

09481: Inteligencia Artificial

Profesor del curso: Breyner Posso, Ing. M.Sc.
e-mail: breyner.posso1@u.icesi.edu.co

Programa de Ingeniería de Sistemas.
Departamento TIC.
Facultad de Ingeniería.
Universidad Icesi.
Cali, Colombia.

Agenda

1. Introducción
2. Regresión Logística

1. Introducción

DATOS:

Materia prima.

MODELO:

Implementación del
modelo de analítica.
Ajuste del modelo.

EVALUACION:

Viabilidad del negocio.
Validación criterios de
éxito.

DESPLIEGUE:

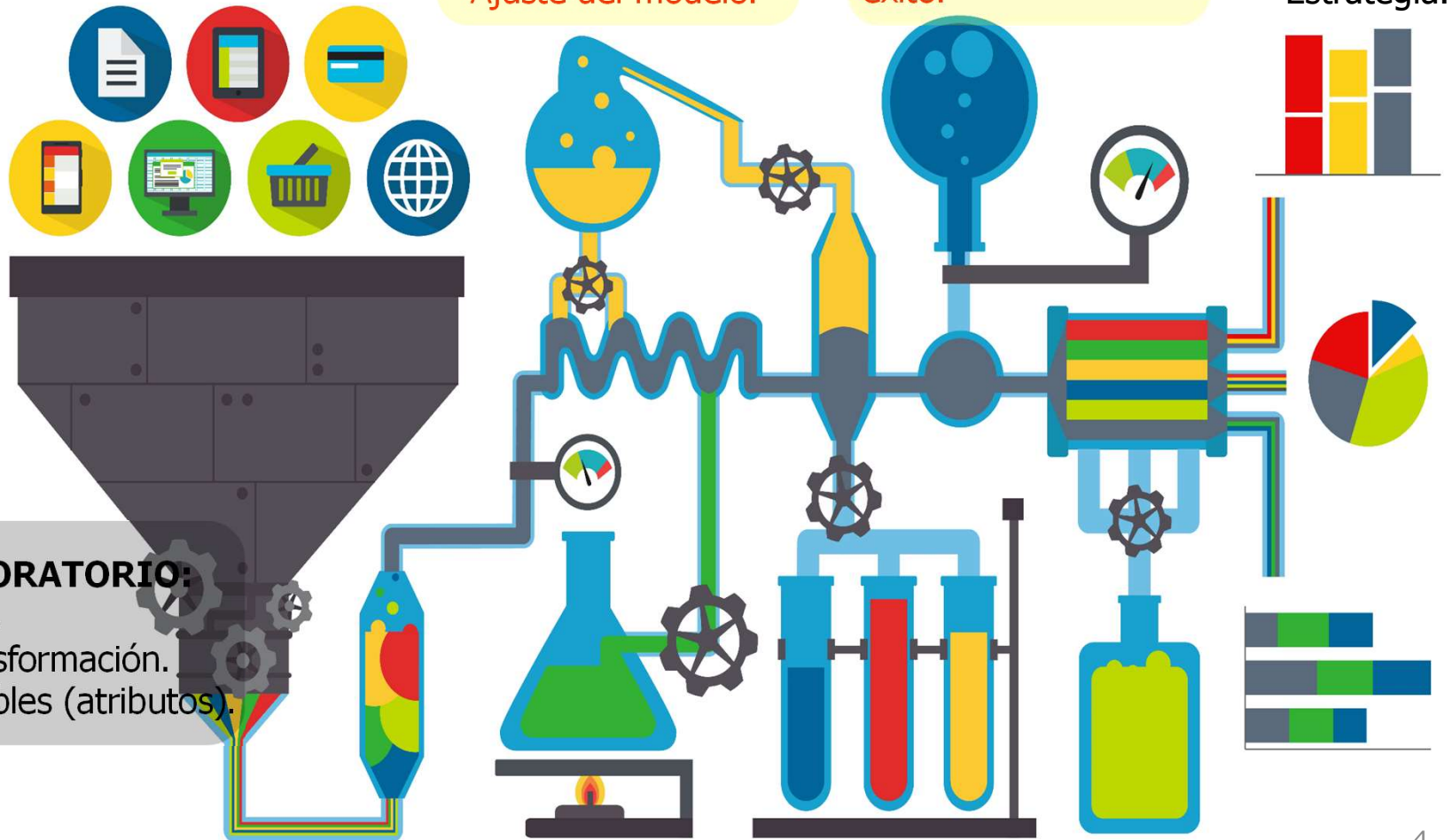
Resultados.
Conocimiento.
Estrategia.

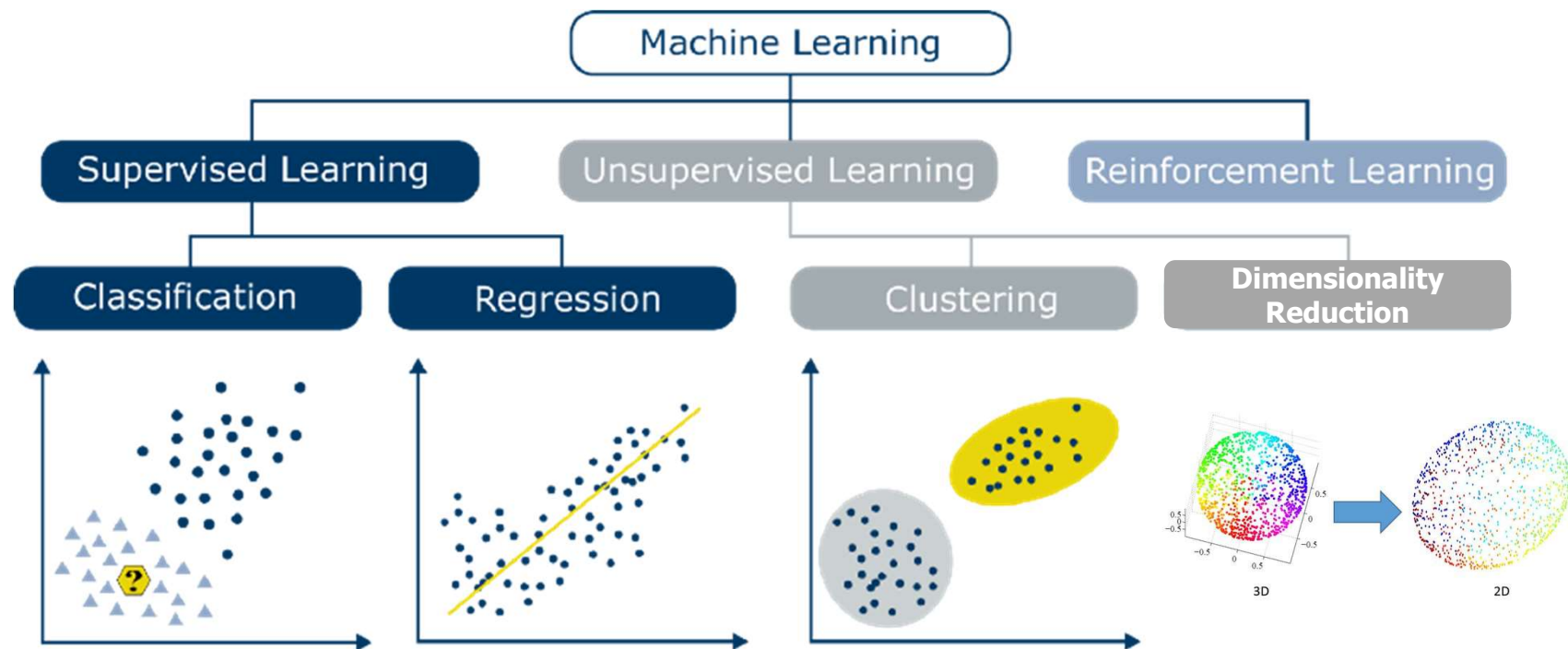
PREGUNTAS:

¿Cómo?
¿Cuáles?
¿Cuándo?
¿Por qué? *

ANALISIS EXPLORATORIO:

Limpieza de datos.
Preparación / transformación.
Selección de variables (atributos).





