

Clase N°6. Clasificador de Bayesiano ingenuo. (Naïve Bayes)

El modelo se utiliza como método de clasificación cuando la variable respuesta es categórica. Está basado en el teorema de Bayes utilizado para inferir la probabilidad de ocurrencia de un suceso dado que ha ocurrido previamente otro evento.

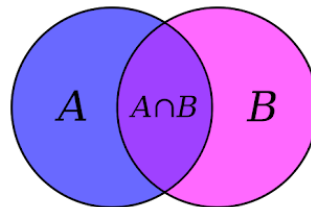
Se llama Bayes ingenuo ya que presume que los factores del modelo son linealmente independientes y este hecho en la práctica no siempre ocurre con frecuencia.

Revisión de conceptos de probabilidad condicional bayesiana:

Dados dos sucesos **mutuamente excluyentes** A y B (no pueden ocurrir al mismo tiempo), se tiene que:

$$\text{con } P(B) \neq 0,$$

la probabilidad que ocurra el suceso A dado que ocurrió el suceso B es igual a la probabilidad que ocurran ambos sucesos sobre la probabilidad que ocurra el suceso B.



En la **probabilidad condicional** la información previa tiene relación con los sucesos que ocurren posteriormente e influyen en su probabilidad de ocurrencia.

Ejemplo 1:

¿Cuál es la probabilidad que llueva dado que está nublado, hay baja presión, y un gran porcentaje de humedad?

¿Cuál es la probabilidad que un avión despegue dado que hay tormenta y cae granizo en la pista?

Ejemplo 2: Sistema de admisión en una universidad de USA.

En un test de admisión a una universidad en USA, los alumnos efectuaron una denuncia ante presunta discriminación en el ingreso de estudiantes extranjeros. Siendo 259 estudiantes nativos (N) y 48 extranjeros (E). Al tomar una persona al azar la probabilidad que sea extranjero y nativos es respectivamente:

$$P(E) = 48/307 = 0,16, \quad P(N) = 259/307 = 0,84.$$

De los estudiantes extranjeros solo 26 aprobaron el examen y el resto falló. Por otra parte de los estudiantes nativos 206 promovieron y el resto no alcanzó. Entonces, a partir de un total de 232 alumnos que aprobaron, resulta que:

La probabilidad que una persona apruebe el examen dado que es extranjera es:

$$P(A/E) = P(A \cap E)/P(E) = 26/48 = 0,54.$$

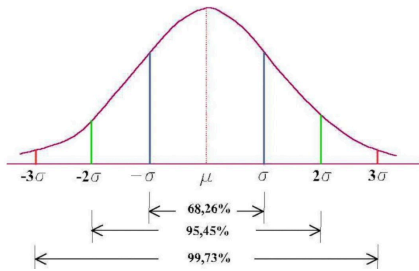
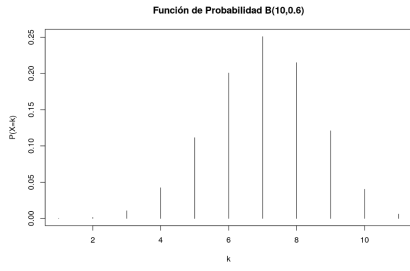
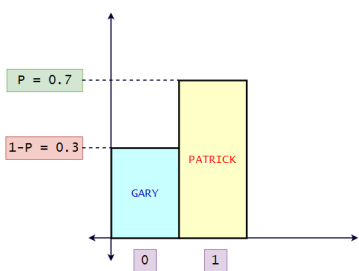
Por otra parte, la probabilidad que una persona supere el examen dado que es nativa es:

$$P(A/N) = P(A \cap N)/P(N) = 206/259 = 0,80.$$

En este caso, existe una mayor probabilidad que un nativo apruebe el examen.

Tipos de modelos Naïve Bayes

Dependiendo del tipo de dato de entrada (variable observable o característica) asociada con cada clase que se quiera predecir, existen diferentes modelos de Naïve Bayes, entre los más utilizados se pueden mencionar:

GaussianNB	MultinomialNB	BernoulliNB
La variable observable es un dato continuo. (3.14, 6.87...)	La variable observable es un dato discreto (1, 2, 3, 4...)	La variable observable es un dato binómico. (0,1)
<p>Distribución normal.</p> 	<p>Distribución multinomial.</p> 	<p>Distribución de Bernoulli.</p> 
Clasifica características numéricas.	Clasifica características de conteo. Frecuencia de palabras (PNL).	Clasifica características binómicas. Presencia o ausencia de palabras.
<p>Sea (μ) la media aritmética y (σ) el desvío standard de xi datos, la función de densidad de una distribución normal es:</p> $f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$	<p>Sea p_i, la probabilidad del evento i y n_i la cardinalidad de cada suceso, entonces la función de densidad de una distribución multinomial es:</p> $P = \frac{n!}{(n_1!)(n_2!)\dots(n_x!)} P_1^{n_1} P_2^{n_2} \dots P_x^{n_x}$	<p>Sea p, la probabilidad de éxito y $q=1-p$, la probabilidad de fracaso, el nro. de pruebas y de éxitos, entonces la función de probabilidad es:</p> $F(x) = \sum_{k=0}^x \binom{n}{k} p^k \cdot q^{n-k}$

