Naive Bayes Classifier

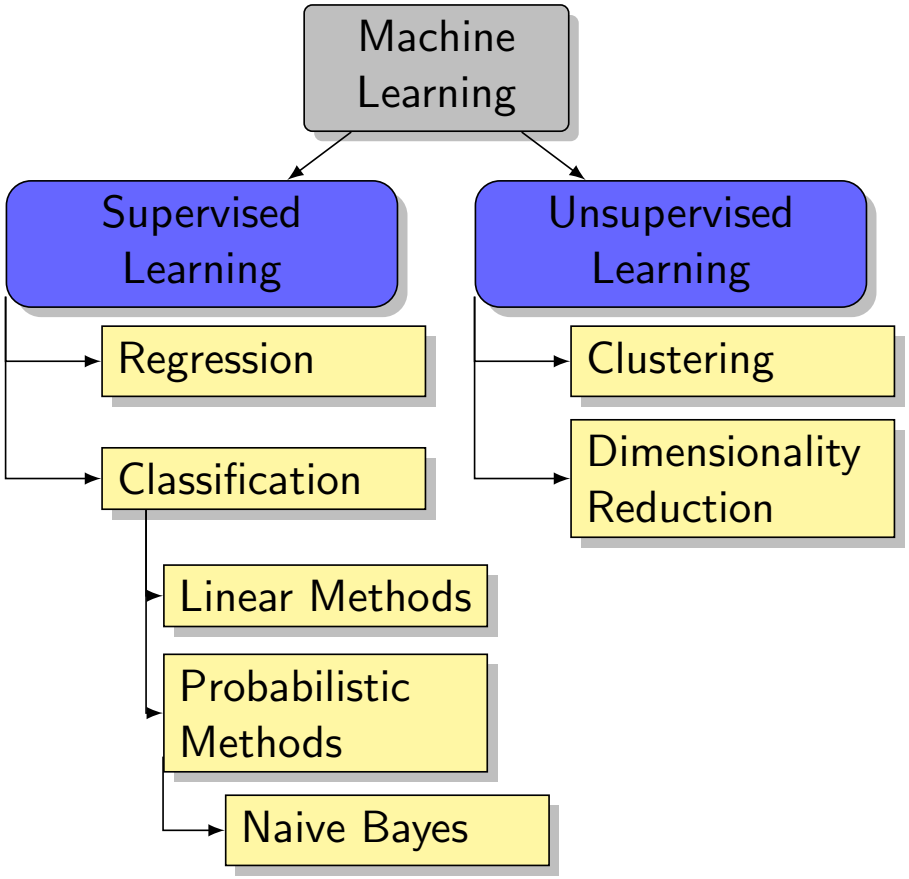
A probabilistic classifier

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python

Table of Contents

Introducción



Introducción

Clasificador Naive Bayes

Naive Bayes es un algortimo de clasificación binaria y multiclase. Se llama *Naive* Bayes o Bayes ingenuo porque se hacen suposiciones para simplificar los cálculos de probabilidades por cada clase.

Enfoque probabilístico

A diferencia de los clasificadores lineales que buscan una frontera de decisión que separe los datos en el espacio, un clasificador probabilístico busca estimar

$$P(\text{clase}_j | x_i)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase 0 si observamos los datos x.

Enfoque probabilístico

A diferencia de los clasificadores lineales que buscan una frontera de decisión que separe los datos en el espacio, un clasificador probabilístico busca estimar

$$P(\text{clase}_j | x_i)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase 0 si observamos los datos x.

En un problema de clasificación binaria, predecimos que un dato x pertenece a la clase 0 si

$$P(\text{clase}_0 | x) > P(\text{clase}_1 | x).$$

Enfoque probabilístico

A diferencia de los clasificadores lineales que buscan una frontera de decisión que separe los datos en el espacio, un clasificador probabilístico busca estimar

$$P(\text{clase}_j | x_i)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase 0 si observamos los datos x.