Métodos

KNN

Regresión Lineal Simple
Regresión Lineal Múltiple
Regresión Polinomial

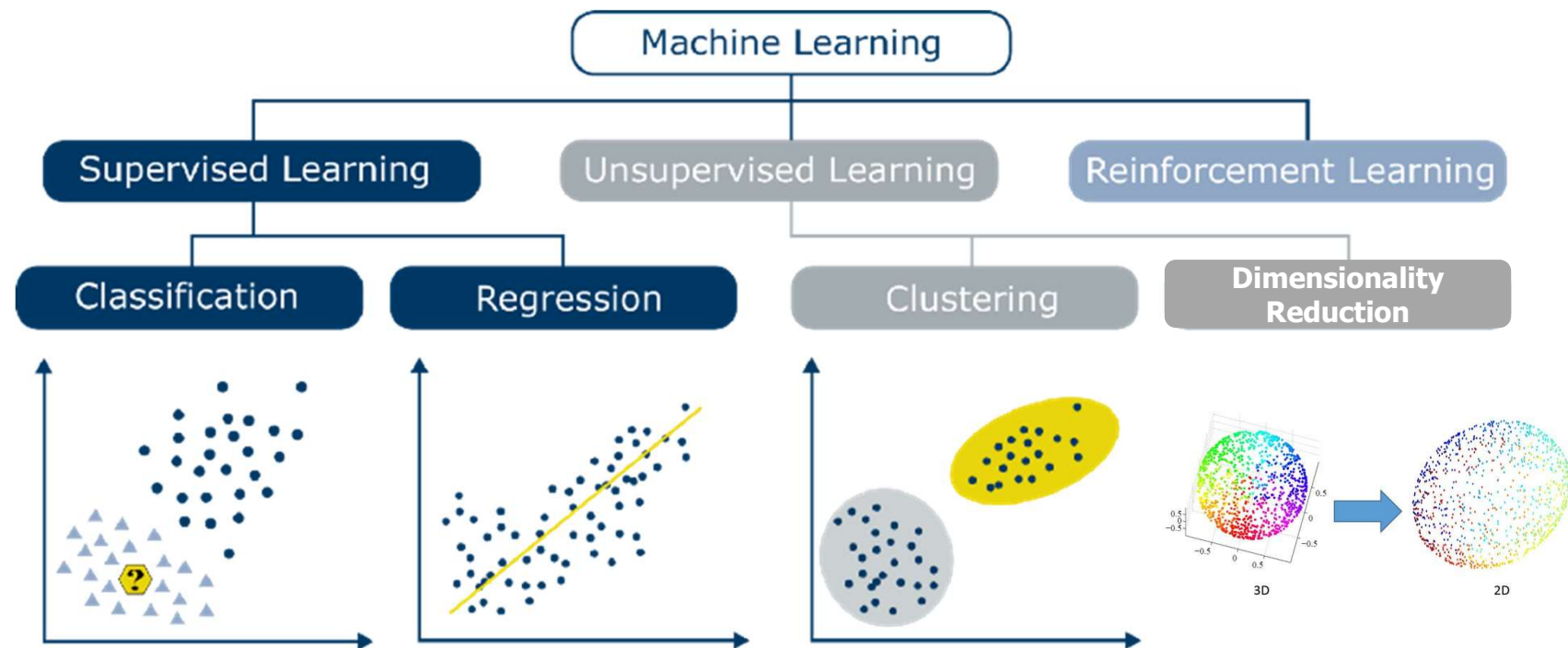
K-Means

Evaluación

Accuracy, Precision,
Recall, F1 score,
ROC, etc.

MSE, RMSE, R^2 , etc.

Método del codo
Método de la silueta
Calinski-Harabasz



Métodos

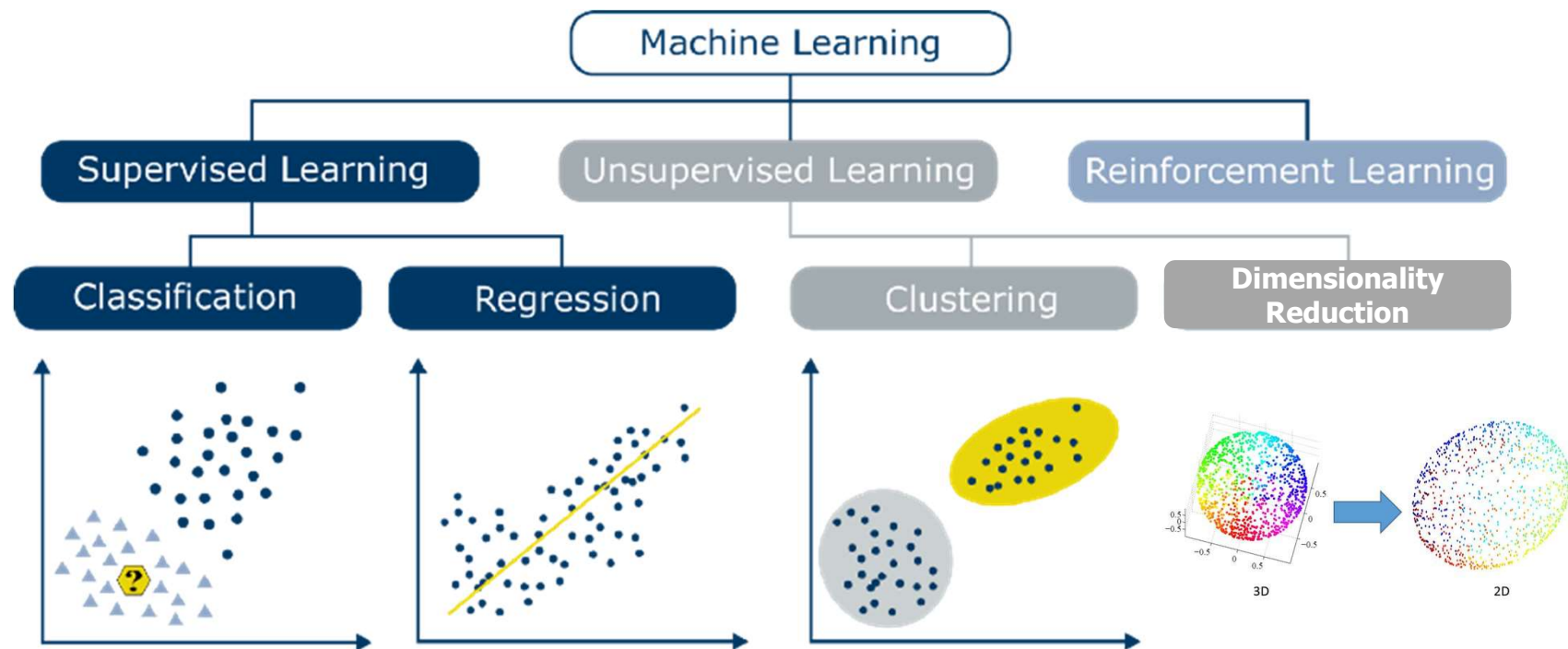
KNN

- Regresión logística
- Bayes ingenuo
- Árboles de decisión
- SVM
- Redes neuronales
- ...

Regresión Lineal Simple
Regresión Lineal Múltiple
Regresión Polinomial

- Regresión Ridge
- Regresión Lasso
- ...

K-Means



Métodos

KNN

- Regresión logística
- Bayes ingenuo
- Árboles de decisión
- SVM
- Redes neuronales
- ...

Regresión Lineal Simple
Regresión Lineal Múltiple
Regresión Polinomial

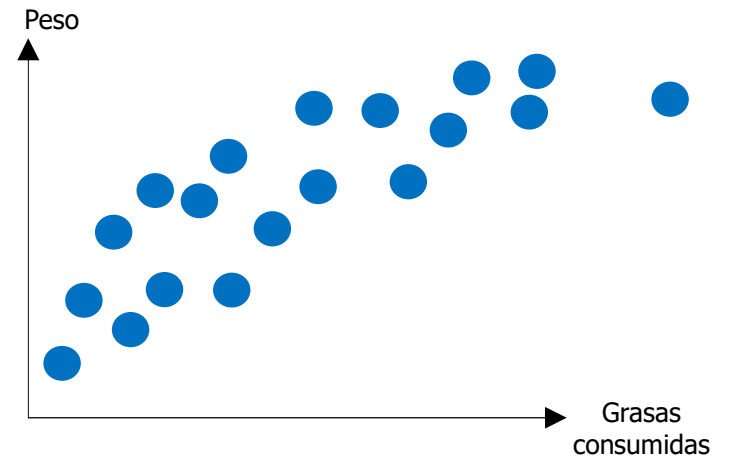
- Regresión Ridge
- Regresión Lasso
- ...

