Procesamiento de Lenguaje Natural

Clasificación de textos – Naive Bayes

Dra. Helena Gómez Adorno
helena.gomez@iimas.unam.mx

Dr. Orlando Ramos
orlando.ramos@aries.iimas.unam.mx

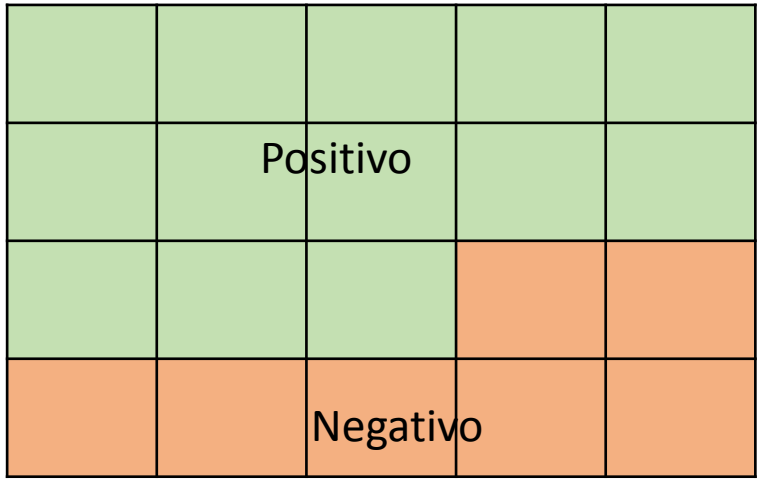
Correo del curso:
pln.cienciadedatos@gmail.com

Aprendizaje Probabilístico: Clasificadores Naïve Bayes

- Es una familia de clasificadores probabilísticos simple.
- Estos clasificadores se denominan "ingenuos" porque asumen que las características son condicionalmente independientes, dada la clase.
  - Asumen que, para todas las instancias de una clase determinada, las características tienen poca o ninguna correlación entre sí.
- Aprendizaje y predicción altamente eficientes.
- Pero el rendimiento de la generalización puede ser peor que los métodos de aprendizaje más sofisticados.
- Puede ser competitivo para algunas tareas.