En un problema de clasificación binaria, predecimos que un dato x pertenece a la clase 1 si

$$P(\text{clase}_1 | x) > P(\text{clase}_0 | x).$$

Table of Contents

Conceptos básicos

- **Probabilidad Marginal**. La probabilidad de un evento independiente del resultado de otras variables aleatorias, $P(A)$.
  Si la variable es independiente, es la probabilidad del evento directamente. Si es dependiente de otras variables,

$$P(A) = \sum_{Y} P(A, Y).$$

Conceptos básicos

- **Probabilidad Marginal**. La probabilidad de un evento independiente del resultado de otras variables aleatorias, $P(A)$.
  Si la variable es independiente, es la probabilidad del evento directamente. Si es dependiente de otras variables,

$$P(A) = \sum_{Y} P(A, Y).$$

- **Probabilidad Conjunta**. Probabilidad de varios eventos simultaneos:

$$P(A, B).$$

Conceptos básicos

- **Probabilidad Marginal**. La probabilidad de un evento independiente del resultado de otras variables aleatorias, $P(A)$.
  Si la variable es independiente, es la probabilidad del evento directamente. Si es dependiente de otras variables,

$$P(A) = \sum_{Y} P(A, Y).$$

- **Probabilidad Conjunta**. Probabilidad de varios eventos simultaneos:

$$P(A, B).$$

- **Probabilidad Condicional**. Probabilidad de un evento dado que otro evento ha ocurrido, para dos variables dependientes

$$P(A|B).$$

Conceptos básicos

- **Regla del producto**. La probabilidad conjunta puede ser calculada usando la probabilidad condicional:

$$P(A, B) = P(A|B) \cdot P(B).$$

Por lo tanto, la probabilidad condicional puede ser calculada usando la probabilidad conjunta:

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

Conceptos básicos

- **Regla del producto**. La probabilidad conjunta puede ser calculada usando la probabilidad condicional:

$$P(A, B) = P(A|B) \cdot P(B).$$

Por lo tanto, la probabilidad condicional puede ser calculada usando la probabilidad conjunta:

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

- Si las variables son independientes

$$P(A, B) = P(A) \cdot P(B).$$

Teorema de Bayes

- **Teorema de Bayes**. Podemos calcular la probabilidad condicional sin usar la probabilidad conjunta:

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

- Teorema de Bayes para la clasificación:

$$P(Y_j|X_i) = \frac{P(X_i|Y_j) \cdot P(Y_j)}{P(X_i)},$$

$P(X_i|Y_j)$ es la función de similitud, la cual nos dice qué tanto la clase j explica, o hace creíbles, el dato X_i.

La distribución multinomial

Supongamos que se realiza un experimento consistente en extraer n bolas de k colores diferentes de una bolsa, sustituyendo las bolas extraídas después de cada extracción. Las bolas del mismo color son equivalentes.

La distribución multinomial

Supongamos que se realiza un experimento consistente en extraer n bolas de k colores diferentes de una bolsa, sustituyendo las bolas extraídas después de cada extracción. Las bolas del mismo color son equivalentes.

Denotemos por X_i la variable que denota el número de bolas extraídas de color i, y como p_i la probabilidad de que una extracción dada sea de color i. La función de masa de probabilidad de esta distribución multinomial es:

$$f(x_1, ..., x_k; n, p_1, ..., p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdot \cdots \cdot p_k^{x_k},$$

para enteros no negativos $x_1, ..., x_k$.

La distribución multinomial

En el caso de un problema de clasificación de documentos, consideramos el vocabulario del conjunto de documentos. Cada documento es considerado como la secuencia del número de ocurrencias de cada palabra del vocabulario dentro del documento.

La distribución multinomial

En el caso de un problema de clasificación de documentos, consideramos el vocabulario del conjunto de documentos. Cada documento es considerado como la secuencia del número de ocurrencias de cada palabra del vocabulario dentro del documento.

Entonces, la distribución multinomial nos dice que la probabilidad de que un documento pertenezca a la clase j es

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{k} p(x_i|\theta) = n! \cdot \frac{\theta_1^{x_1}}{x_1!} \cdot \ldots \cdot \frac{\theta_k^{x_k}}{x_k!}$$

donde $\mathbf{x}$ es el documento dado por la secuencia $(x_1, ..., x_k)$, k es el tamaño del vocabulario y $n = \sum x_i$. Además, $\theta = (\theta_1, ..., \theta_k)$ son las probabilidades de que cada palabra aparezca en la clase j.

