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.

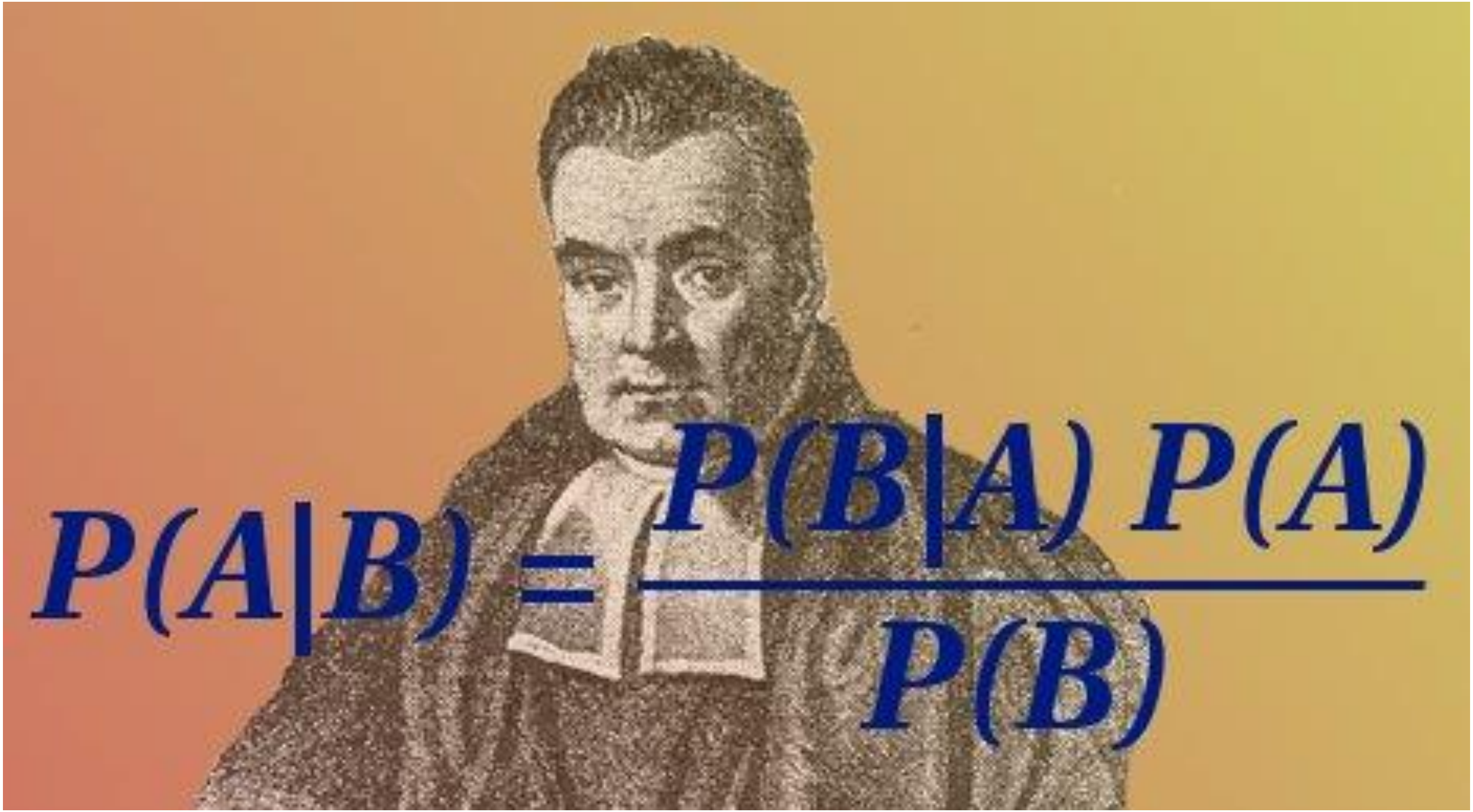
EGADE Business School
Tecnológico de Monterrey

