



EGADE Business School
Tecnológico de Monterrey

Aplicaciones de analítica de datos a los negocios II

PROF: JUAN C. BUSTAMANTE

JUCBUSTAM@TEC.MX

Normas para la conexión síncrona:

1. La clases tiene un back-up garantizado (Grabación disponible en la nube de Zoom).
2. Ingresar a la clase con la cámara del equipo de computo encendida.
3. La **cámara deberá permanecer encendida a lo largo de la clase.**
4. Al ingresar a la clase deben silenciar el micrófono del equipo de computo.
5. Levantar la mano es una opción cuando se quiere preguntar algo durante la sesión de clase, pero les recomiendo que mejor hagamos uso intensivo del chat del canal general para hacer preguntas.
6. En caso de necesitar hacer una pregunta, puede interrumpir la clase sin problema, activando el micrófono de vuestro equipo de computo, luego de la pregunta desactíVELO nuevamente.
7. Para una buena clase online es indispensable **el debate, así que foméntelo!!!**.
8. Toda la información se gestiona en CANVAS LMS.



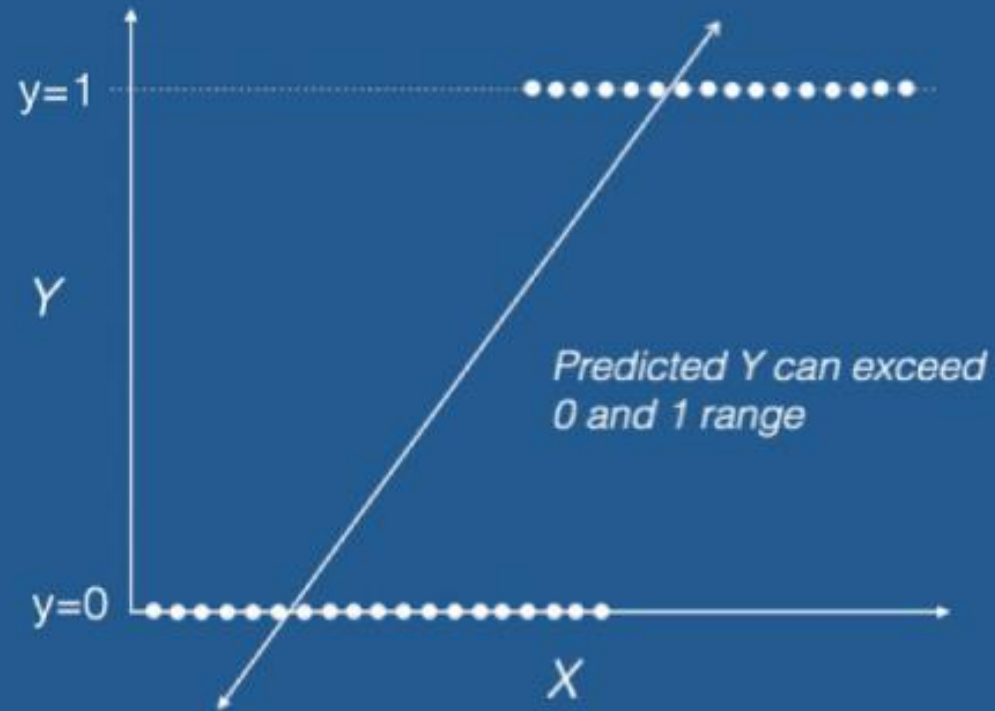
Cronograma de trabajo:

Sesiones	Contenidos	Actividad		Fecha
1	Información general del curso	Utility of classification algorithms		Martes 18/04
2	Algoritmo de regresión logística	Ejecutar script	Solución caso: Retention modelling at Scholastic Travel Company (A) and (B)	Martes 25/04
3	Algoritmo Naïve Bayes	Ejecutar script		Martes 02/05
4	Algoritmo k-nearest-neighbors (KNN)	Ejecutar script		Martes 09/05
5	Algoritmo Support vector machine	Ejecutar script		Martes 16/05
6	Algoritmo Decision Trees	Ejecutar script		Martes 23/05
7	Algoritmo Random Forest	Ejecutar script		Martes 30/05
8	Modelo RFM	Ejecutar script	Solución caso: CD Now	Martes 06/06
9	Modelo valor de vida del cliente (I)	Ejecutar script		Martes 13/06
10	Modelo valor de vida del cliente (II)	Ejecutar script		Martes 20/06
11	Análisis de series de tiempo	Ejecutar script		Martes 27/06
12	Proyecto final	Presentación en equipos		Martes 04/07
	Evaluación final			