K-Means

2. Regresión logística

¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

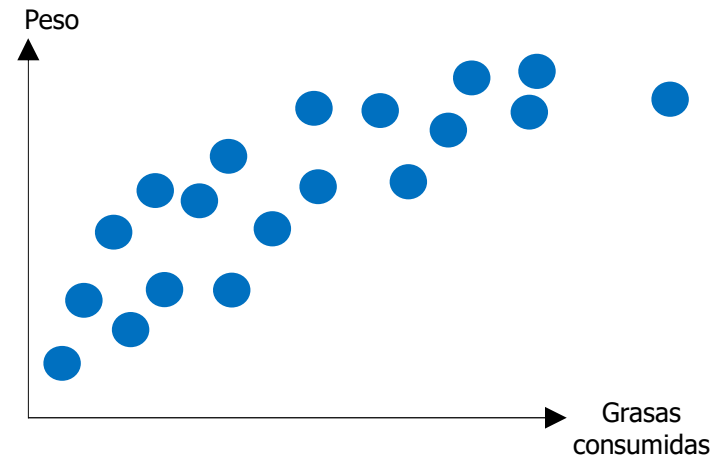


¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

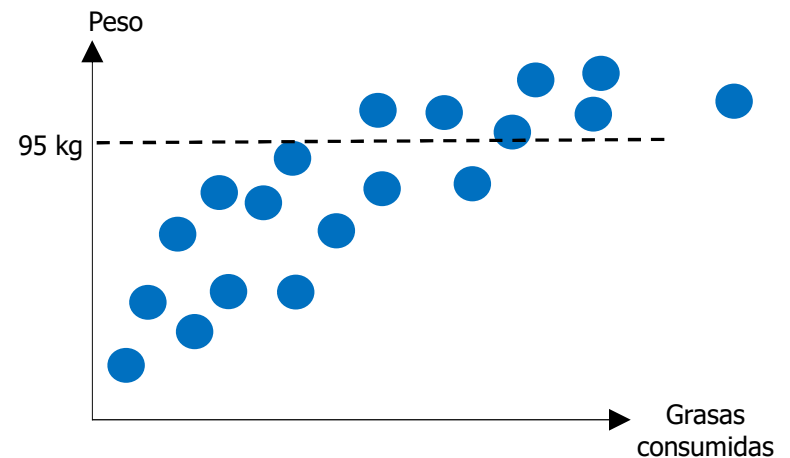


¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.



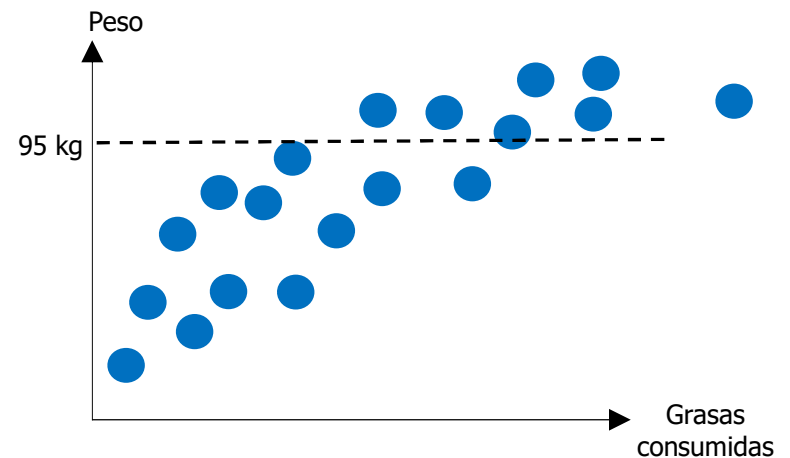
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.



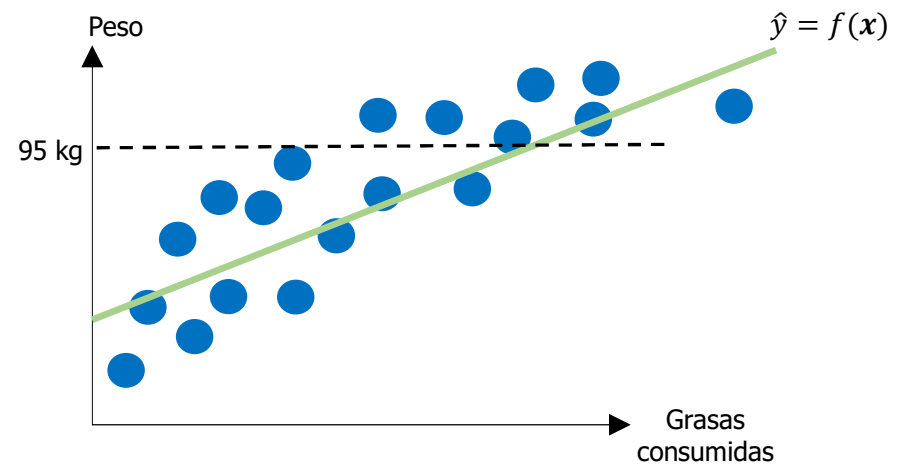
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.



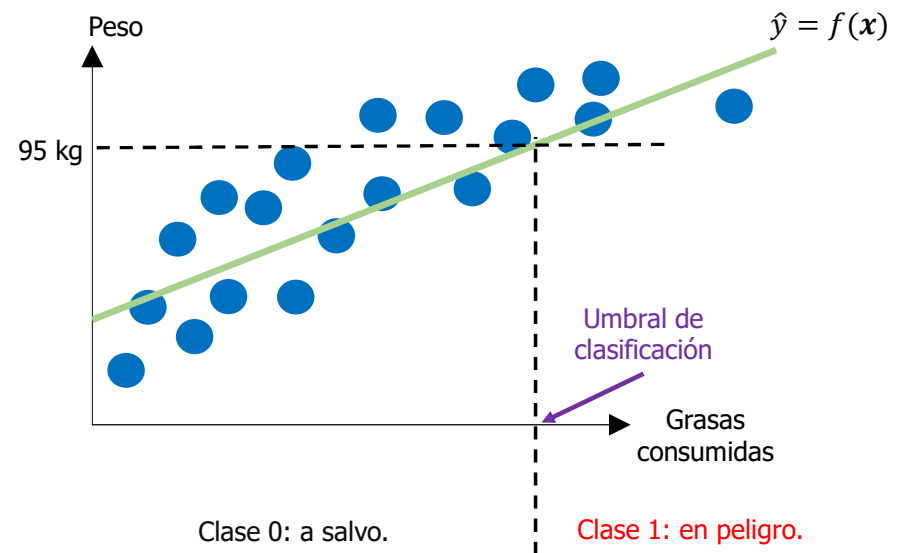
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.



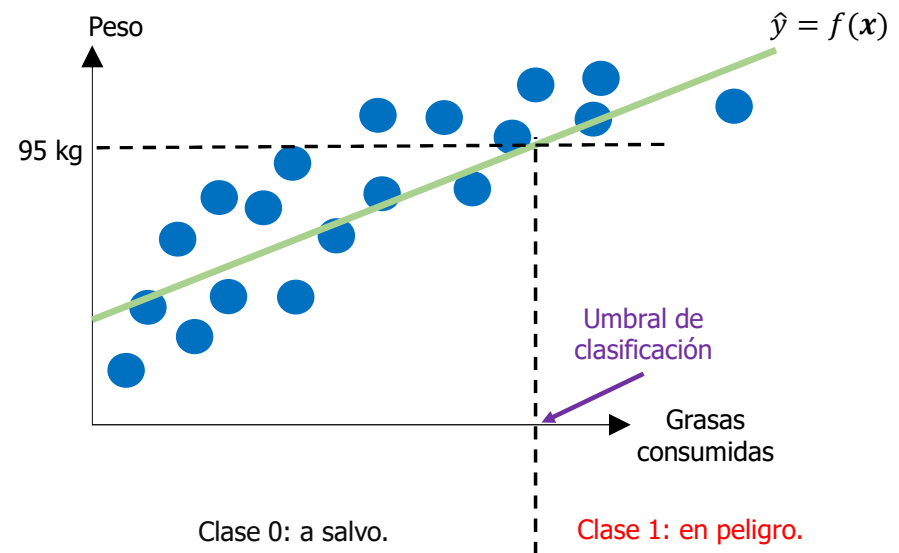
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.
- Pero no se pueden interpretar sus predicciones como probabilidades (valores no están entre $[0, 1]$).