La distribución multinomial

Es decir:

n:	longitud del documento (en palabras)
k:	número de palabras posibles
p_i:	probabilidad de que la palabra aparezca en la clase.

Supongamos que se realiza un experimento consistente en extraer n bolas de k colores diferentes de una bolsa, sustituyendo las bolas extraídas después de cada extracción. Las bolas del mismo color son equivalentes.

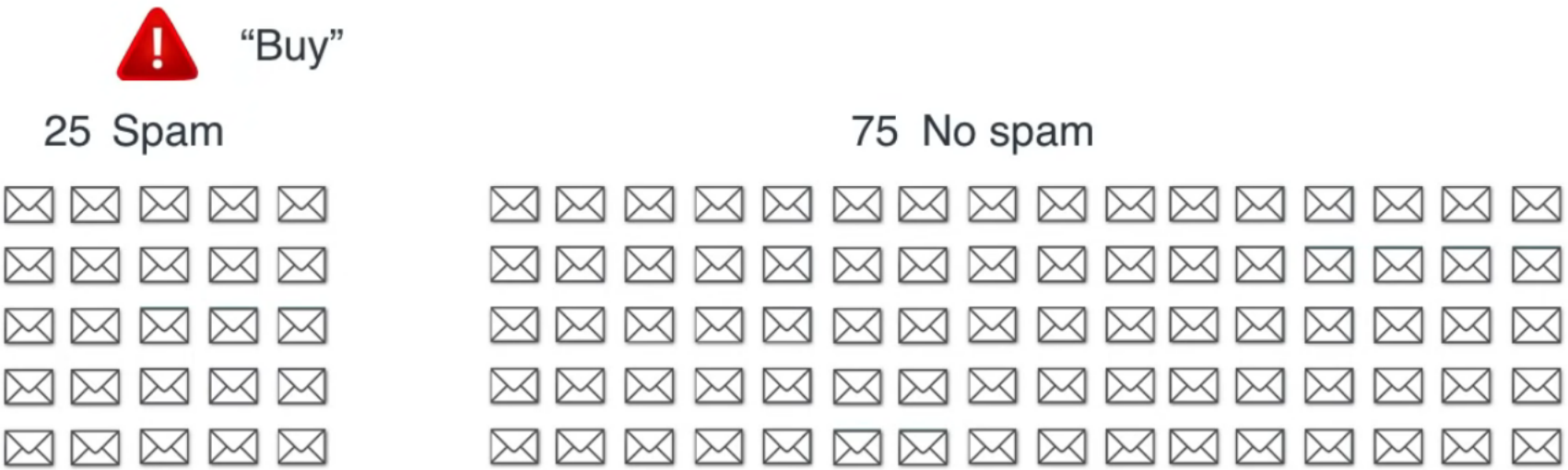
Table of Contents

Un ejemplo: Detección de SPAM

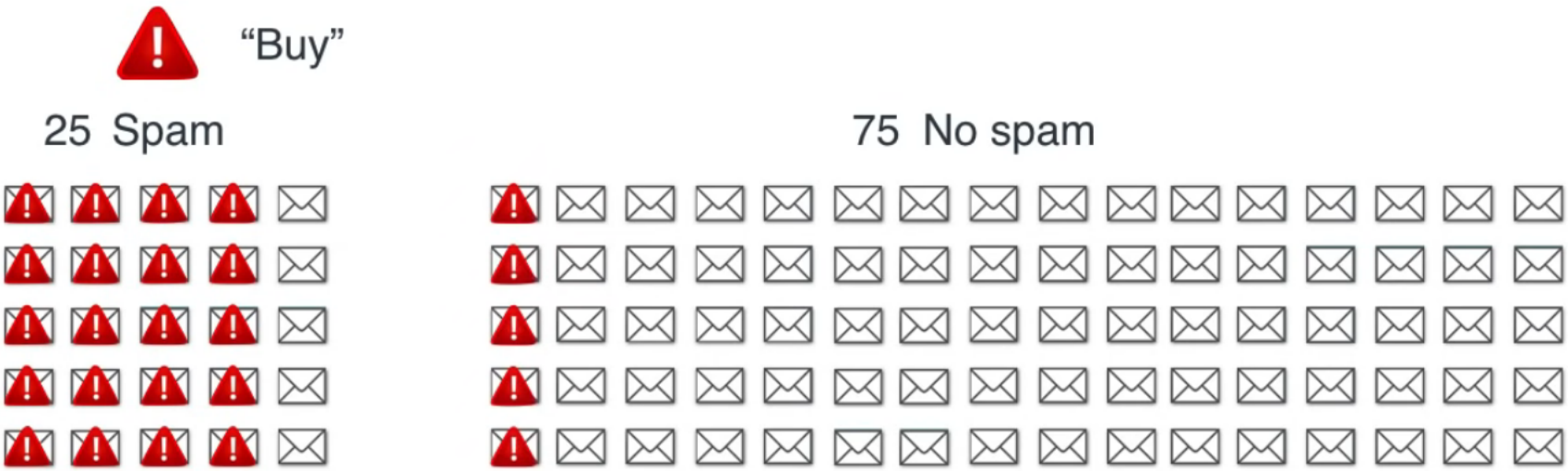
Un ejemplo: Detección de SPAM

Un ejemplo: Detección de SPAM

Buscamos propiedades que se correlacionen con que el correo sea SPAM o no. Por ejemplo, la aparición de ciertas palabras.