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

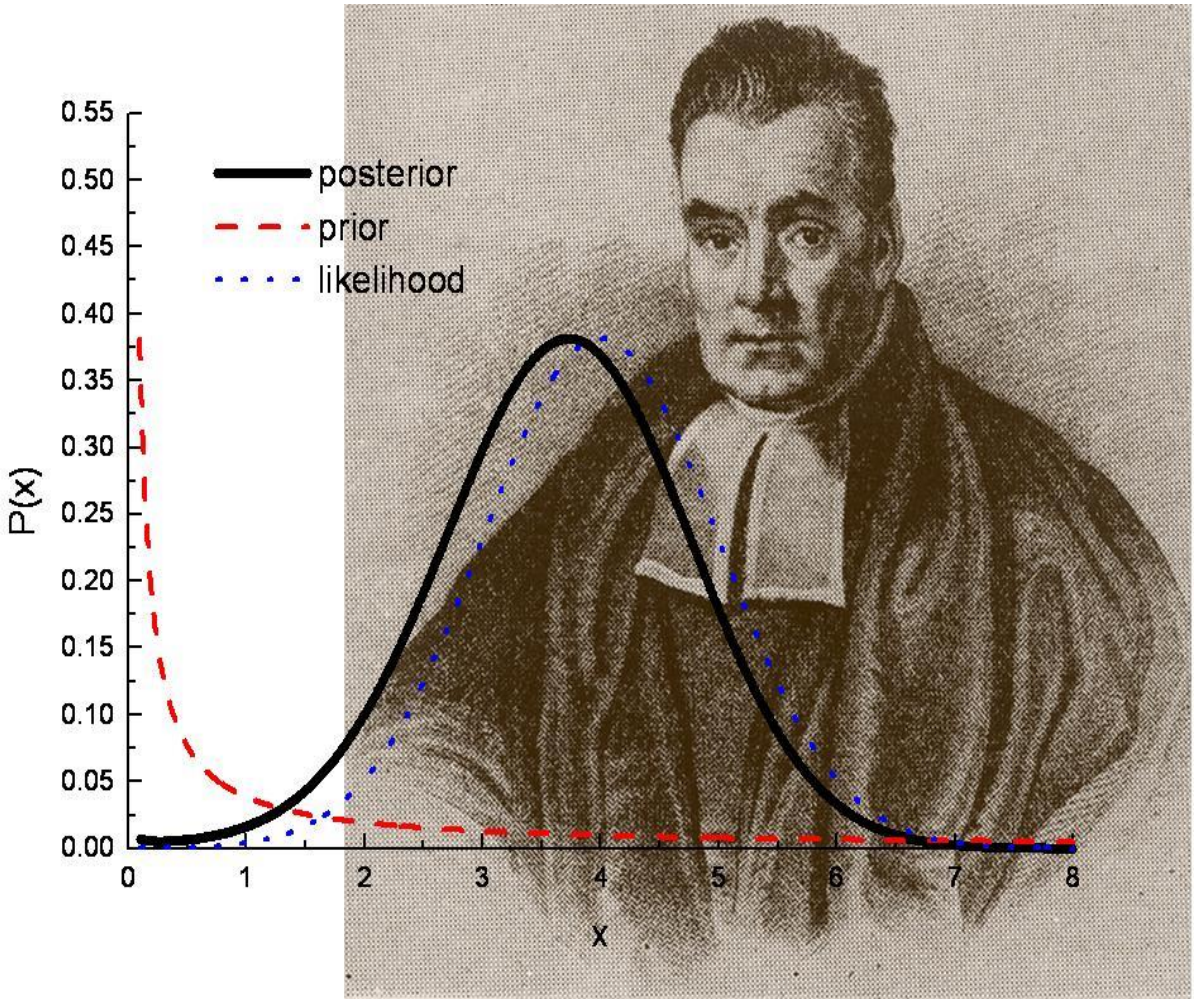
〈 Naïve Bayes.


