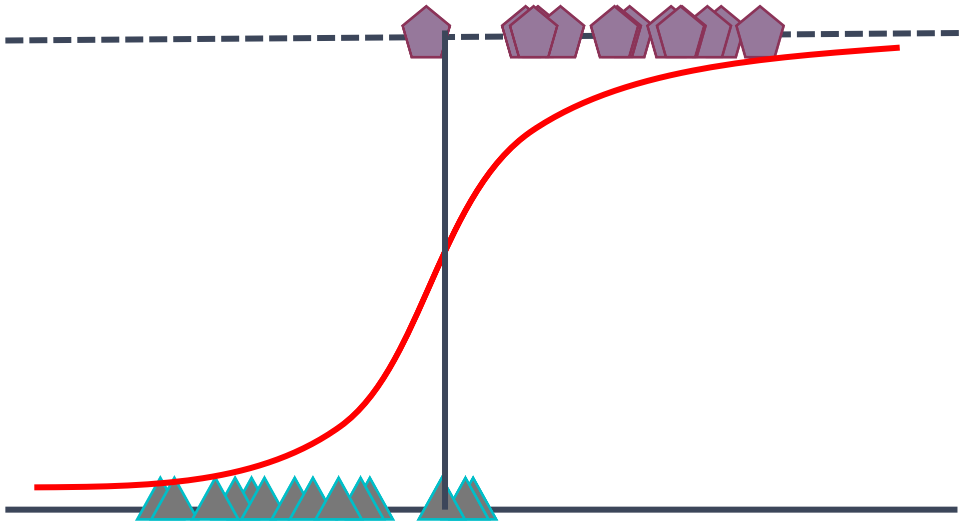


Función logística

- Para poder generar un modelo continuo en este nuevo espacio, se utiliza la función logística:

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

- Dicha función es cercana a cero cuando u es negativa (es decir $x\beta$), y cercana a uno cuando u es positiva.



Ajuste del modelo

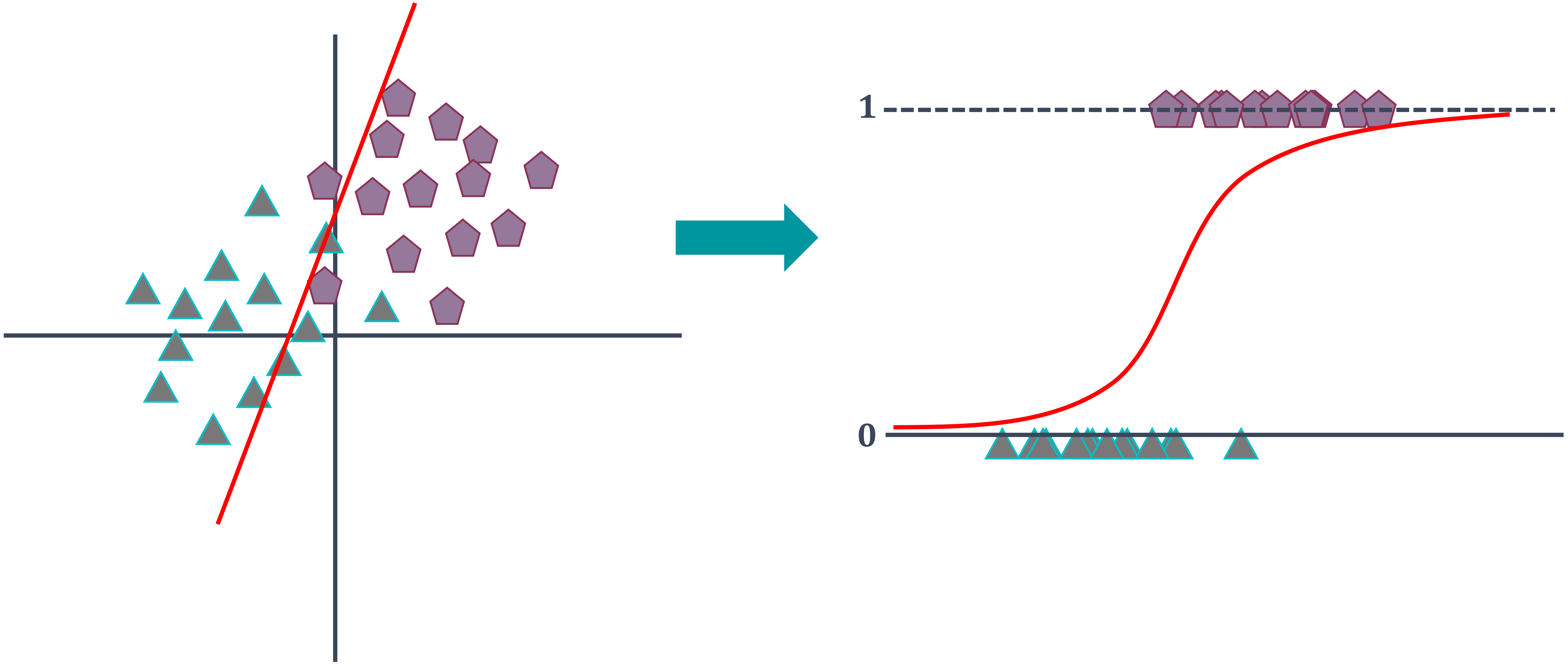
Ajuste del modelo de regression logística