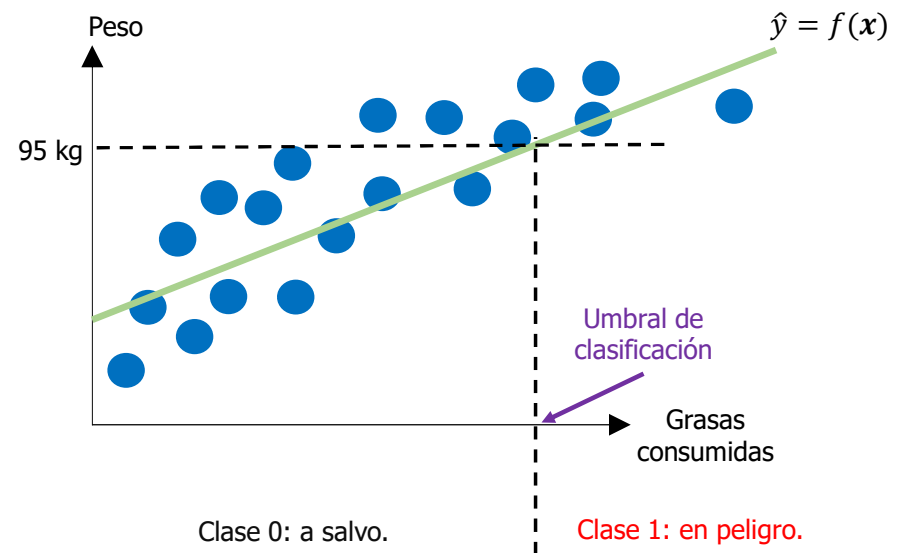
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.
- Pero no se pueden interpretar sus predicciones como probabilidades (valores no están entre $[0, 1]$).
- Además, podría no ser muy robusto...



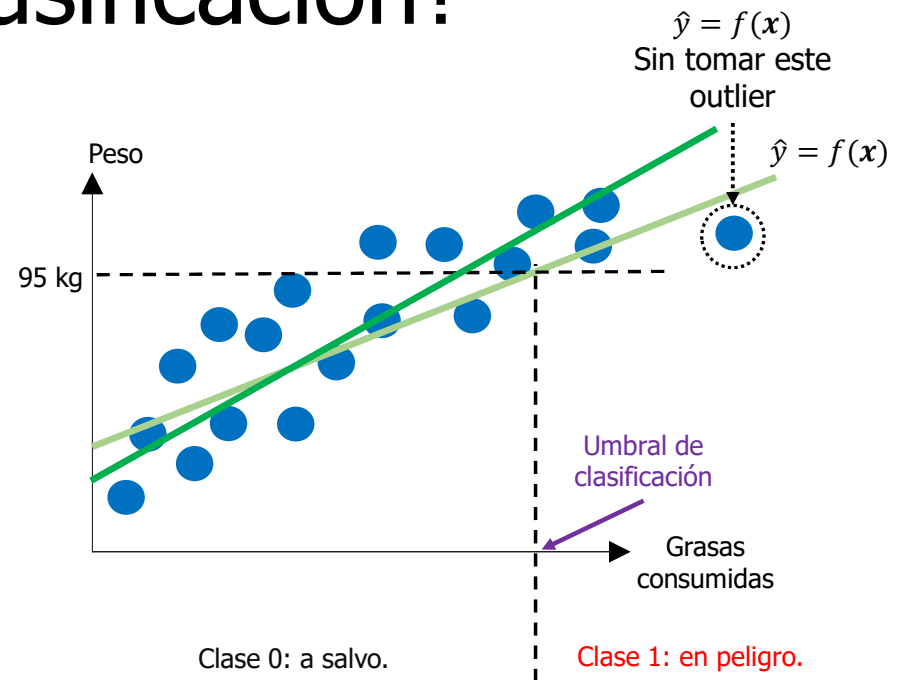
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.
- Pero no se pueden interpretar sus predicciones como probabilidades (valores no están entre $[0, 1]$).
- Además, podría no ser muy robusto...



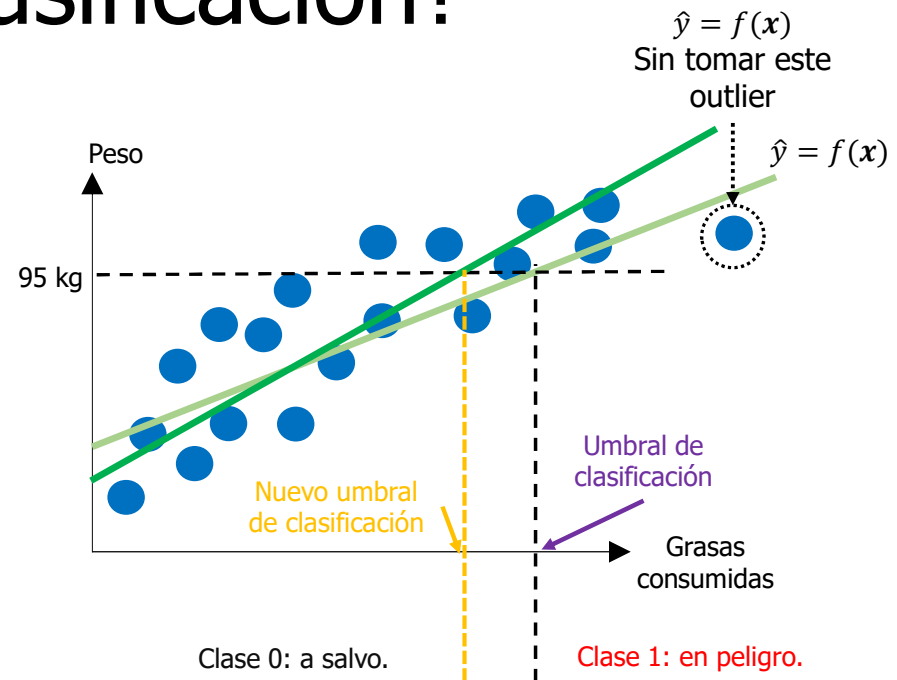
¿Regresión lineal para clasificación?

Ejemplo: tenemos datos que relacionan la cantidad de grasa consumida con la masa corporal de las personas, es decir un problema de regresión.

- Si un doctor estima que más de 95 kg implica riesgo de diabetes, ahora el problema se convierte en uno de clasificación.

0: a salvo, 1: en peligro.

- Una regresión lineal podría ayudar a estimar el límite sobre el cual se estaría en peligro de diabetes.
- Pero no se pueden interpretar sus predicciones como probabilidades (valores no están entre $[0, 1]$).
- Además, podría no ser muy robusto...



Regresión logística

- A pesar de su nombre, es un algoritmo de **clasificación**, no de **regresión!!!!**
- Parte de la idea de la regresión lineal, pero se modifica su resultado para obtener una salida continua **que varía entre 0 y 1**: sólo permite distinguir entre 2 clases:
 - Cliente que abandona vs. cliente que se queda.
 - Cliente que compra vs. cliente que no compra.
 - Cliente valioso vs. cliente no valioso.
 - Paciente con diabetes o no diabetes.
 - ...
- Se agrega una transformación al resultado de la regresión lineal (z) a partir de una función logística, también conocida como función **logit** o **sigmoide**.