$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

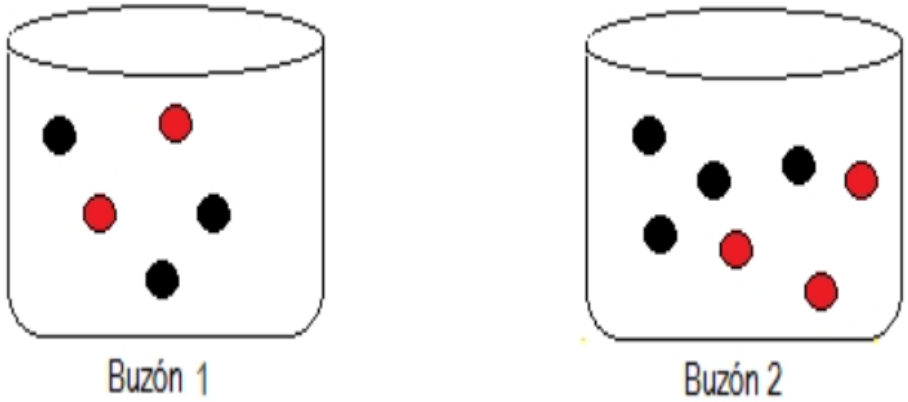
Teorema de Bayes

Teorema (Bayes, 1764):
Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por p(A) y p(B) respectivamente, verificándose que p(B) > 0. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, p(A) y p(B), así como la probabilidad condicionada del suceso B dado el suceso A, es decir p(B|A). La probabilidad a posteriori del suceso A conocido que se verifica el suceso B, es decir p(A|B), puede calcularse a partir de la siguiente formula:

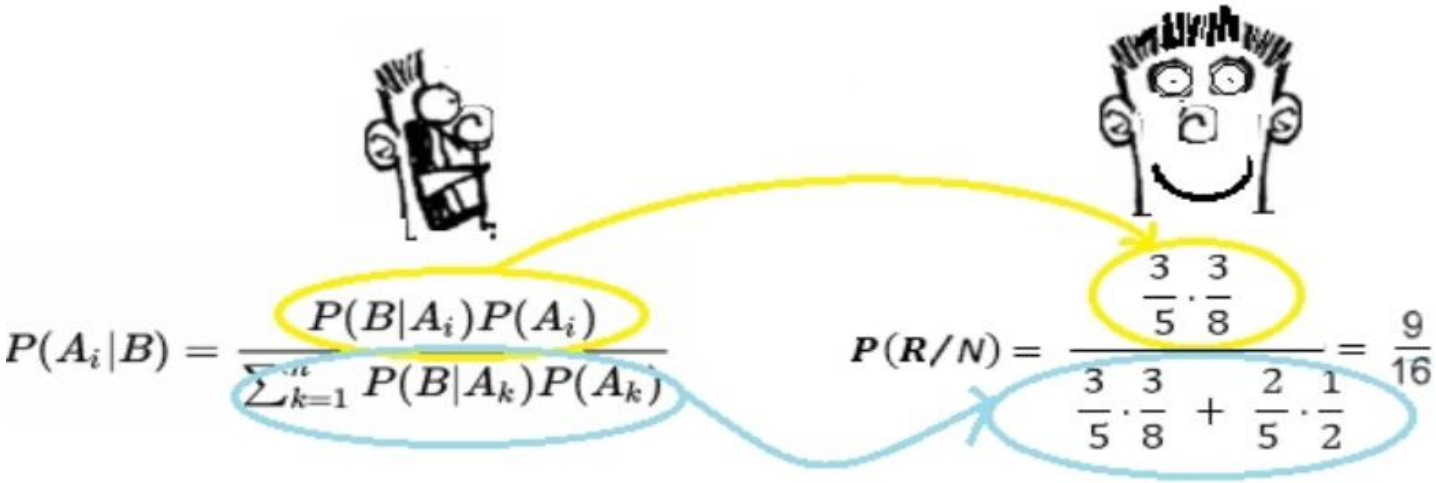
$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)p(B|A)}{\sum_{A'} p(A')p(B|A')}$$



Teorema de Bayes

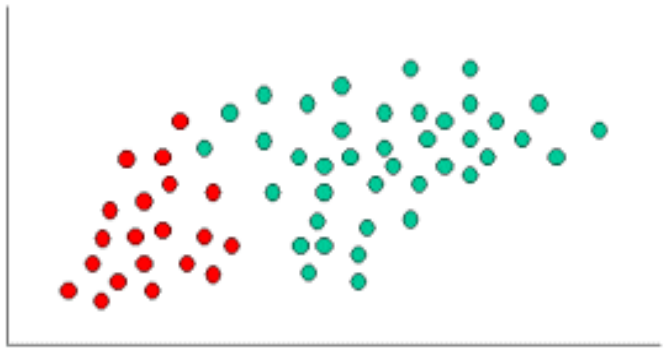


¿Cuál es la probabilidad que la bolita del segundo buzón sea roja, si la bolita del primer buzón salió negra?