Ejemplo 3:

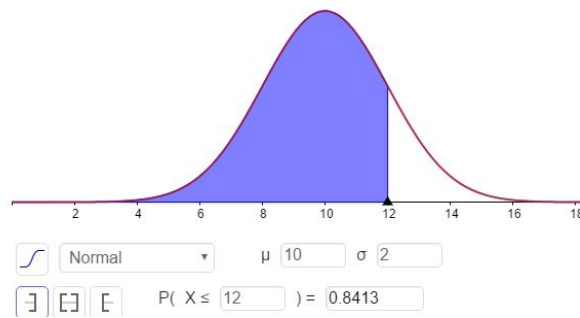
En un laboratorio se toma una muestra de un conjunto de datos aleatorios (x) que presentan una distribución normal, cuya media es 10 y su desviación standard es de 2. Hallar la probabilidad que al tomar una variable al azar sea menor o igual que 12.

Sea $N(x; \mu, \sigma)$, entonces, la probabilidad de obtener un 12 es:

$$N(12, 10, 2), z = (x - \mu)/\sigma = (12 - 10)/2 = 1$$

Utilizando una tabla de distribución normal, la probabilidad es:

z	0.00	0.01
0.0	0.5000	0.5040
0.1	0.5398	0.5438
0.2	0.5793	0.5832
0.3	0.6179	0.6217
0.4	0.6554	0.6591
0.5	0.6915	0.6950
0.6	0.7257	0.7291
0.7	0.7580	0.7611
0.8	0.7881	0.7910
0.9	0.8159	0.8186
1.0	0.8413	0.8438
1.1	0.8643	0.8665
1.2	0.8849	0.8869
1.3	0.9032	0.9049
1.4	0.9192	0.9207



Ejemplo 4:

En un puesto de celulares hay 5 Samsung, 3 Nokia y 2 iPhone. ¿Cuál es la probabilidad de tomar al azar: 2 Samsung, 2 Nokia y 1 iPhone?

$$p_1 = 5/10 = 0.5$$

$$x_1 = 2$$

$$p_2 = 3/10 = 0.3$$

$$x_2 = 2$$

$$p_3 = 2/10 = 0.2$$

$$x_3 = 1$$

$$n = 5$$

$$P(x) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

$$\frac{5!}{(2!)(2!)(1!)} (0,5)^2 (0,3)^2 (0,2)^1 = ,135 \quad 13\%$$

Ejemplo 5:

Se arroja 5 veces una moneda al aire y se pretende determinar la probabilidad de obtener cara 3 veces.

Sea p , la probabilidad que salga cara (éxito) $= 1/2$ y sea q , la probabilidad que no salga cara (fracaso) $= 1/2$, entonces,

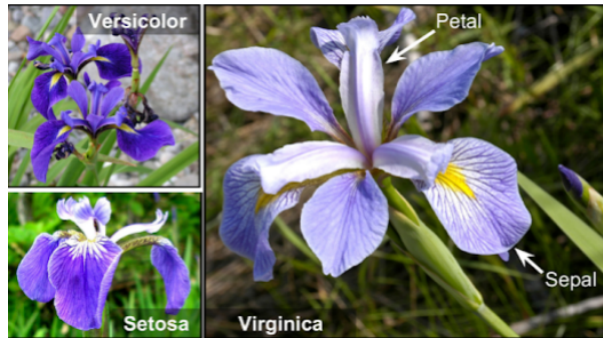
$$P(X = 3) = f(3) = \binom{5}{3} (0,5)^3 (0,5)^{5-3}$$

$$P(X=3) = [5! / (3! \cdot 2!)] \cdot 0,5^3 \cdot 0,5^2 = 10 \cdot 0,125 \cdot 0,25 = \mathbf{0,3125}$$

La probabilidad sería por lo tanto un 31,25%

Ejemplo 6: base de datos iris.