Un ejemplo: Detección de SPAM



Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra buy, cuál es la probabilidad de que sea SPAM?

Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra buy, cuál es la probabilidad de que sea SPAM? Podemos calcular directamente la probabilidad:

$$P(\text{Spam}|\text{buy}) = \frac{20}{20+5} = \frac{4}{5} = 0.8$$

$$\text{buy} \rightarrow 80\%$$

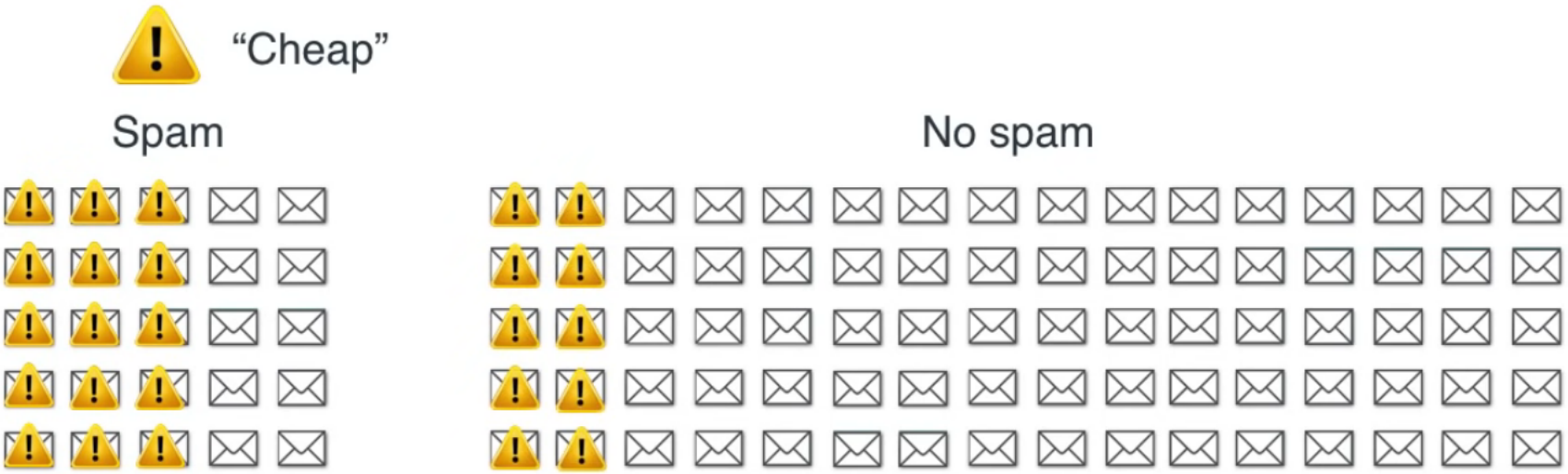
Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra buy, cuál es la probabilidad de que sea SPAM? Podemos usar también el teorema de Bayes:

$$P(\text{Spam}|\text{buy}) = \frac{P(\text{buy}|\text{Spam}) \cdot P(\text{Spam})}{P(\text{buy})} = \frac{\frac{20}{25}\frac{25}{100}}{\frac{25}{100}} = 0.8$$