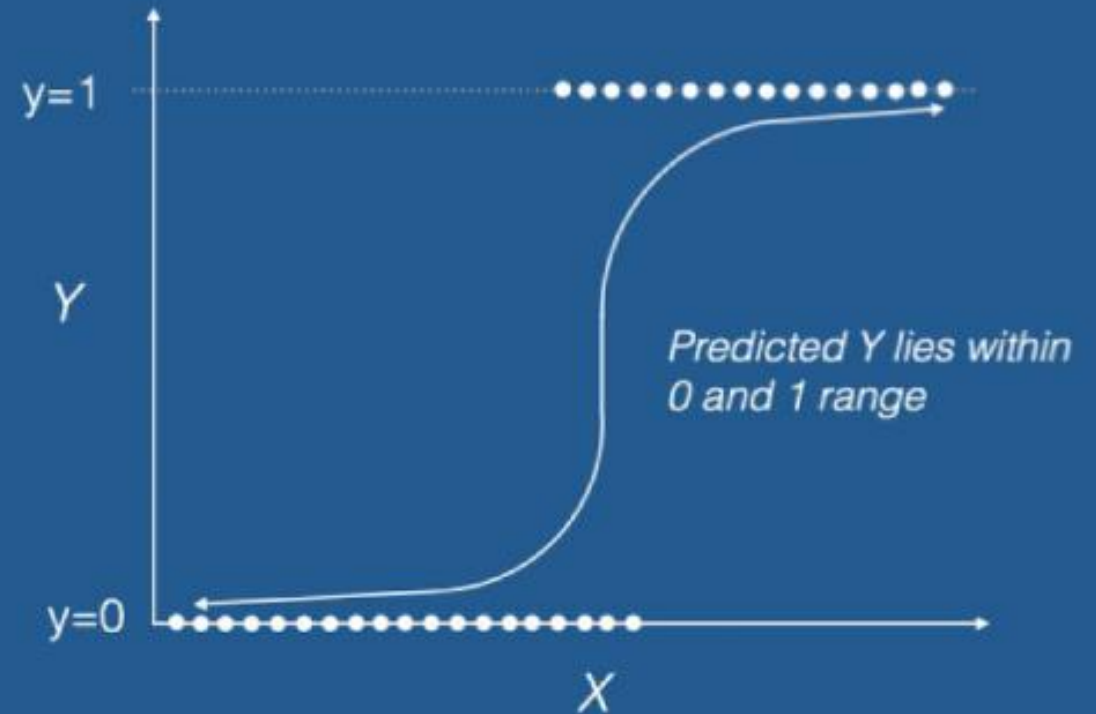
1

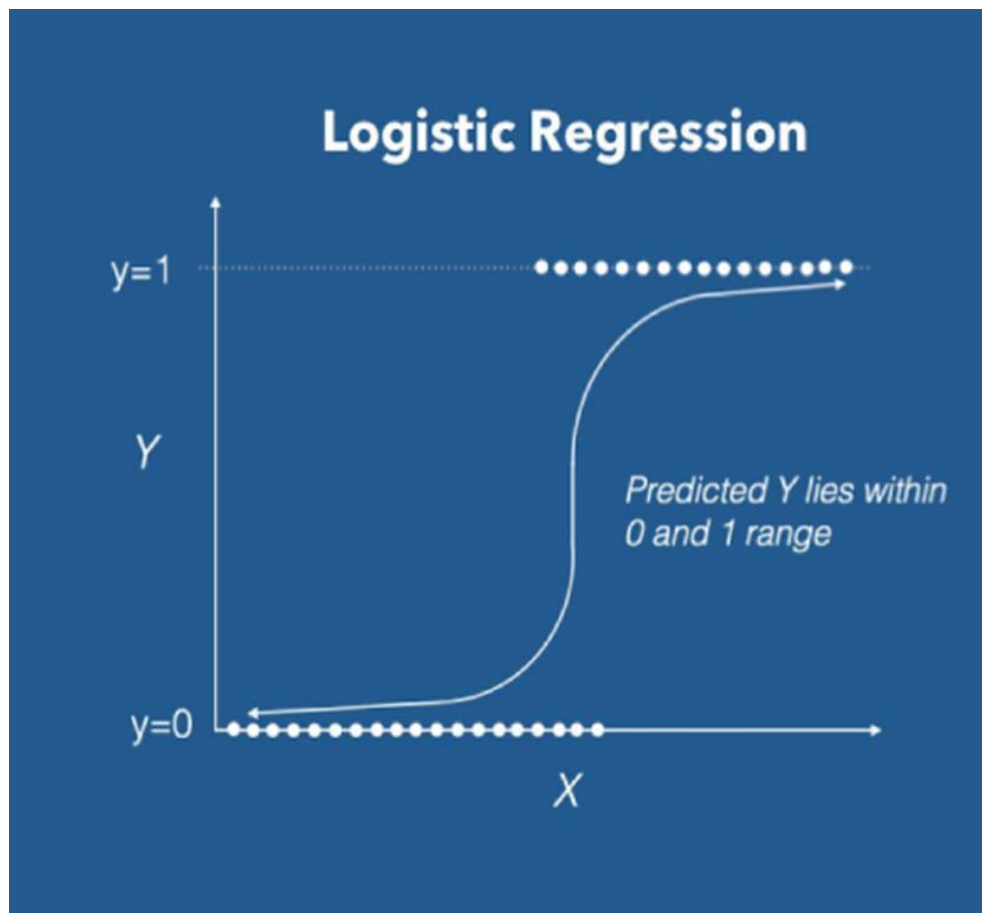
< **Introducción a la
regresión logística.**

Linear Regression



Logistic Regression





- La Regresión Logística Simple, es un método de regresión que permite estimar la **probabilidad** de una variable cualitativa binaria en función de una variable cuantitativa.
- Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.
- Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

¿Por qué regresión logística y no lineal?

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 x$.

El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1]

Para evitar estos problemas, la regresión logística transforma el valor devuelto por la regresión lineal ($\beta_0 + \beta_1 X$) empleando una función cuyo resultado está siempre comprendido entre 0 y 1.

Existen varias funciones que cumplen esta descripción, una de las más utilizadas es la función logística (también conocida como función sigmoide):

$$\text{función sigmoide} = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Sustituyendo la x de la ecuación 1 por la función lineal ($\beta_0 + \beta_1 X$) se obtiene que:

$$P(Y = k | X = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Consideraciones sobre datos de regresión logística

1

Datos. La variable dependiente debe ser dicotómica. Las variables independientes pueden estar a nivel de intervalo o ser categóricas; si son categóricas, deben ser variables auxiliares o estar codificadas como indicadores (existe una opción en el procedimiento para recodificar automáticamente las variables categóricas).

2

Supuestos. La regresión logística no se basa en supuestos distribucionales en el mismo sentido en que lo hace el análisis discriminante. Sin embargo, la solución puede ser más estable si los predictores tienen una distribución normal multivariante

3

Procedimientos relacionados. Utilice el procedimiento Diagrama de dispersión para mostrar en pantalla sus datos para multicolinealidad. Si se cumplen los supuestos de normalidad multivariante y de matrices de varianzas-covarianzas iguales, puede obtener una solución más rápida utilizando el procedimiento Análisis discriminante

Métricas

1. **True Positives (TP):** cuando la clase real del punto de datos era 1 (Verdadero) y la predicha es también 1 (Verdadero)
2. **Verdaderos Negativos (TN):** cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).
3. **False Positives (FP):** cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True).
4. **False Negatives (FN):** Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso).

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Precision = $\frac{TP}{TP + FP}$

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Recall = $\frac{TP}{TP + FN}$

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Specificity = $\frac{TN}{TN + FP}$

