Clasificación Naive Bayes y Procesamiento del Lenguaje Natural

Introducción

- En esta parte se usarán modelos de Machine Laearning con cadenas de texto sin formato, a esta idea se le denomina como "Procesamiento del Lenguaje Natural".
- Se trata de extraer características de datos tipo texto.

Naive Bayes

- Naive Bayes es como se le conoce a un conjunto de algoritmos que usan el Teorema de Bayes para la clasificación con aprendizaje supervisado.
- El Teorema de Bayes es una fórmula de probabilidad que aprovecha las probabilidades previamente conocidas para definir la probabilidad de que ocurran eventos relacionados.

Naive Bayes

- Métodos Naive Bayes son un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes.

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

Naive Bayes

- Métodos Naive Bayes son un conjunto de algoritmos de aprendizaje supervisado basados en la aplicación del teorema de Bayes.

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

Naive Bayes

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

A y B son eventos

P(A|B) es la probabilidad del evento B dado que B es verdadero

P(B|A) es la probabilidad del evento A dado que A es verdadero

P(A) es la probabilidad de que ocurra A

P(B) es la probabilidad de que ocurra B

Naive Bayes

Imagine la siguiente situación:

- Cada departamento de un edificio tiene un sistema de alarma contra incendios.
- Sin embargo, existen falsas alarmas cuando se detecta humo, pero no proviene de fuego peligroso (por ejemplo el humo proveniente de un horno).

Naive Bayes

Probabilidades asociadas:

- Ocurren llamas de fuego peligroso solamente el 1% de las veces.
- Los detectores de humo no son buenos y se activan el 10% de las veces.
- Cuando existe fuego peligroso, el 95% de las veces las alarmas de humo se activan.

Naive Bayes

Si se activo la alarma de humo al detectar fuego, ¿cuál es la probabilidad que realmente sea fuego peligroso?

- Evento A: Fuego peligroso
- Evento B: Se activa la alarma de humo.
- P(A|B): Probabilidad de fuego peligroso dado que se activo la alarma de humo.
- P(B|A): Probabilidad de que se active la alarma de humo dado fuego peligroso.

Naive Bayes

Si se activo la alarma de humo al detectar fuego, ¿cuál es la probabilidad que realmente sea fuego peligroso?

- Evento A: Fuego peligroso. $P(A) = 0.01$
- Evento B: Se activa la alarma de humo. $P(B) = 0.10$
- $P(A|B)$: Probabilidad de fuego peligroso dado que se activo la alarma de humo. $P(A|B) = ?$
- $P(B|A)$: Probabilidad de que se active la alarma de humo dado fuego peligroso. $P(B|A) = 0.95$

Naive Bayes

- Evento A: Fuego peligroso. P(A) = 0.01
- Evento B: Se activa la alarma de humo. P(B) = 0.10
- P(A|B): Probabilidad de fuego peligroso dado que se activo la alarma de humo. P(A|B) = ?
- P(B|A): Probabilidad de que se active la alarma de humo dado fuego peligroso. P(B|A) = 0.95
- Con el teorema de Bayes: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{0.95 \cdot 0.01}{0.10} = 0.095 = 9.5\%$

Procesamiento del Lenguaje Natural

- Modelo para la probabilidad de pertenecer a una clase dado un vector de características.

$$\mathbf{x} = (x_1, \ldots, x_n)$$

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \quad \Rightarrow \quad p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Procesamiento del Lenguaje Natural

- La regla de la cadena puede reescribir el numerador como una serie de productos de probabilidades condicionales:

$$
\begin{aligned}
p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)
\end{aligned}
$$

Procesamiento del Lenguaje Natural

- Asumimos que todas las características x son mutuamente independientes una de cada otra, por tanto:

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)$$

Procesamiento del Lenguaje Natural

- Luego el modelo de Naive Bayes queda como:

$$
\begin{aligned}
p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\
&\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k) \cdots \\
&\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,,
\end{aligned}
$$