Un ejemplo: Detección de SPAM

Ahora, veamos la palabra cheap:



Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra cheap, cuál es la probabilidad de que sea SPAM?

Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra cheap, cuál es la probabilidad de que sea SPAM? La calculamos directamente:

$$P(\text{Spam}|\textit{cheap}) = \frac{15}{15+10} = \frac{15}{25} = 0.6$$

$$\textit{cheap} \rightarrow 60\%$$

Un ejemplo: Detección de SPAM

¿Si un correo contiene la palabra cheap, cuál es la probabilidad de que sea SPAM? Usamos el teorema de Bayes

$$P(\text{Spam}|\textit{cheap}) = \frac{P(\textit{cheap}|\text{Spam}) \cdot P(\text{Spam})}{P(\textit{cheap})} = \frac{\frac{15}{25}\frac{25}{100}}{\frac{25}{100}} = 0.6$$

Un ejemplo: Detección de SPAM

¿Qué pasa si quieremos basar la clasificación en dos palabras buy y cheap?

Si un correo contiene las palabras buy y cheap, ¿cuál es la probabilidad de que sea SPAM? La calculamos directamente:

$$P\left(\text{Spam} \mid \begin{array}{c} \textit{buy} \\ \textit{cheap} \end{array}\right) = \frac{12}{12+0} = \frac{12}{12} = 1$$

Un ejemplo: Detección de SPAM

Si un correo contiene las palabras buy y cheap, ¿cuál es la probabilidad de que sea SPAM? Usando el teorema de Bayes:

$$P\left(\text{Spam} \mid \begin{matrix} buy \\ cheap \end{matrix}\right) = \frac{P\left(\begin{matrix} buy \\ cheap \end{matrix} \mid \text{Spam}\right) P(\text{Spam})}{P(buy, cheap)}$$