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

Regresión logística

- El modelo pasa de:

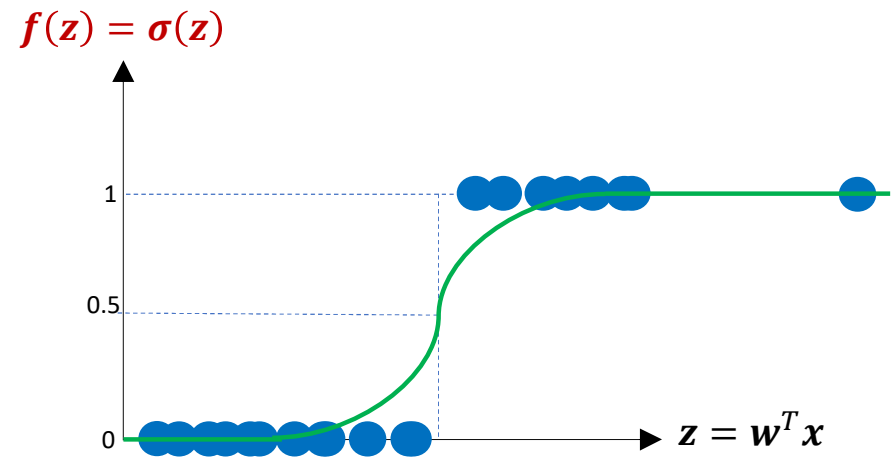
$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1x_1 + \dots + w_nx_n$$

a

$$f(\mathbf{z}) = \sigma(\mathbf{z}) = \sigma(\mathbf{w}^T \mathbf{x}) = \sigma(w_0 + w_1x_1 + \dots + w_nx_n),$$

Donde:

- $\mathbf{z} = \mathbf{w}^T \mathbf{x} = w_0 + w_1x_1 + \dots + w_nx_n$
- $\sigma(\mathbf{z})$ es la función **sigmoide** o **logística**.
- $\max(\sigma(\mathbf{z})) = 1$ y $\min(\sigma(\mathbf{z})) = 0$



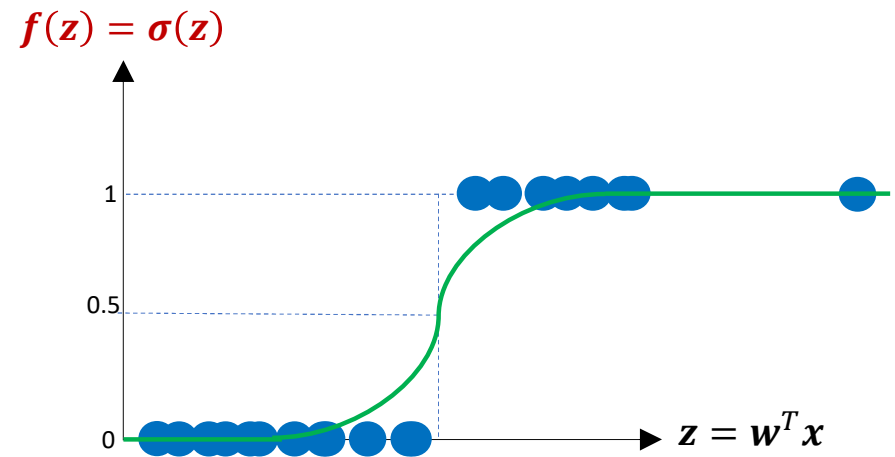
$$f(\mathbf{z}) = \sigma(\mathbf{z}) = \frac{1}{1 + e^{-\mathbf{z}}}$$

Regresión logística

- Se pueden interpretar los valores de $\sigma(\mathbf{z})$ como **probabilidades** que una instancia con atributos \mathbf{x} pertenezca a la clase $Y=1$:

$$P(Y = 1 | x_1, \dots, x_n) = p_1(\mathbf{x}) = \sigma(w_0 + w_1 x_1 + \dots + w_n x_n)$$

$$f(\mathbf{z}) = \sigma(\mathbf{z}) = \frac{1}{1 + e^{-z}} = p_1(\mathbf{x}) = \frac{1}{1 + e^{-w^T \mathbf{x}}}$$



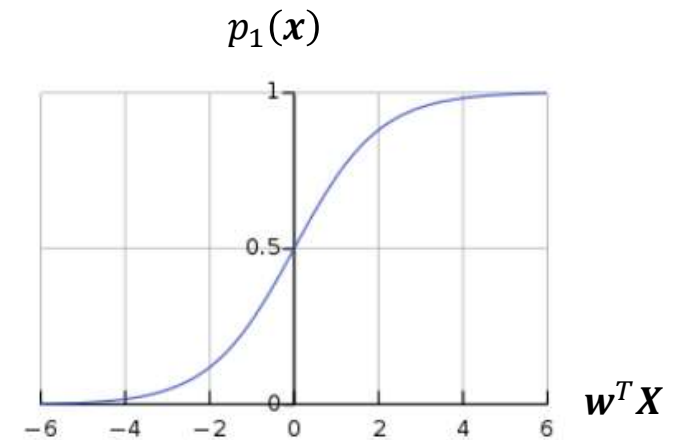
$$p_1(\mathbf{x}) = f(\mathbf{z}) = \sigma(\mathbf{z}) = \frac{1}{1 + e^{-z}}$$

Regresión logística

- Comportamiento:

- ✓ Si $y=1$, queremos que $p_1(x) \approx 1$ y por lo tanto $w^T X \gg 0$

- ✓ Si $y=0$, queremos que $p_1(x) \approx 0$ y por lo tanto $w^T X \ll 0$

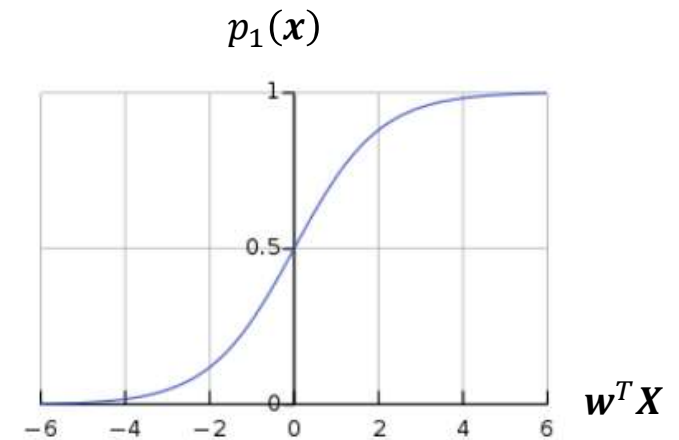


Fuente: Wikipedia.

$$p_1(x) = \frac{1}{1 + e^{-w^T x}}$$

Regresión logística

- Comportamiento:
 - ✓ Si $y=1$, queremos que $p_1(x) \approx 1$ y por lo tanto $w^T X \gg 0$
 - ✓ Si $y=0$, queremos que $p_1(x) \approx 0$ y por lo tanto $w^T X \ll 0$
- Para hacer **Predicción**, se debe establecer un valor de **umbral**.



Fuente: Wikipedia.

$$p_1(x) = \frac{1}{1 + e^{-w^T x}}$$

Regresión logística

- Comportamiento:

- ✓ Si $y=1$, queremos que $p_1(x) \approx 1$ y por lo tanto $w^T X \gg 0$

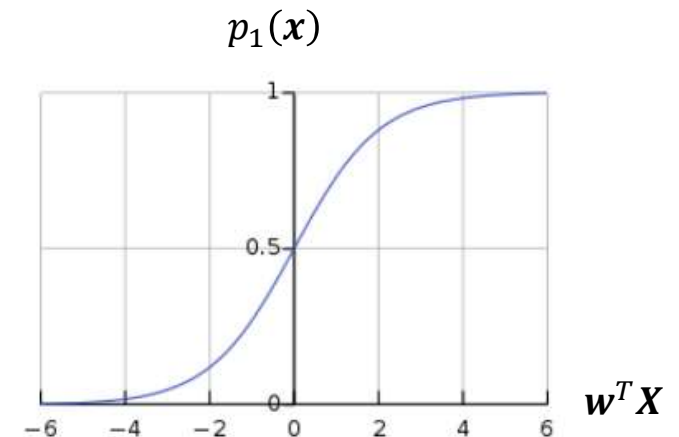
- ✓ Si $y=0$, queremos que $p_1(x) \approx 0$ y por lo tanto $w^T X \ll 0$

- Para hacer **Predicción**, se debe establecer un valor de **umbral**.

- Por ejemplo, ¿qué sucede si el umbral se fija por en 0.5?

- ✓ Si $p_1(x) \geq 0.5$, se predice clase 1.

- ✓ Si $p_1(x) < 0.5$, se predice clase 0.