Procesamiento del Lenguaje Natural

Existen muchas variaciones del modelo de Naive Bayes, incluyendo:

- Naive Bayes Multinomial
- Naive Bayes Gaussiano
- Naive Bayes Complemento
- Naive Bayes de Bernoulli
- Naive Bayes Categórico

Procesamiento del Lenguaje Natural

- Nos enfocaremos en Naive Bayes Multinomial, ya que este es el más usado frecuentemente en el contexto del procesamiento del lenguaje natural.
- Imaginemos que queremos crear un sitio web de agregación de reseñas de películas en el que necesitemos clasificar las reseñas de películas en dos categorías: positivas o negativas.

Procesamiento del Lenguaje Natural

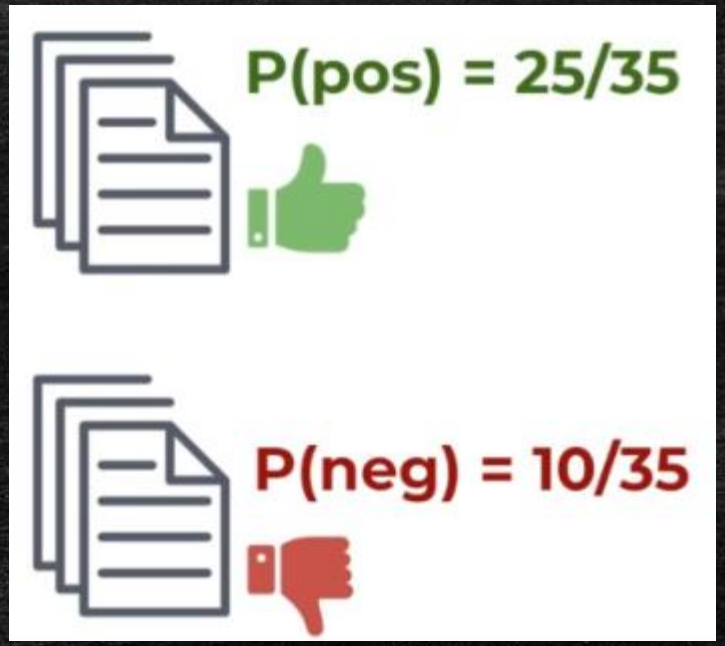
- Basándonos en revisiones anteriores, podemos etiquetarlos manualmente para tener un conjunto de datos etiquetados.
- Luego podemos usar un algoritmo de Machine Learning para clasificar automáticamente un nuevo texto de revisión.
- Bayes Multinomial trabaja muy bien como un modelo de vectorización de conteo (contando la frecuencia de cada palabra en el documento)

Procesamiento del Lenguaje Natural

Documentos separados de cada clase:

Procesamiento del Lenguaje Natural

Probabilidades "previas" para cada clase:



Procesamiento del Lenguaje Natural

Vectorización de conteo en cada clase:

10	2	8	4
movie	actor	great	film

8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Probabilidades Condicionales

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

8	10	0	2
movie	actor	great	film

P(movie|pos) = 10/24 = 0.42
P(actor|pos) = 2/24 = 0.08
P(great|pos) = 8/24 = 0.33
P(film|pos) = 4/24 = 0.17

Procesamiento del Lenguaje Natural

Probabilidades Condicionales

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

8	10	0	2
movie	actor	great	film

P(movie|neg)= 8/20 = 0.4
P(actor|neg) = 10/20 = 0.5
P(great|neg) = 0/20 = 0
P(film|neg) = 2/20 = 0.1

Procesamiento del Lenguaje Natural

Se ha creado una nueva revisión:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

“movie actor”

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Se ha creado una nueva revisión:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

“movie actor”

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Se ha creado una nueva revisión:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

“movie actor”

P(pos)×P(movie|pos)×P(actor|pos)

