Introducción al clasificador bayesiano ingenuo para detección de correo basura

Supongamos que tenemos los siguientes mensajes ya clasificados como correo basura o correo normal:

basura	mándanos tu clave
normal	mándanos tu coche
normal	revisión de tu coche
basura	revisión de tu clave
basura	revisión de coche

Si recibimos un nuevo mensaje, utilizando la inferencia clásica, estimaríamos la probabilidad de que sea correo basura como:

$$\mathrm{P}(\text{basura}) = \frac{3}{5}$$

y igualmente, estimaríamos la probabilidad de que sea un correo normal como:

$$\mathrm{P}(\text{normal}) = 1 - \mathrm{P}(\text{basura}) = \frac{2}{5}.$$

Ahora, supongamos que el un nuevo mensaje contiene la palabra "revisión". Estimamos que la probabilidad de que un correo basura contenga la palabra es:

$$\mathrm{P}(\text{revisión}|\text{basura}) = \frac{2}{3}$$

y la probabilidad de que un correo normal contenga la palabra como

$$\mathrm{P}(\text{revisión}|\text{normal}) = \frac{1}{2}.$$

Luego, por el teorema de Bayes,

$$\mathrm{P}(\text{basura}|\text{revisión}) = \frac{\mathrm{P}(\text{revisión}|\text{basura})\mathrm{P}(\text{basura})}{\mathrm{P}(\text{revisión})}$$

donde el denominador es:

$$\mathrm{P}(\text{revisión}) = \mathrm{P}(\text{revisión}|\text{basura})\mathrm{P}(\text{basura}) + \mathrm{P}(\text{revisión}|\text{normal})\mathrm{P}(\text{normal})$$

Haciendo los cálculos, tenemos:

$$\mathrm{P}(\text{revisión}) = \frac{2}{3} * \frac{3}{5} + \frac{1}{2} * \frac{2}{5} = \frac{3}{5}$$

$$\mathrm{P}(\text{basura}|\text{revisión}) = \frac{\frac{2}{3} * \frac{3}{5}}{\frac{3}{5}} = \frac{2}{3}.$$

Ahora supongamos que el mensaje entero es “mándanos tu revisión de coche”.

Estimando la probabilidad de la frase entera a través de la tabla, tendríamos:

$$\mathrm{P}(\text{mándanos tu revisión de coche}|\text{basura}) = 0,$$

$$\mathrm{P}(\text{mándanos tu revisión de coche}|\text{normal}) = 0,$$

porque la frase exacta no ocurre nunca entre los mensajes que tenemos en la tabla y luego es imposible calcular $\mathrm{P}(\text{basura}|\text{mándanos tu revisión de coche})$.

Claramente no es muy útil, ya que es el nuevo mensaje que acabamos de recibir y tiene que ser o normal o correo basura. Luego necesitamos una simplificación.

Observamos primero que es fácil estimar la probabilidad de ocurrencia de cada palabra individual:

Palabra	P(palabra|basura)	P(palabra|normal)
mándanos	1/3	1/2
tu	2/3	2/2
clave	2/3	0/2
coche	1/3	2/2
revisión	2/3	1/2
de	1/3	1/2

El clasificador bayesiano ingenuo supone que las palabras en cada mensaje son *sucesos independientes* condicionado en el tipo de mensaje.

Luego, estimamos que las probabilidades condicionadas de ver el mensaje entero son:

$$\mathrm{P}(\text{mándanos tu revisión de coche}|\text{basura}) = \frac{1}{3} * \frac{2}{3} * \left(1 - \frac{2}{3}\right) * \frac{1}{3} * \frac{2}{3} * \frac{1}{3} \approx 0,0055$$

$$\mathrm{P}(\text{mándanos tu revisión de coche}|\text{normal}) = \frac{1}{2} * \frac{2}{2} * \left(1 - \frac{0}{2}\right) * \frac{2}{2} * \frac{1}{2} * \frac{1}{2} = 0,125$$

Observamos que tenemos que tomar en cuenta tanto las palabras que sí ocurren en el mensaje como los que no ocurren.

Ahora, podemos calcular la probabilidad marginal de ver el mensaje mediante el teorema de la probabilidad total:

$$\mathrm{P}(\text{mándanos tu revisión de coche}) = \frac{3}{5} * 0,0055 + \frac{2}{5} * 0,125 \approx 0,0533$$

y finalmente, por la regla de Bayes, la probabilidad de que el mensaje sea basura es:

$$\mathrm{P}(\text{basura}|\text{mándanos tu revisión de coche}) = \frac{\frac{3}{5} * 0,0055}{0,0533} \approx 0,062.$$