Algoritmo Naïve Bayes

El algoritmo clasificador Naïve-Bayes (NBC), es un clasificador probabilístico simple con fuerte suposición de independencia. Aunque la suposición de la independencia de los atributos es generalmente una suposición pobre y se viola a menudo para los conjuntos de datos verdaderos. A menudo proporciona una mejor precisión de clasificación en conjuntos de datos en tiempo real que cualquier otro clasificador. También requiere una pequeña cantidad de datos de entrenamiento. El clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior. También es útil para datos dimensionales altos ya que la probabilidad de cada atributo se estima independientemente.



Algoritmo Naïve Bayes

Ventajas:

- Naive Bayes es uno de los algoritmos de Machine Learning más rápidos y sencillos para predecir una clase de conjuntos de datos.
- Se utiliza para clasificaciones binarias y de clases múltiples.
- Funciona mejor que otros algoritmos cuando hablamos de predicciones multiclase.
- Es la opción más popular para problemas de clasificación de texto.

Desventajas:

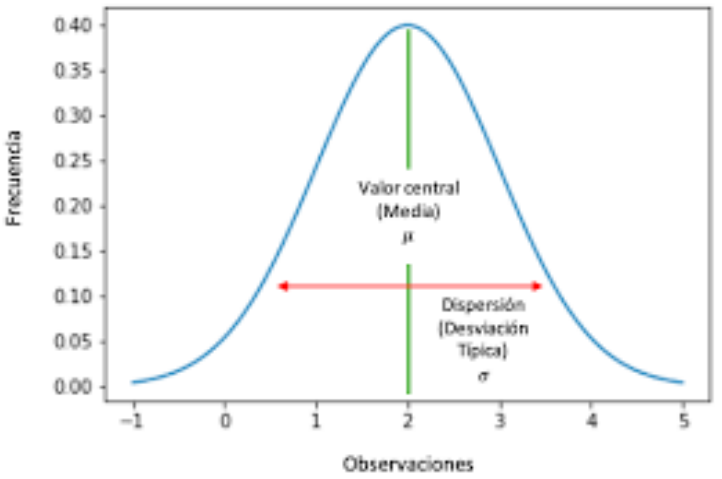
- Naive Bayes asume que todas las características son independientes entre sí, de modo que nunca podrá aprender la relación existente entre ellas.
- Además, a pesar de ser muy buenos clasificadores, los algoritmos Naive Bayes son conocidos por ser estimadores pobres.

Algoritmo Naïve Bayes: Tipos

Gaussiano: según este modelo, las características siguen una distribución normal. De modo que, en caso de que los predictores tomen valores continuos en lugar de discretos, el modelo asume que estos valores se muestrean a partir de la distribución gaussiana.