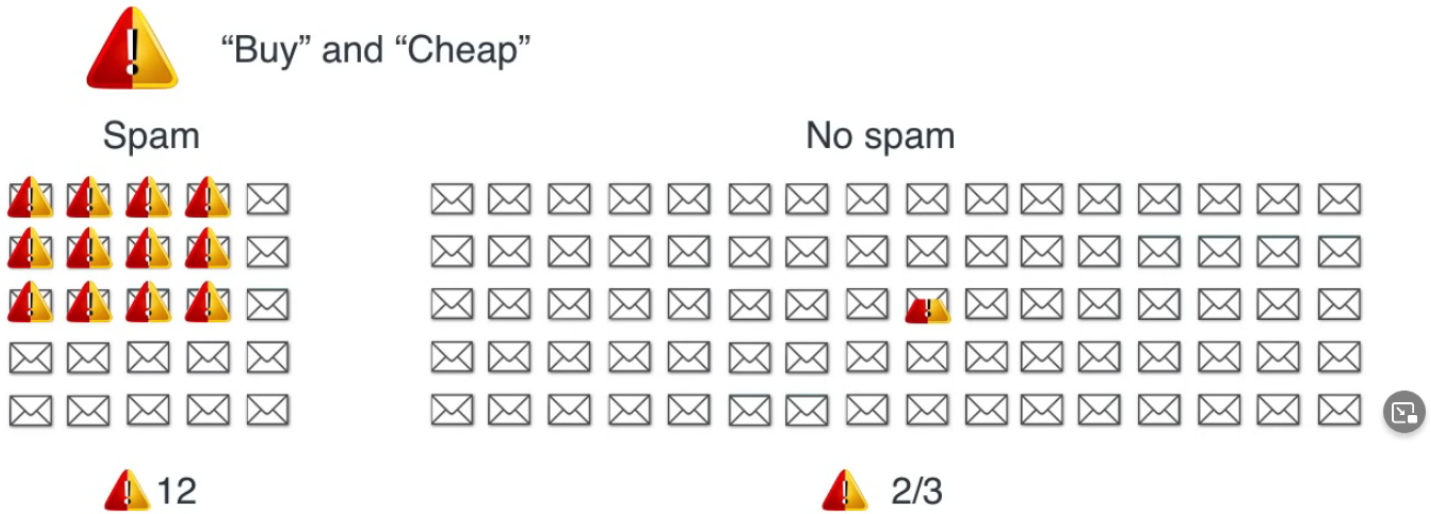
Un ejemplo: Detección de SPAM

Queremos calcular $P(\textit{buy}, \textit{cheap})$. Suponemos independencia de las variables:

$$\begin{aligned} P(\textit{buy}, \textit{cheap}) &= P(\textit{buy}) \cdot P(\textit{cheap}) \\ &= \frac{5}{75} \cdot \frac{10}{75} = \frac{2}{225} = 0.008 \end{aligned}$$

Un ejemplo: Detección de SPAM

Usamos esta hipótesis de independencia en ambas clases.



Si un correo contiene las palabras buy y cheap, ¿cuál es la probabilidad de que sea SPAM?

$$P\left(\text{Spam} \mid \begin{matrix} buy \\ cheap \end{matrix}\right) = \frac{12}{12 + \frac{2}{3}} = \frac{36}{38} = 0.947$$

buy y cheap $\rightarrow$ 94.7%

Table of Contents

Clasificación Naive-Bayes

- Encontrar rasgos en los datos que estén correlacionados con las clases.

Clasificación Naive-Bayes

- Encontrar rasgos en los datos que estén correlacionados con las clases.
- Clasificar estimando $P(\text{clase}_j|x_i)$.

Clasificación Naive-Bayes

- Encontrar rasgos en los datos que estén correlacionados con las clases.
- Clasificar estimando $P(\text{clase}_j|x_i)$.
- Usar el teorema de Bayes

Clasificación Naive-Bayes

- Encontrar rasgos en los datos que estén correlacionados con las clases.
- Clasificar estimando $P(\text{clase}_j|x_i)$.
- Usar el teorema de Bayes
- Usar la hipótesis *naive* de independecia.

Clasificación Naive-Bayes

Diferencias con la implementación en un ejemplo *real*:

- Nos basaremos en todas las palabras del vocabulario.

Clasificación Naive-Bayes

Diferencias con la implementación en un ejemplo *real*:

- Nos basaremos en todas las palabras del vocabulario.
- Contaremos las ocurrencias en cada documento.

Clasificación Naive-Bayes

Diferencias con la implementación en un ejemplo *real*:

- Nos basaremos en todas las palabras del vocabulario.
- Contaremos las ocurrencias en cada documento.
- La hipótesis *naive* de independecia permite usar la distribución multinomial de probabilidad.

$$p(\mathrm{x}|\theta) = \prod_{i=1}^{k} p(x_i|\theta) = n! \cdot \frac{\theta_1^{x_1}}{x_1!} \cdot \ldots \cdot \frac{\theta_k^{x_k}}{x_k!}$$

con $\theta^j = (\theta_1^j, ..., \theta_k^j)$

$$\theta_i^j = \frac{n_i + 1}{|S_j| + k}.$$

Esto último es el suavizado.

Table of Contents

¿Cuándo usar Naive-Bayes?

Algunas consideraciones generales:

- Funciona en clasificación binaria y multiclase.
- Suele ser adecuado para diversos problemas de clasificación de documentos de texto: filtros de spam, análisis de sentimientos, etc.
- Es barato computacionalmente.
- No requiere tantos datos de entrenamiento.
- Tiene la limitante de suponer que las diferentes variables predictoras son independientes.
- Suele ser afectado por el desbalance de clases.

Referencias

Ejemplos ilustrativos:

- https://www.youtube.com/watch?v=HZGCoVF3YvM
- https://www.youtube.com/watch?v=Q8l0Vip5YUw
- https://www.youtube.com/watch?v=l3dZ6ZNFjo0