Naive Bayes

Algoritmo que gracias a un simple supuesto puede ser usado para hacer predicciones

Se basa en el teorema de Bayes:

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

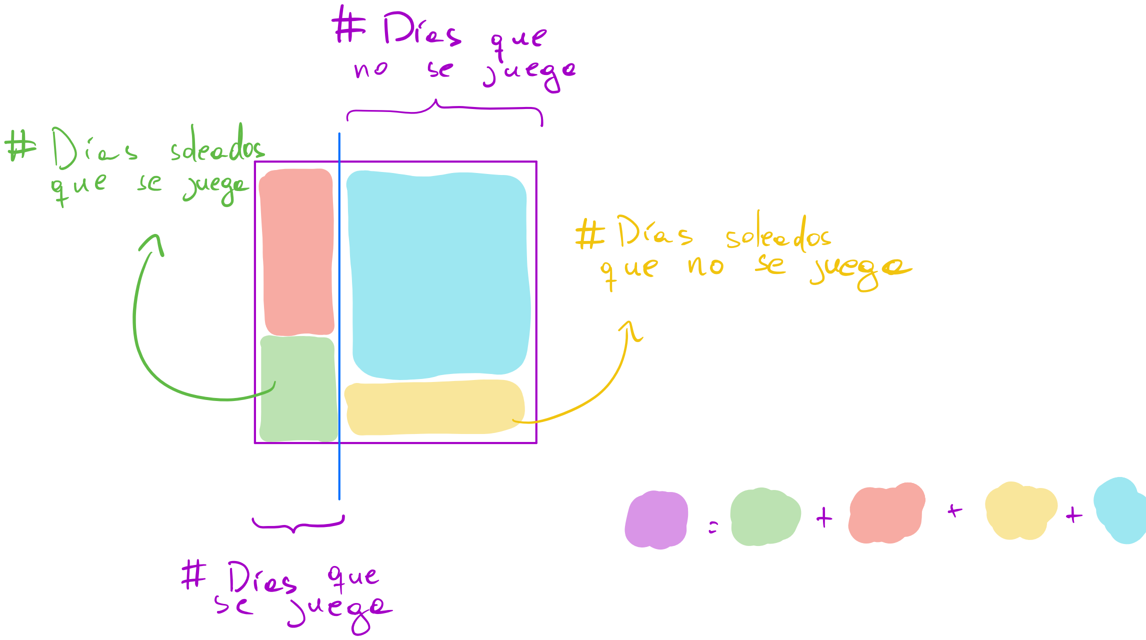
Pero, ¿qué nos dice este teorema?
Supongamos dos eventos relacionados:

J es el evento jugar al fútbol un día
S es el evento el día está soleado

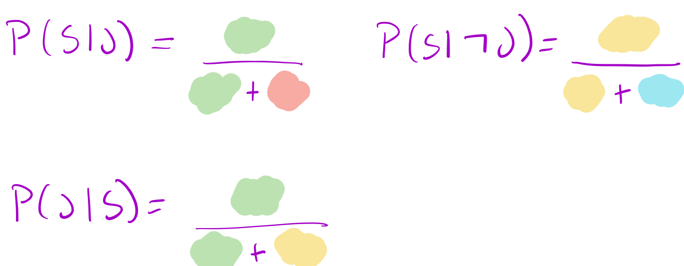
Entonces:

$$P(J|S) = \frac{P(J)\, P(S|J)}{P(S)}$$

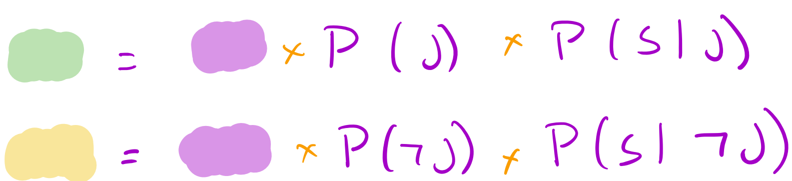
Intentemos ver esto de forma gráfica:



Queremos $P(J|S)$



Pero en términos de probabilidades,
¿Cuanto valen y ?



$$P(J|S) = \frac{\times P(J) \times P(S|J)}{\times P(J) \times P(S|J) + \times P(\neg J) \times P(S|\neg J)}$$

$$P(J|S) = \frac{P(J) \times P(S|J)}{P(J) \times P(S|J) + P(\neg J) \times P(S|\neg J)}$$

Pero también $+ \; = \; \times P(S)$

Entonces:

$$P(J|S) = \frac{P(J)\, P(S|J)}{P(S)}$$

Ahora, ¿cómo usamos esto para clasificar?