Probabilidad y regla de Bayes

Dataset: Tweets



A -> Tweet Positivo

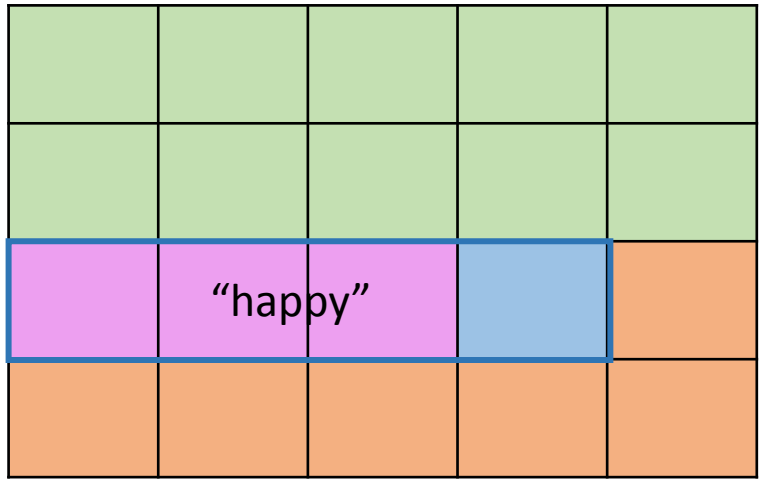
B -> Tweet Negativo

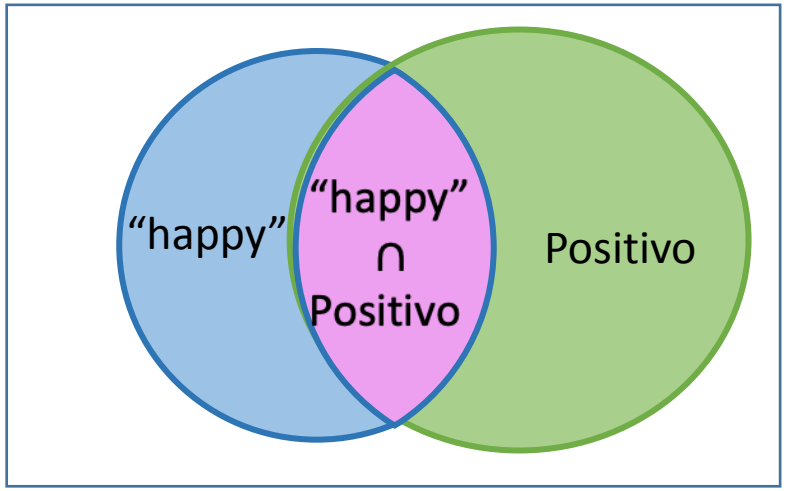
$P(A) = N_{pos}/N = 13/20 = 0.65$

$P(B) = 1 - P(A) = 0.35$

Probabilidad y regla de Bayes

- Para calcular la probabilidad de que ocurra un determinado evento, se toma el **recuento** de ese evento específico y se divide por la suma de todos los eventos. Además, la suma de todas las probabilidades debe ser igual a 1.
- Para calcular la probabilidad de que sucedan 2 eventos, como “happy" y "positivo" en la imagen de abajo, estaría mirando la intersección o superposición de eventos. En este caso, los cuadros rosa y azul se superponen en 3 cuadros. Entonces la respuesta es 3/20.

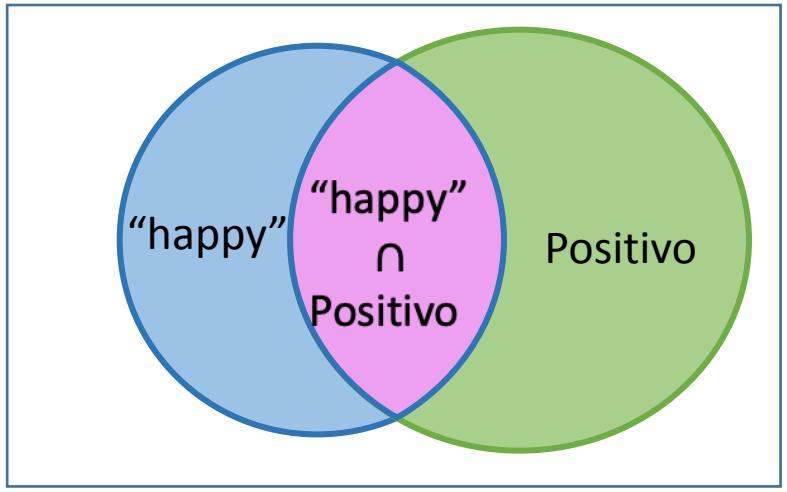




$P(A \cap B) = P(A, B) = 3/20 = 0.15$

Regla de Bayes

- Las probabilidades condicionales nos ayudan a reducir el espacio de búsqueda muestral. Por ejemplo, dado que un evento específico ya sucedió, es decir, sabemos que la palabra es “happy”:



$$P(Positivo \mid \text{“}happy\text{”}) = \frac{P(Positivo \cap \text{“}happy\text{”})}{P(\text{“}happy\text{”})}$$

Regla de Bayes (cont.)

- Entonces solo buscaría en el círculo azul de arriba. El numerador será la parte roja y el denominador será la parte azul. Esto nos lleva a concluir lo siguiente:

$$P(Positivo \mid "happy") = \frac{P(Positivo \cap "happy")}{P\ ("happy")}$$

$$P("happy" \mid Positivo) = \frac{P("happy" \cap Positivo)}{P\ (Positivo)}$$

- Sustituyendo el numerador en el lado derecho de la primera ecuación, obtienes lo siguiente:

$$P(Positivo \mid "happy") = P("happy" \mid Positivo) \times \frac{P(Positivo)}{P\ ("happy")}$$

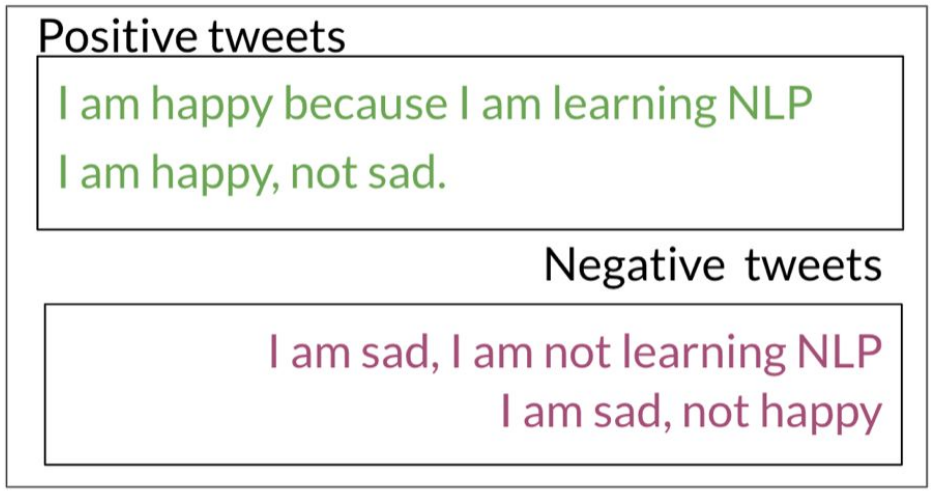
- Tener en cuenta que multiplicamos por P(positivo) para asegurarnos de no cambiar nada. Eso concluye la regla de Bayes que se define como P (X|Y) = P (Y|X)P(X) / P (Y)