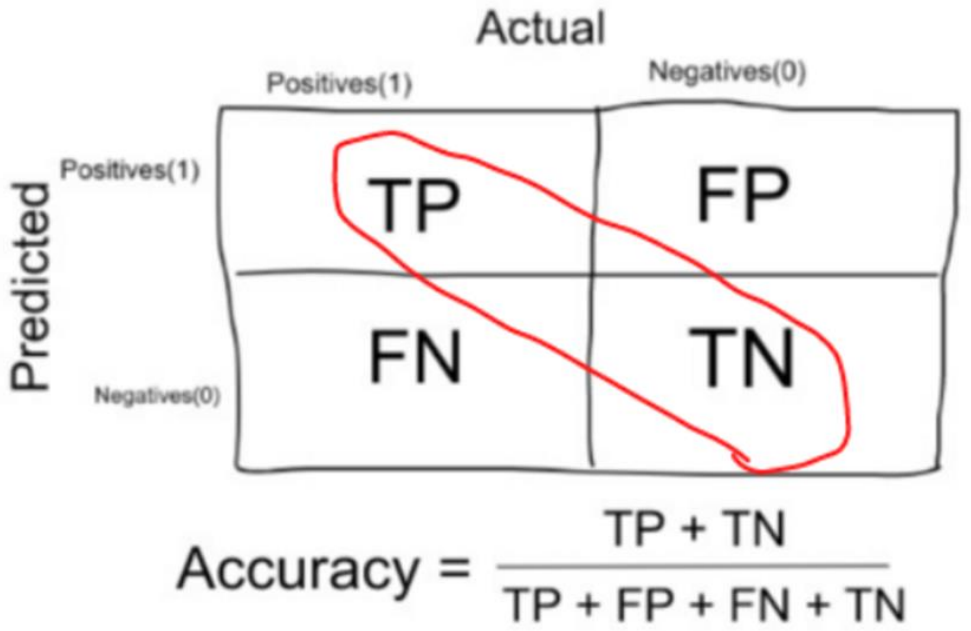
Multinomial: este modelo se usa cuando los datos cuentan con una distribución multinomial y se utilizan, principalmente, para resolver problemas de clasificación de documentos.

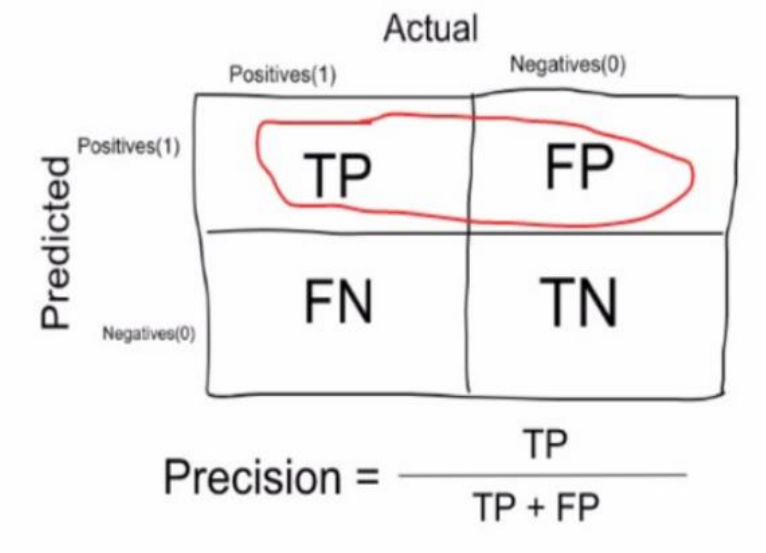
Bernoulli: el tipo Bernoulli tiene un funcionamiento parecido al multinomial, pero las variables predictoras son las variables booleanas independientes.



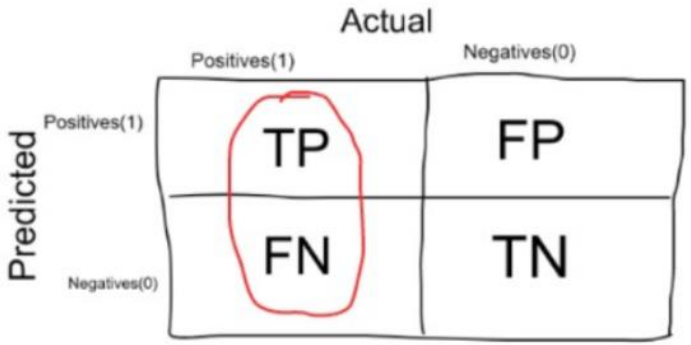
Métricas

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).