Supongamos que queremos predecir si un suceso Y sucederá o no $(0, 1)$ igual a i, dado que sabemos $X_i, \ldots, X_p$.

Esto es:

$$P(Y \mid X_i, \ldots, X_p)$$

Si seguimos la mecánica hasta ahora, tenemos:

$$P(Y=i|X_1,\dots,X_p) = \frac{P(Y=i)\,P(X_1,\dots,X_p|Y=i)}{P(Y=0)\,P(X_1,\dots,X_p|Y=0)+P(Y=1)\,P(X_1,\dots X_p|Y=1)}$$

Pero aquí usamos la parte "naive":
vamos a asumir que:

$$\boxed{P(X_1,\dots,X_p|Y=i) = \prod_{j=1}^{p} P(X_j|Y=i)}$$

$$\underbrace{\prod_{j=1}^{p} P(X_j|Y=i)}_{P(X_1|Y=i)\times\dots\times P(X_p|Y=i)}$$

Veamos un ejemplo:

Queremos predecir si este sábado habrá partido de la liga de computación UAI según estos datos:

Día de la semana	Clima	Partido
Lunes	Sol	No
Martes	Lluvia	No
Miércoles	Lluvia	Sí
Jueves	Sol	No
Viernes	Sol	Sí
Sábado	Sol	Sí
Lunes	Lluvia	Sí
Miércoles	Sol	No
Viernes	Sol	No
Sábado	Sol	No
Lunes	Lluvia	Sí
Sábado	Lluvia	Sí
Martes	Sol	Sí
Sábado	Sol	Sí
Sábado	Lluvia	No

Queremos $P(J = \text{"Sí"} \mid D = \text{"S"}, C = \text{"Sol"})$

$$= \frac{P(J=\text{"Sí"})\,P(D=\text{"S"} \mid J=\text{"Sí"})\,P(C=\text{"Sol"} \mid J=\text{"Sí"})}{P(J=\text{"Sí"})\,P(D=\text{"S"} \mid J=\text{"Sí"})\,P(C=\text{"Sol"} \mid J=\text{"Sí"}) + P(J=\text{"No"})\,P(D=\text{"S"} \mid J=\text{"No"})\,P(C=\text{"Sol"} \mid J=\text{"No"})}$$

$$P(J=\text{"Si"}) = \frac{8}{15} \qquad P(J=\text{"No"}) = \frac{7}{15}$$

$$P(C=\text{"Sol"} \mid J=\text{"Si"}) = \frac{4}{8}$$

$$P(D=\text{"S"} \mid J=\text{"Si"}) = \frac{3}{8}$$

$$P(C=\text{"Sol"} \mid J=\text{"No"}) = \frac{5}{7}$$

$$P(D=\text{"S"} \mid J=\text{"No"}) = \frac{2}{7}$$

$$\frac{P(J=\text{"Si"})\,P(D=\text{"S"} \mid J=\text{"Si"})\,P(C=\text{"Sol"} \mid J=\text{"Si"})}{P(J=\text{"Si"})\,P(D=\text{"S"} \mid J=\text{"Si"})\,P(C=\text{"Sol"} \mid J=\text{"Si"}) + P(J=\text{"No"})\,P(D=\text{"S"} \mid J=\text{"No"})\,P(C=\text{"Sol"} \mid J=\text{"No"})}$$

$$= \frac{\frac{8}{15} \cdot \frac{3}{8} \cdot \frac{4}{8}}{\frac{8}{15} \cdot \frac{3}{8} \cdot \frac{4}{8} + \frac{7}{15} \cdot \frac{2}{7} \cdot \frac{5}{7}}$$