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Este cálculo es proporcional a P(pos|"movie actor)"

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

"movie actor"

(0.71)×(0.42)×(0.08) = 0.024

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Este cálculo es proporcional a P(pos|"movie actor)"

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

"movie actor"

0.024 ∝ P(pos| "movie actor")

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Lo mismo sucede con las clases negativas:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

“movie actor”

$$P(neg) \times P(movie|neg) \times P(actor|neg)$$

Procesamiento del Lenguaje Natural

Lo mismo sucede con las clases negativas:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

“movie actor”

$(10/35)\times(0.4)\times(0.5) = 0.057$

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Lo mismo sucede con las clases negativas:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

“movie actor”

$0.057 \propto P(neg|\text{“movie actor”})$

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

Procesamiento del Lenguaje Natural

Se comparan ambos resultados:

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

“movie actor”

$0.057 \propto P(\text{neg}|\text{“movie actor”})$
$0.024 \propto P(\text{pos}|\text{“movie actor”})$

Procesamiento del Lenguaje Natural

Se clasifica como una revisión negativa

0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

“movie actor”

$0.057 \propto P(\text{neg}|\text{“movie actor”})$
$0.024 \propto P(\text{pos}|\text{“movie actor”})$

Procesamiento del Lenguaje Natural

¿Qué sucede con las palabras de conteo cero?

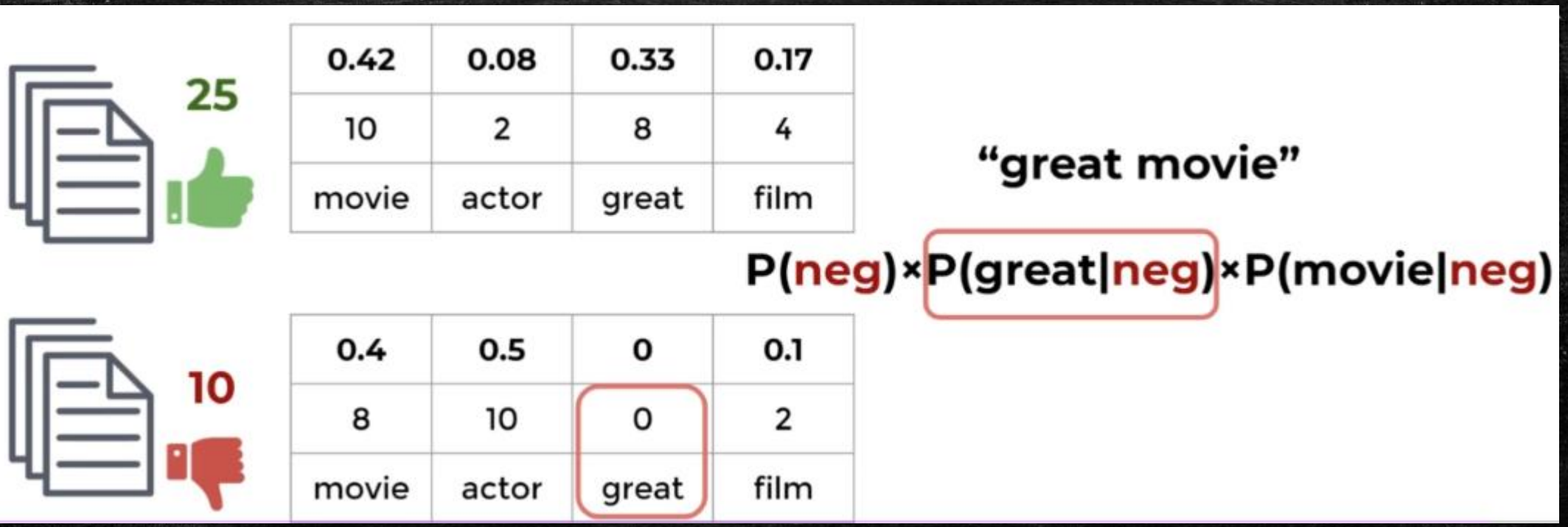
0.42	0.08	0.33	0.17
10	2	8	4
movie	actor	great	film