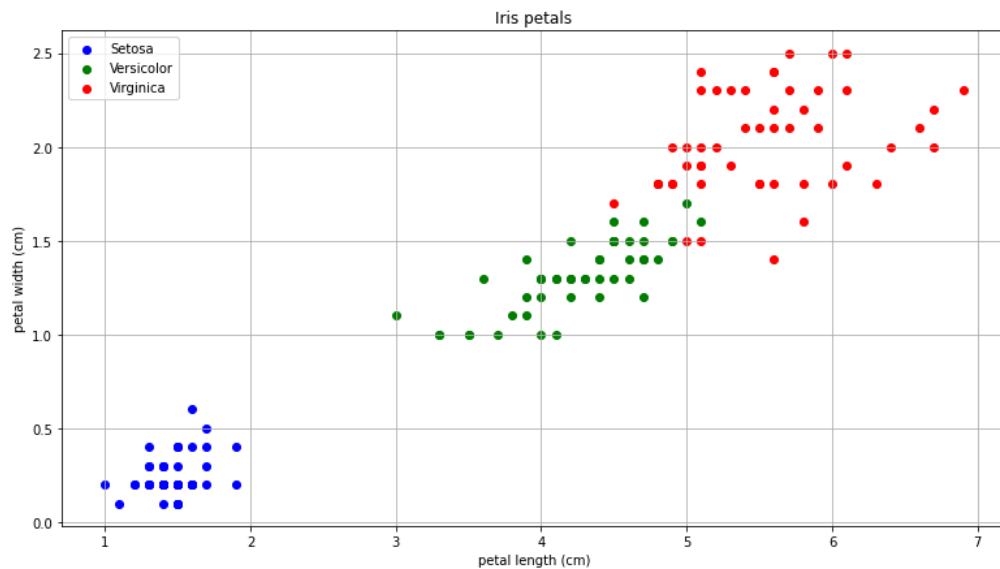
La base de datos Iris, contiene información sobre 150 especies de flores clasificadas como: Virginica, Setosa y Versicolor según el ancho y el largo del sépalos y del pétalo.



La base de datos contiene la sig. información:

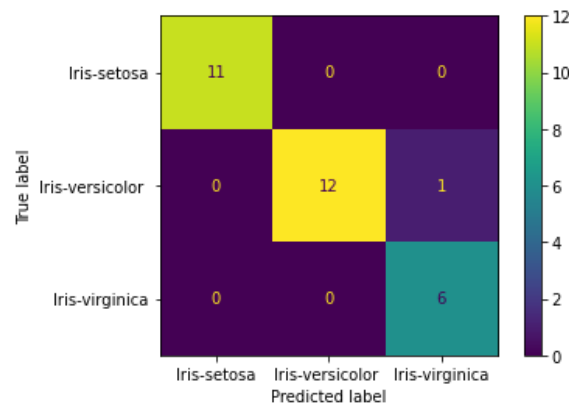
	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

En base a los datos de la tabla, se muestra el tipo de especie en relación al largo y ancho del pétalo:



Luego aplicando un modelo predictivo Naïve Bayes, se puede estimar el tipo de especie en relación al largo y ancho del pétalo.

Al comparar el valor real con el valor predicho de la variable respuesta “target” se obtiene la sig. matriz de confusión:



Donde las especies Iris-Setosa e Iris-Virginica son correctamente clasificadas, mientras que la variante Iris-Versicolor fue una vez mal catalogada como Iris-Virginica. Esta predicción arroja un 97% de exactitud en la predicción, como lo reflejan las sig. métricas:

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	11
Iris-versicolor	1.00	0.92	0.96	13
Iris-virginica	0.86	1.00	0.92	6
accuracy			0.97	30
macro avg	0.95	0.97	0.96	30
weighted avg	0.97	0.97	0.97	30

Métricas del modelo Naïve Bayes

En base a la matriz de confusión se pueden obtener las principales métricas del modelo.

		Actual (True) Values	
		Positive	Negative
Predicted Values	Positive	TP	FP
	Negative	FN	TN

- **TP: True Positive:** casos positivos correctamente predichos.
- **TN: True Negative:** casos negativos correctamente predichos.
- **FP: False Positive:** casos positivos predichos de manera errónea.
- **FN: False Negative:** casos negativos predichos de manera errónea.

Precision	Mide cuantas de las predicciones son correctas en relación a la totalidad de los casos positivos. Precision=TP/ (TP + FP).
Recall/Sensitivity	Mide el porcentaje de casos positivos son correctamente clasificados Recall=TP/ (TP + FN).
F1-score	Métrica que combina la media armónica entre la precisión y la sensibilidad. F1-score=2*(Precision*Recall)/ (Precision+Recall)
Accuracy	Número de predicciones correctas sobre el total de predicciones. Accuracy= (TP+TN)/Data set size.

Macro avg	Estima la precisión y la sensibilidad del modelo teniendo en cuenta todas las categorías juntas.
Weighted avg	Estima la precisión y sensibilidad del modelo considerando Un porcentaje ponderado de cada categoría del data set.

Actividad:

1. Tomando la base de datos Iris, clasificar el tipo de flor en relación al largo y ancho del sépalo.
2. Crear una tabla comparando el valor real y el predicho y representarla en una matriz de confusión.
3. Hallar las métricas y analizar sus resultados.