Fuente: Wikipedia.

$$p_1(x) = \frac{1}{1 + e^{-w^T x}}$$

Regresión logística

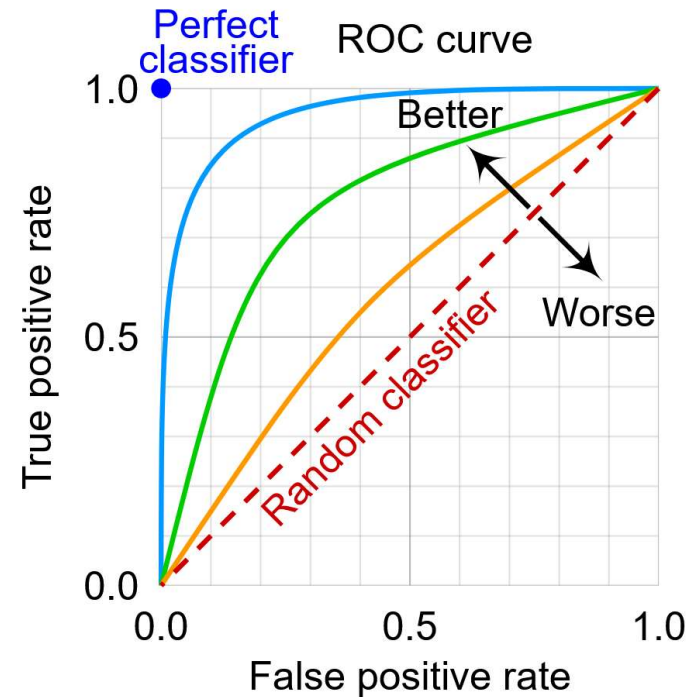
- ¿Qué impacto tiene este **umbral** en los falsos positivos y falsos negativos?

Regresión logística

- ¿Qué impacto tiene este **umbral** en los falsos positivos y falsos negativos?
- Una herramienta que se tiene para determinar el impacto de este umbral es la curva ROC (Receiver Operating Characteristic).

Regresión logística

- ¿Qué impacto tiene este **umbral** en los falsos positivos y falsos negativos?
- Una herramienta que se tiene para determinar el impacto de este umbral es la curva ROC (Receiver Operating Characteristic).



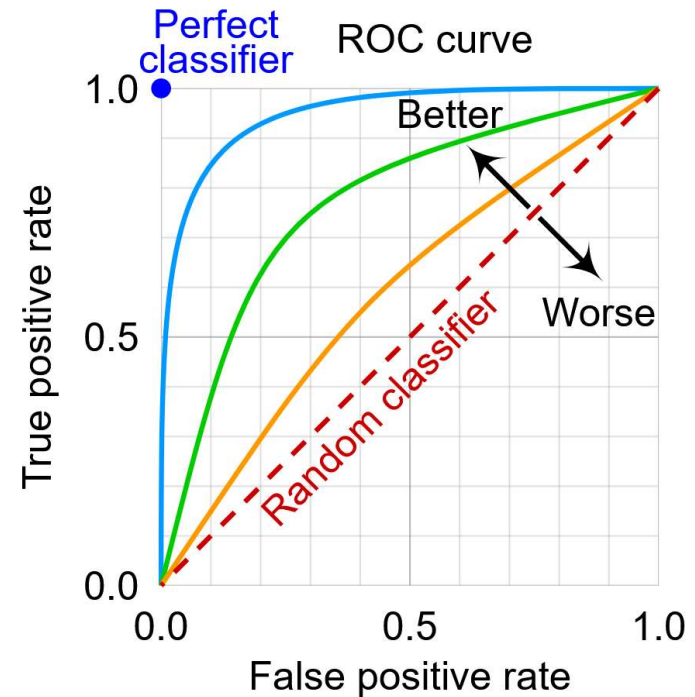
		Clase que se predijo (predicted class)	
		1	0
Clase verdadera (actual class)	1	Verdaderos positivos (true positive)	Falsos negativos (false negative)
	0	Falsos positivos (false positive)	Verdaderos negativos (true negative)

$$\text{TPR} = \text{sensitivity} = \text{recall} = \frac{TP}{TP+FN}$$

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{FP+TN}$$

Regresión logística

- ¿Qué impacto tiene este **umbral** en los falsos positivos y falsos negativos?
- Una herramienta que se tiene para determinar el impacto de este umbral es la curva ROC (Receiver Operating Characteristic).
- A partir de la curva ROC se puede calcular el AUC (Area Under the Curve) y usarlo para comparar diferentes modelos.



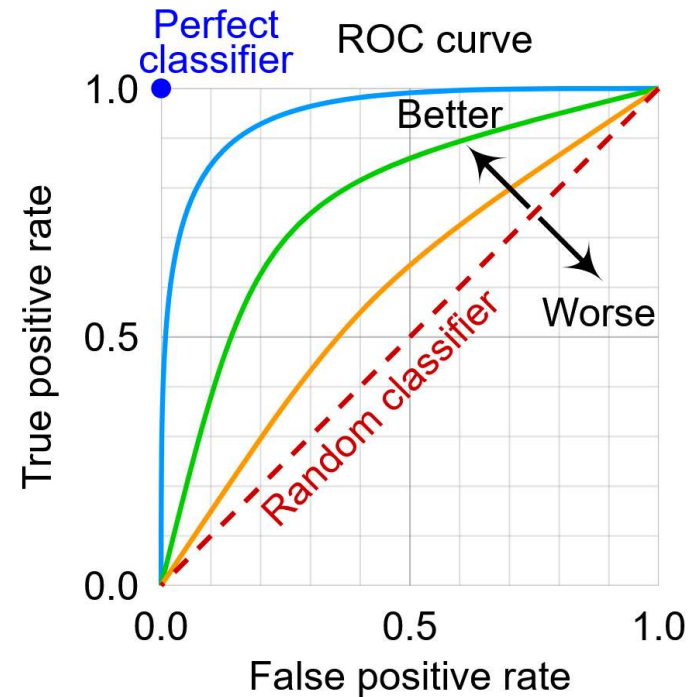
		Clase que se predijo (predicted class)	
		1	0
Clase verdadera (actual class)	1	Verdaderos positivos (true positive)	Falsos negativos (false negative)
	0	Falsos positivos (false positive)	Verdaderos negativos (true negative)

$$\text{TPR} = \text{sensitivity} = \text{recall} = \frac{TP}{TP+FN}$$

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{FP+TN}$$

Regresión logística

- ¿Qué impacto tiene este **umbral** en los falsos positivos y falsos negativos?
- Una herramienta que se tiene para determinar el impacto de este umbral es la curva ROC (Receiver Operating Characteristic).
- A partir de la curva ROC se puede calcular el AUC (Area Under the Curve) y usarlo para comparar diferentes modelos.
- ***El umbral se escoge dependiendo del contexto de la clasificación.***



		Clase que se predijo (predicted class)	
		1	0
Clase verdadera (actual class)	1	Verdaderos positivos (true positive)	Falsos negativos (false negative)
	0	Falsos positivos (false positive)	Verdaderos negativos (true negative)