- El problema en regresión logística es encontrar la línea de proyección tal que las clases se separen lo mejor posible en el espacio donde se modelan los datos proyectados.

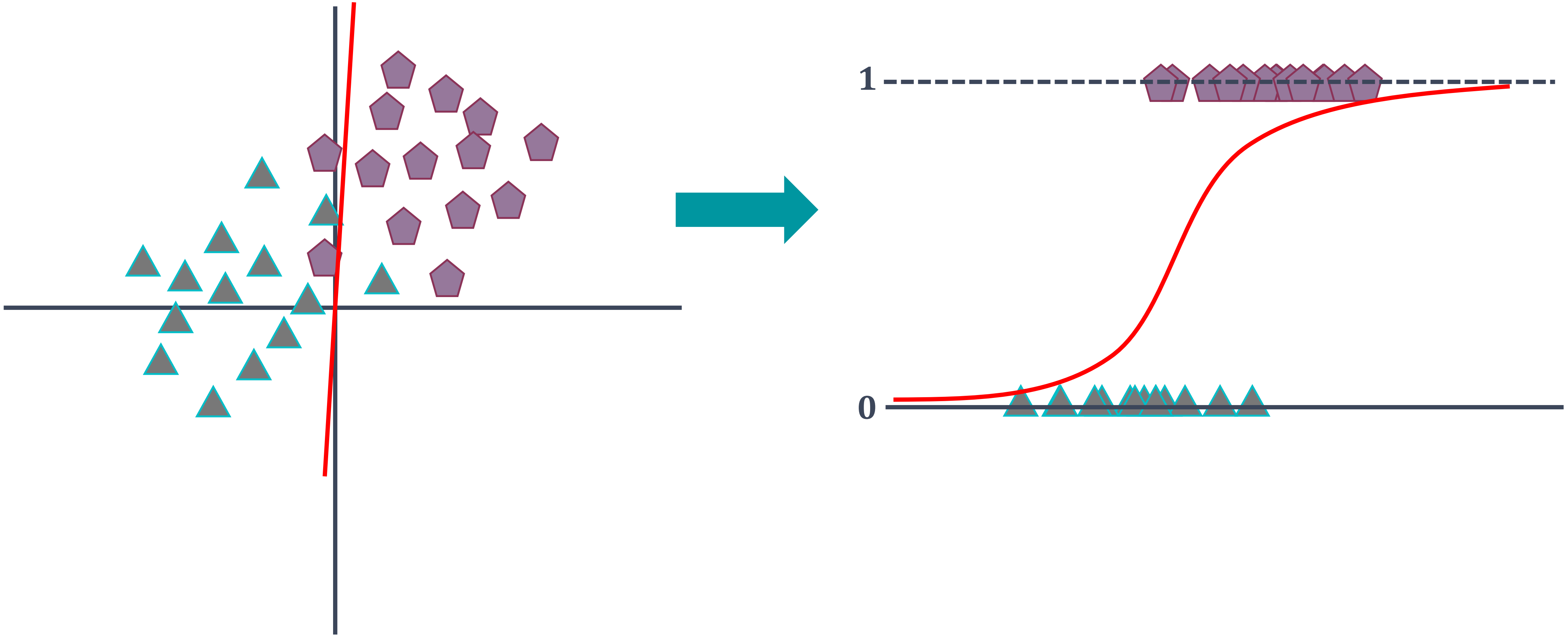
- En otras palabras, queremos que, en lo posible, para la mayoría de las observaciones de la clase 0 su respectiva probabilidad $P(L = 0|X = x)$ sea 1, mientras que para los datos de la clase 1 su respectiva probabilidad $P(L = 1|X = x)$ sea 1.