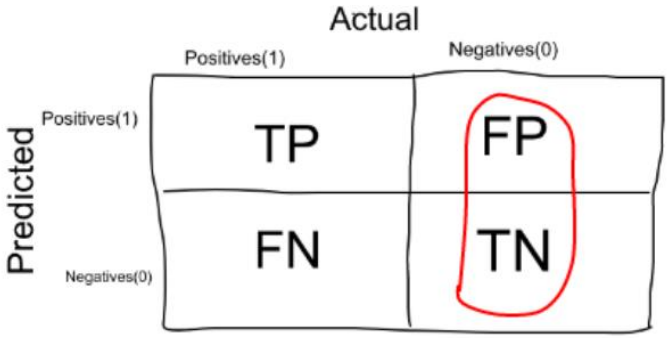




$$\text{Precision} = \frac{TP}{TP + FP}$$



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



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

Métricas

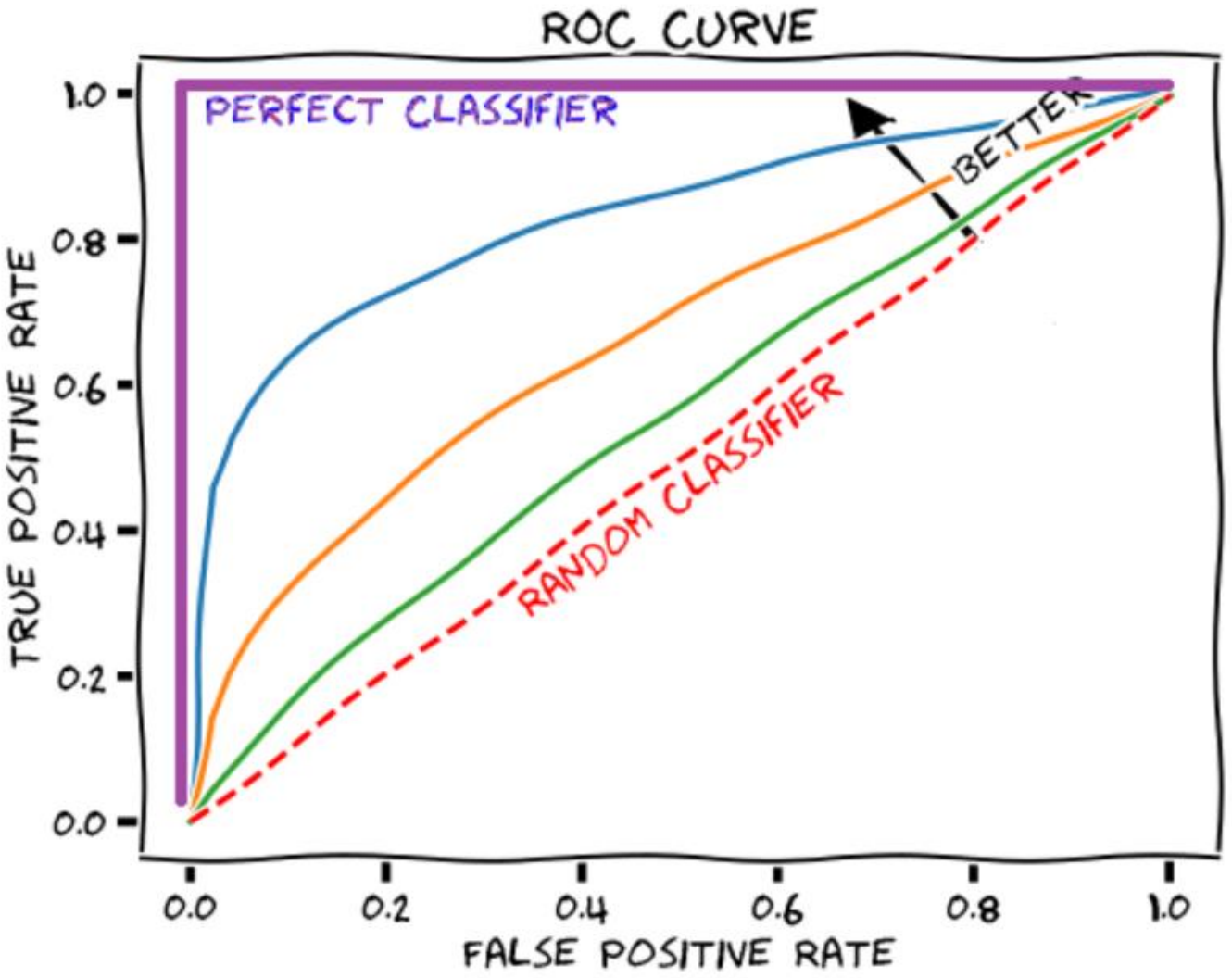
Métricas

Curva ROC

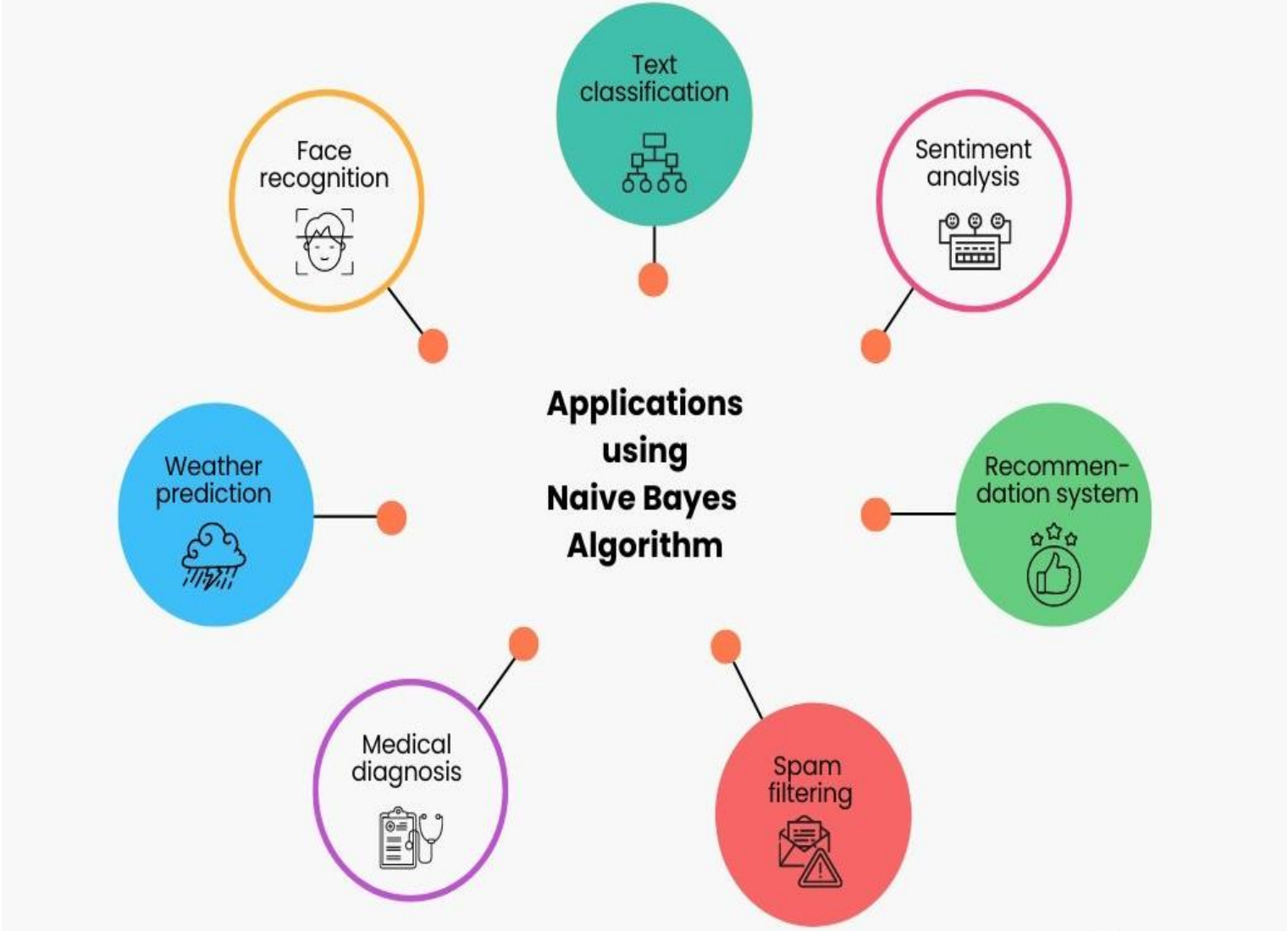
ROC es un acrónimo para Receiver Operating Characteristic (Característica Operativa del Receptor). Es una gráfica que enfrenta el ratio de falsos positivos (eje x) con el ratio de falsos negativos (eje y).

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 Naïve Bayes



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