$$\text{TPR} = \text{sensitivity} = \text{recall} = \frac{TP}{TP+FN}$$

$$\text{FPR} = 1 - \text{specificity} = \frac{FP}{FP+TN}$$

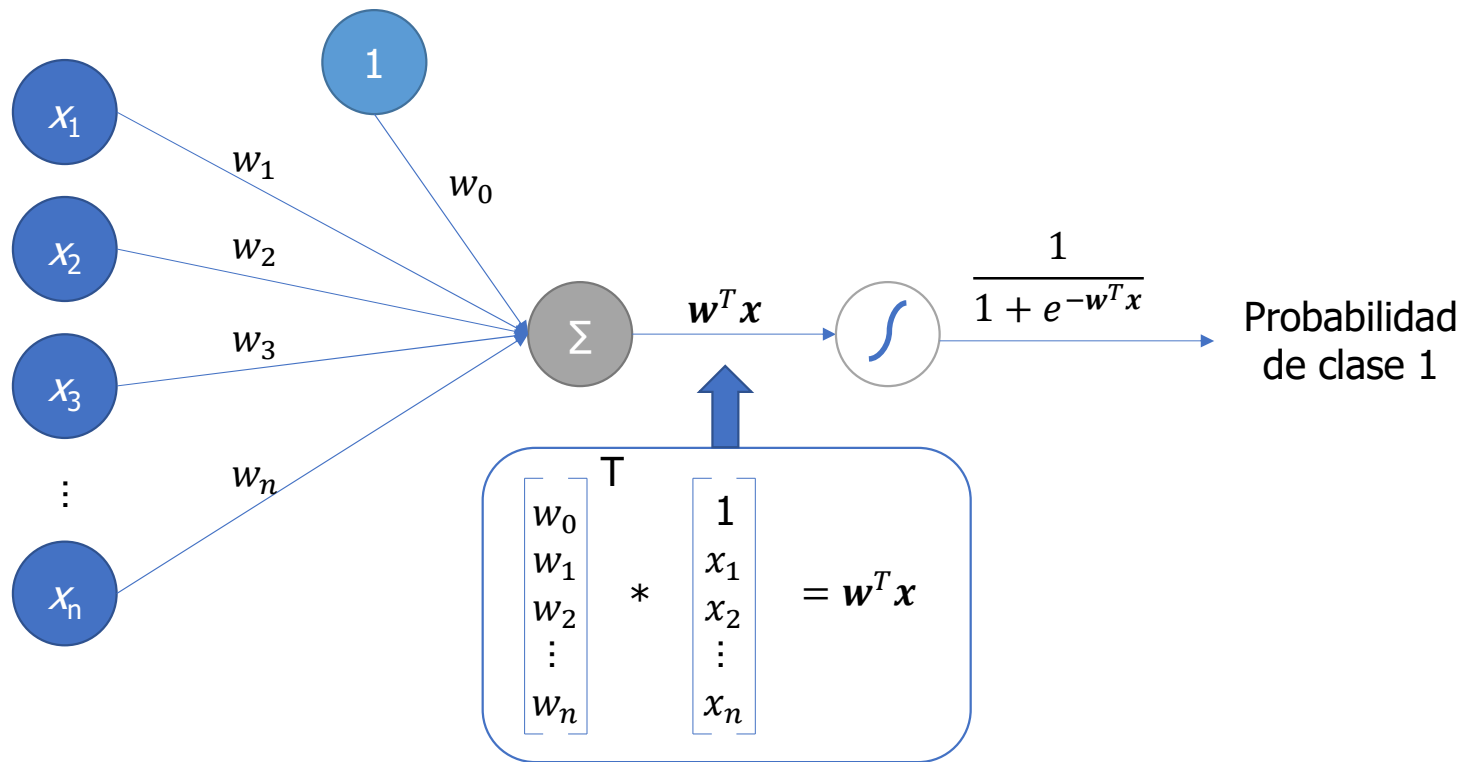
Regresión logística

- Coeficientes w_i :
 - $\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = w_0 + w_1x_1 + \dots + w_nx_n$
 - Relación **lineal** entre los coeficientes y el logaritmo de la razón de probabilidades (*odds ratio*).
 - Un crecimiento de una unidad de x_1 indica que el log de la razón de probabilidades va a crecer w_1 unidades.
 - El **signo** indica la dirección de la influencia.
 - Análisis de sensibilidad de $p(y=1)$ con respecto a una variable, fijando las otras en sus valores promedios.
 - Prueba de hipótesis para evaluar la **significancia** de cada coeficiente (diferencia de 0).

- Razón de probabilidades
 - A probabilidades altas, razón alta y viceversa

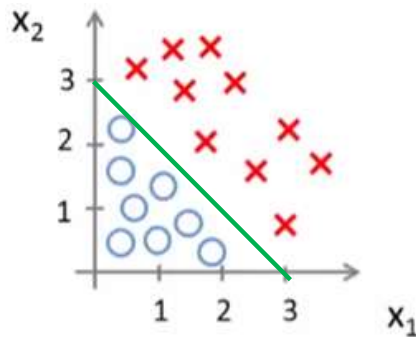
$p_1(X)$	odds
1,0	+Inf
0,99	99
0,75	3
0,5	1
0,25	0,33
0	0

Regresión logística



Regresión logística

- El algoritmo de regresión logística determina una **frontera de decisión lineal**.



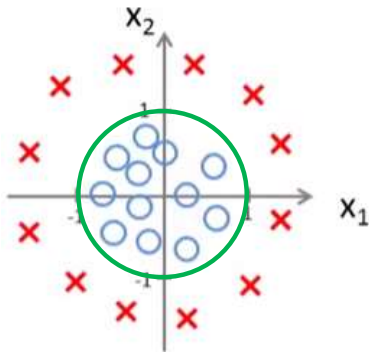
$$h_w(x) = f(-3 + x_1 + x_2)$$

Predecir la clase roja de cruz cuando:

$$h_w(x) \geq 0.5$$

$$f(-3 + x_1 + x_2) \geq 0.5$$

- Para **fronteras de decisión no lineales**: usar polinomios de un mayor orden.



$$h_w(x) = f(-1 + x_1^2 + x_2^2)$$

Predecir la clase roja de cruz cuando:

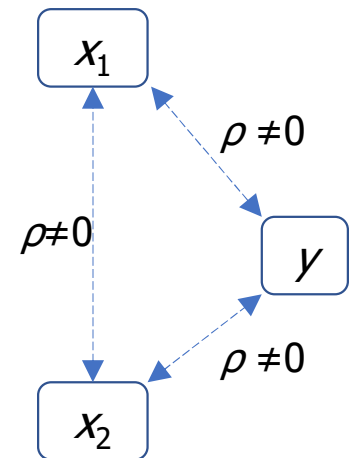
$$h_w(x) \geq 0.5$$

$$f(-1 + x_1^2 + x_2^2) \geq 0.5$$

Regresión logística

Variables de confusión (confounding variables)

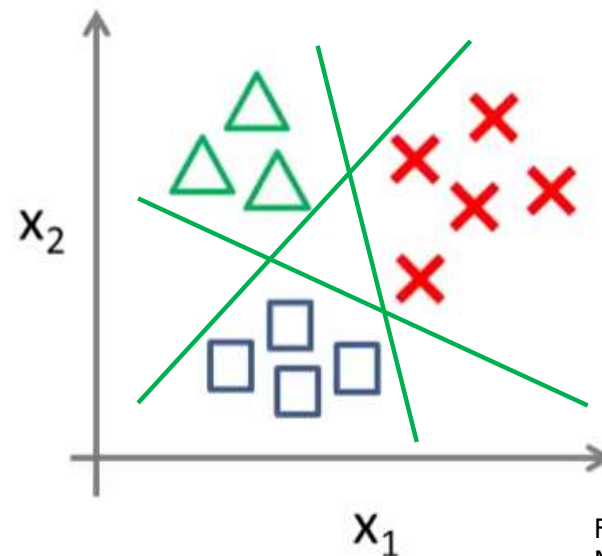
- Problema que ocurre cuando un modelo de regresión no considera variables independientes relevantes.
- Posibles efectos en la relación entre variables independientes y dependiente:
 - Sobrestimar / subestimar la fortaleza de una relación.
 - Cambiar la dirección de una relación.
 - Esconder un efecto que en realidad existe.
- Causas
 - La variable que se omite está correlacionada con la variable dependiente.
 - La variable que se omite está correlacionada con al menos una de las variables independientes del modelo.



Regresión logística

¿Qué se puede hacer si se tienen más de 2 clases?

- Para problemas de clasificación con más de dos clases, se puede utilizar la aproximación de **uno contra todos**.
- Se requiere un clasificador de regresión logística para cada clase.
- Para una nueva instancia, la clase con la mayor probabilidad en su propio modelo es la que se retorna.