Métricas

Métricas

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

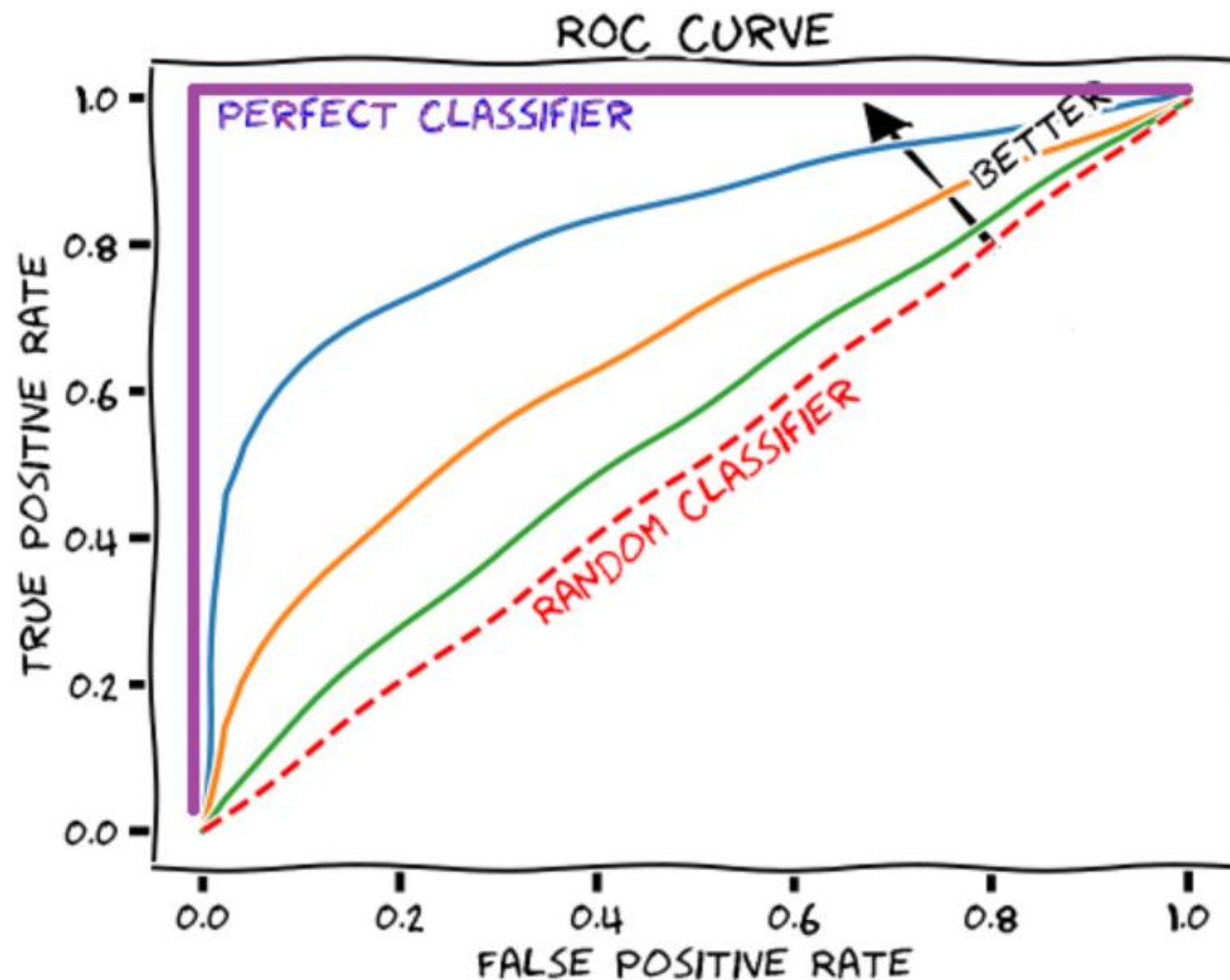
Métricas

Curva ROC

ROC es un acrónimo para Receiver Operating Characteristic (Característica Operativa del Receptor). Es una gráfica que enfrenta el ratio de falsos positivos (eje x) con el ratio de falsos negativos (eje y).

La curva ROC es útil por dos principales motivos:

- Permite comparar diferentes modelos para identificar cual otorga mejor rendimiento como clasificador
- El área debajo de la curva (AUC) puede ser utilizado como resumen de la calidad del modelo



Aplicaciones del algoritmo de regression logistica



Evalúe el riesgo crediticio



Cree perfiles de los consumidores de productos envasados



Incremente las ganancias en la industria de banca

Bibliotecas



Manipulación y análisis de datos



Creación de vectores y matrices
Colección de funciones matemáticas



Generación de gráficos



Biblioteca de visualización basada en Matplotlib



Biblioteca de aprendizaje automático

Reto



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Kaggle · 11,727 teams · Ongoing · ID 3136