0.4	0.5	0	0.1
8	10	0	2
movie	actor	great	film

“great movie”

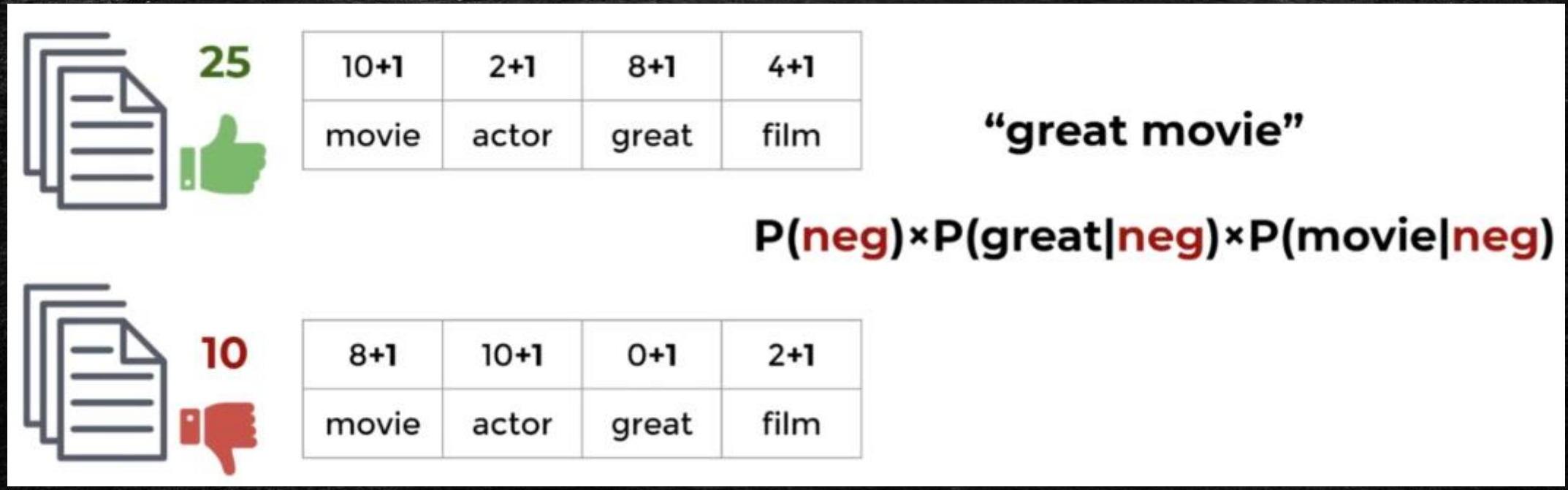
Procesamiento del Lenguaje Natural

La probabilidad es cero sin importar el resto del texto, esto no es muy adecuado



Procesamiento del Lenguaje Natural

Parámetro de suavizado Alfa (Alpha) para agregar al conteo.



Procesamiento del Lenguaje Natural

Se deben recalcular las probabilidades condicionales y seguir con lo descrito previamente.

25

10+1	2+1	8+1	4+1
movie	actor	great	film

10

8+1	10+1	0+1	2+1
movie	actor	great	film

“great movie”

$$P(neg) \times P(great|neg) \times P(movie|neg)$$

Procesamiento del Lenguaje Natural

- Tenga en cuenta cómo un valor alfa más alto se estará más "suavizado", dando a cada palabra una importancia menos distinta.
- La extracción de características es un paso muy importante, ¿existen mejores formas de extraer características que el conteo de la frecuencia de palabras?.

Métodos para extraer características

Los principales métodos para la extracción de características son:

- Vectorización de Conteo
- TF-IDF
  - Term Frecuency – Inverse Document Frecuency