- Esto se resuelve como un problema de optimización.

Ejemplos de proyecciones



Ejemplos de proyecciones



Métodos para el ajuste del modelo

- Estimación de máxima verosimilitud
  - Método favorito, usualmente se integra en las librerías.
- Estimación por mínimos cuadrados
  - No es el preferido, pero algunas librerías lo incluyen.

Método de máxima verosimilitud

- En este método, para un conjunto de datos $D = \{(x_1, l_1), (x_2, l_2), \dots, (x_n, l_n)\}$ con n observaciones, se resuelve el siguiente problema de optimización:

$$\beta^* = \arg\min_{\beta} \left(-\sum_{i=1}^{n} \ln\big(P(L = l_i | X = x_i)\big) \right)$$

- Este problema indica que queremos los coeficientes β tales que minimicen la función de −log-verosimilitud, o su problema equivalente, encontrar los coeficientes que maximizan la función de verosimilitud.

Método de máxima verosimilitud

- Para optimización de gradiente:

$$\nabla L = \sum_{i \in C_0} \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} x_i^T - \sum_{i \in C_1} \frac{e^{-x_i \beta}}{1 + e^{-x_i \beta}} x_i^T$$

donde L es la función de costo (-log-verosimilitud), C_0 es el conjunto de índices con observaciones de la clase 0 y C_1 es el conjunto de observaciones de la clase 1.

Método de mínimos cuadrados

- Para un conjunto de datos $D = \{(x_1, l_1), (x_2, l_2), \dots, (x_n, l_n)\}$ con n observaciones, se resuelve el siguiente problema de optimización:

$$\beta^* = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(l_i - g(x_i \beta)\right)^2 = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(l_i - \frac{1}{1 + e^{-x_i \beta}}\right)^2$$

- En este caso, queremos el conjunto de parámetros β tales que minimicen el error cuadrático medio del modelo al predecir la etiqueta correcta.

Método de mínimos cuadrados

- Para optimización de gradiente:

$$\nabla MSE(D,\beta) = -\frac{2}{n}\sum_{i=1}^{n} r_i \frac{e^{-x_i\beta}}{(1+e^{-x_i\beta})^2} x_i^T$$

Nótese que sólo agregamos el término $\frac{e^{-x_i\beta}}{\left(1+e^{-x_i\beta}\right)^2}$ a la expresión que obtuvimos para regresión lineal.

Para ambos casos, los parámetros del modelo se ajustan con algún método iterativo como descenso de gradiente.

Regularización de parámetros

Regularización

- Al igual que otros modelos de regresión, es posible agregar un término de regularización al problema de optimización.

Para el método de máxima verosimilitud:

$$\beta^* = \arg\min_{\beta}\left(-\sum_{i=1}^{n}\ln\left(P(L = l_i | X = x_i)\right) + \lambda\sum_{i=1}^{p}\beta_i^2\right)$$

$$\nabla L = \sum_{i\in C_0}\frac{e^{x_i\beta}}{1+e^{x_i\beta}}x_i^T - \sum_{i\in C_1}\frac{e^{-x_i\beta}}{1+e^{-x_i\beta}}x_i^T + 2\lambda\beta$$

Regularización

Para el método de mínimos cuadrados:

$$\beta^* = \arg\min_{\beta} \left(\frac{1}{n} \sum_{i=1}^{n} \left(l_i - \frac{1}{1 + e^{-x_i \beta}} \right)^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right)$$

$$\nabla MSE(D, \beta) = -\frac{2}{n} \sum_{i=1}^{n} r_i \frac{e^{-x_i \beta}}{(1 + e^{-x_i \beta})^2} x_i^T + 2\lambda\beta$$

Modelos multiclase

Regresión logística multiclase

- Regresión logística multinomial
  - Se selecciona la clase k como referencia, y con ella se requieren $k-1$ parámetros β_i:

$$P(L=l|X=x)=\begin{cases}\frac{1}{1+\sum_{i=1}^{K-1}e^{x\beta_i}} & \text{si } l=1,2,3,\ldots k-1\\ \frac{e^{x\beta_k}}{1+\sum_{i=1}^{K-1}e^{x\beta_i}} & \text{si } l=k\end{cases}$$

- Regresión logística softmax
  - Se asume que para todas las clases:

$$P(L=l|X=x)=\frac{e^{x\beta_k}}{\sum_{i=1}^{K-1}e^{x\beta_i}}$$

Bibliografía

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 4

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 4