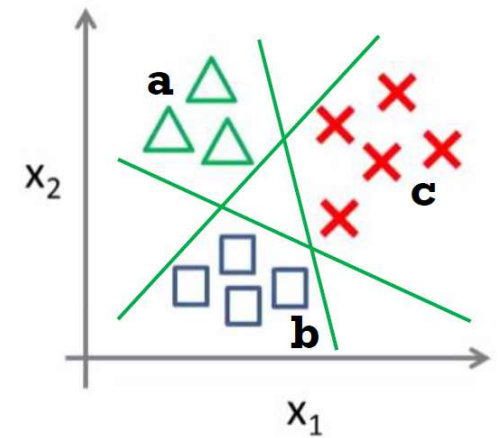
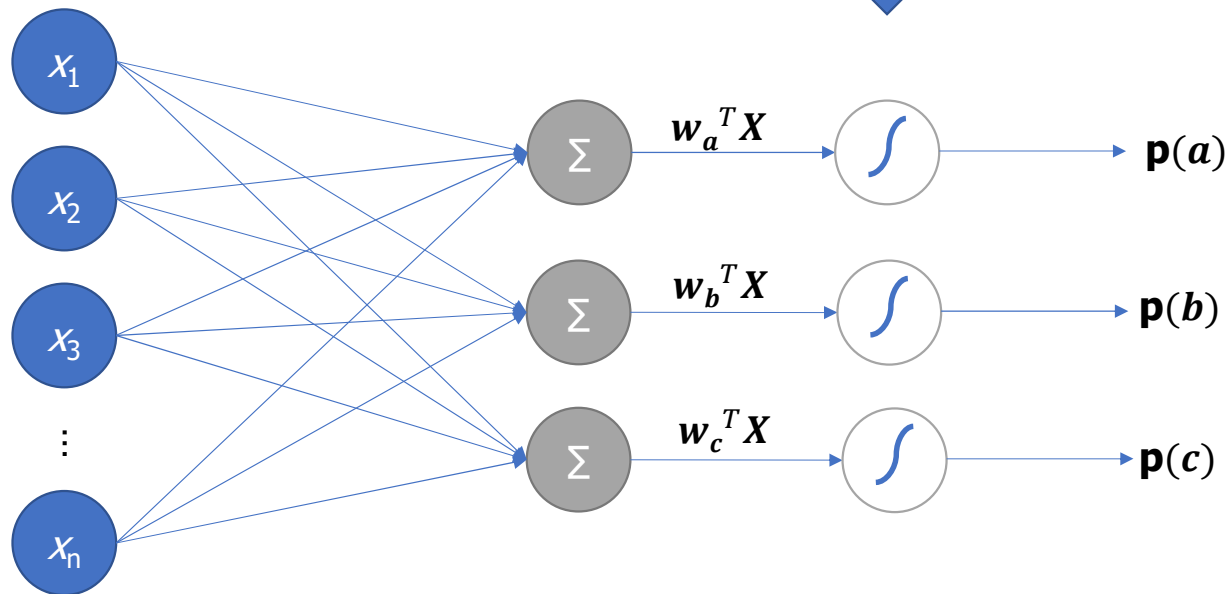
Fuente: Andrew Ng.

También se puede hacer regresión logística **multinomial** con la función **softmax**.

Softmax

Softmax

$$\frac{e^{w^T X^{(i)}}}{\sum_j e^{w^T X^{(j)}}}$$




Softmax

- La función softmax permite tomar un conjunto de medidas (*scores*) de clasificación y convertirlos en probabilidades.

$$s(y^{(i)}) = \frac{e^{y^{(i)}}}{\sum_j e^{y^{(j)}}} = \frac{e^{\mathbf{w}^T \mathbf{X}^{(i)}}}{\sum_j e^{\mathbf{w}^T \mathbf{X}^{(j)}}}$$

- Se puede crear un clasificador logístico basado en softmax y no en la función sigmoide, entrenando los parámetros de una combinación lineal de predictores usando gradiente descendente, a partir de una función de costo.
- La función sigmoide calcula una sola salida, mientras que la softmax calcula múltiples valores intermedios que después se normalizan.
- Se trata de una generalización de la función sigmoide para más de dos clases (en el taller veremos que la función sigmoide es un caso particular de la función softmax).

Softmax

- Noten en los siguientes ejemplos el efecto no proporcional de las diferencias entre los resultados dadas las magnitudes de los puntajes de entrada a la transformación del Softmax:

```
softmax([3.0, 1.0, 0.2]) = [0.8360188 0.11314284 0.05083836]
softmax([3.0, 2.9, 2.8]) = [0.3671654 0.33222499 0.30060961]
softmax([3.0, 0.2, 0.1]) = [0.89619123 0.05449744 0.04931133]
softmax([3.0, 0.02, 0.01]) = [0.908199 0.04613 0.045671]
softmax([3.0, 0.0002, 0.0001]) = [0.90943064 0.04528694 0.04528241]
softmax([3.0, 0.00000002, 0.00000001]) = [0.909443 0.0452785 0.0452785]
softmax([30.0, 0.00000002, 0.00000001]) = [1.00000000e+00 9.35762316e-14 9.35762306e-14]
```

Regresión logística

- Consideraciones:
 - Produce estimación de “probabilidades”.
 - No hay parámetros a afinar, solo las variables independientes a considerar.
 - Permite usar variables independientes, numéricas y categóricas.
 - La estimación de los parámetros es eficiente computacionalmente.
 - No se ve afectado por situaciones de multicolinealidad leves. Casos importantes se pueden resolver con una regularización L_2 (Ridge).
 - Se puede utilizar gradiente descendente para encontrar los parámetros.
 - No es ideal en casos donde se tienen muchas variables categóricas.
 - No es muy flexible (lineal) aunque se puede extender usando ingeniería de atributos como vimos cuando creamos atributos polinomiales.

Regresión logística

- Consideraciones:
 - Las variables predictivas deben ser numéricas.
 - Las variables categóricas debes ser convertidas a variables numéricas:
 - Codificación *uno de n* (*one-hot encoding*): se crea una variable para cada valor posible de cada variable categórica.
 - Variables de contraste o "*dummy*": variables numéricas adicionales para cada valor posible. Se acostumbra a crear una menos que el total de categorías para evitar problemas de multicolinealidad en modelos estadísticos.

Ejemplo: suponga una variable de estrato con tres valores posibles (bajo, medio, y alto).

	Estrato_bajo	Estrato_medio
Valor = bajo	1	0
Valor = medio	0	1
Valor = alto	0	0

Regresión logística

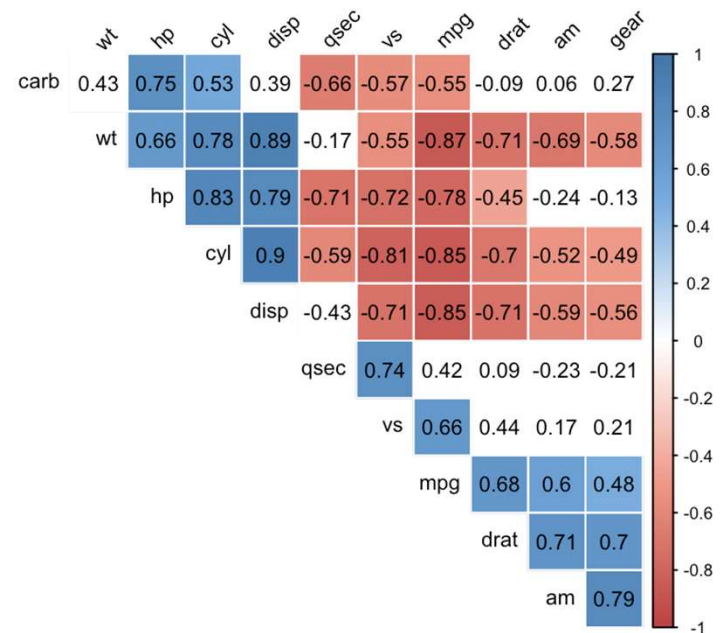
- Consideraciones:
- Supuestos estadísticos para la utilización de una regresión lineal múltiple en:

$$f(\mathbf{z}) = \sigma(\mathbf{z}) = \sigma(\mathbf{w}^T \mathbf{x}) = \sigma(w_0 + w_1 x_1 + \dots + w_n x_n)$$

✓ **Variables predictoras linealmente independientes entre ellas.**

✓ Evitar el problema de la **multicolinealidad**.

- Para validar:
 - ✓ En la matriz de correlación filtrar variables con correlaciones altas (e.g.: $\rho > |0.85|$).



Regresión logística

- Consideraciones:

Si existen varias variables independientes, se puede definir el modelo de regresión múltiple a utilizar, dada una medida de calidad del ajuste:

- **Completo:** se evalúan todas las posibles combinaciones de variables independientes, y se escoge la mejor.
- **Tamaño fijo:** se evalúan todas las posibles combinaciones de K variables independientes y se escoge la mejor.
- **Paso a paso (*stepwise*)**
 - ✓ Hacia adelante (***forward***): se prueba una a una con las variables independientes que aún no se escogen y se evalúa el modelo conjuntamente con las variables seleccionadas previamente. Se detiene cuando la medida de la calidad del ajuste no mejore.
 - ✓ Hacia atrás (***backward***): sigue un proceso contrario al método de búsqueda “hacia adelante”, en este caso, se empieza con todas las variables y se va eliminando la variable que, cuando no se considera, optimiza la medida de calidad del ajuste.
- **PCA:** se transforman los datos a un nuevo espacio vectorial de menor dimensionalidad que el de entrada.

Lecturas Complementarias

- Logistic Regression
(<https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>)
- What is Logistic Regression
(<https://aws.amazon.com/what-is/logistic-regression/>)