Algoritmo Naive Bayes Supervisado

A continuación se listan los pasos que hay que realizar para poder utilizar el algoritmo Naive Bayes en problemas de clasificación:

1. Convertir el conjunto de datos en una tabla de frecuencias.
2. Crear una tabla de probabilidades calculando las correspondientes a que ocurran los diversos eventos.
3. La ecuación de inferencia Naive Bayes se usa para calcular la probabilidad posterior de cada clase.
4. La clase con la probabilidad posterior más alta es el resultado de la predicción.

Convertir el conjunto de datos en una tabla de frecuencias.



word	Pos	Neg
I	3	3
am	3	3
happy	2	1
because	1	0
learning	1	1
NLP	1	1
sad	1	2
not	1	2
N_{class}	13	13

Crear una tabla de probabilidad

$P(w_i|clase)$

word	Pos	Neg
I	3	3
am	3	3
happy	2	1
because	1	0
learning	1	1
NLP	1	1
sad	1	2
not	1	2
N_{class}	13	13

P(I|postivo)=3/13

word	Pos	Neg
I	0.24	0.25
am	0.24	0.25
happy	0.15	0.08
because	0.08	0
learning	0.08	0.08
NLP	0.08	0.08
sad	0.08	0.17
not	0.08	0.17
Suma	**1**	**1**

Inferencia bayesiana

Tweet: I am happy today; I am learning.

$$\prod_{i=1}^{m} \frac{P(w_i|pos)}{P(w_i|neg)} = \frac{0.14}{0.10} = 1.4 \quad > 1$$

$$\frac{0.20}{0.20} * \frac{0.20}{0.20} * \frac{0.14}{0.10} * \frac{0.20}{0.20} * \frac{0.20}{0.20} * \frac{0.10}{0.10}$$

word	Pos	Neg
I	0.20	0.20
am	0.20	0.20
happy	0.14	0.10
because	0.10	0.05
learning	0.10	0.10
NLP	0.10	0.10
sad	0.10	0.15
not	0.10	0.15

- Un puntaje mayor que 1 indica que la clase es positiva, sino es negativa

Laplacian Smoothing

- Calculamos la probabilidad de una palabra dada una clase de la siguiente manera:

$$P\left(\mathrm{w_i} \mid \text{ class }\right) = \frac{\mathrm{freq}\left(\mathrm{w_i}, \text{ class }\right)}{\mathrm{N_{class}}} \qquad \text{class } \in \{ \text{ Positive, Negative } \}$$

- Si una palabra no aparece en el entrenamiento, automáticamente obtiene una probabilidad de 0, para solucionar esto agregamos suavizado:

$$P\left(\mathrm{w_i} \mid \text{class}\right) = \frac{\mathrm{freq}(\mathrm{w_i}, \text{ class }) + 1}{(\mathrm{N_{class}} + \mathrm{V})}$$

- Agregamos un **1** en el numerador y, dado que hay **V** palabras para normalizar, agregamos V en el denominador.

Laplacian Smoothing

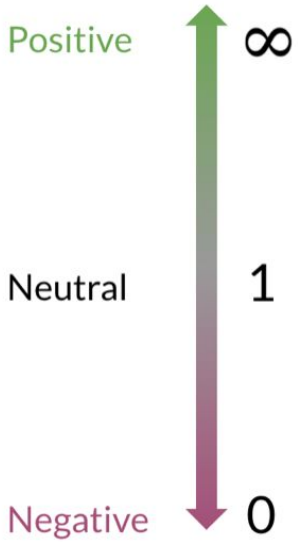
word	Pos	Neg
I	3	3
am	3	3
happy	2	1
because	1	0
learning	1	1
NLP	1	1
sad	1	2
not	1	2
N_{class}	13	13

P(I|postivo)=3 +1 / 13 + 8

word	Pos	Neg
I	0.19	0.20
am	0.19	0.20
happy	0.14	0.10
because	0.10	0.05
learning	0.10	0.10
NLP	0.10	0.10
sad	0.10	0.15
not	0.10	0.15
Sum	1	1

Logaritmo de la probabilidad (log likelihood)

- Para calcular el logaritmo de la probabilidad, necesitamos obtener las proporciones y usarlas para calcular una puntuación que nos permitirá decidir si un tweet es positivo o negativo. Cuanto mayor sea la proporción, más positiva será la palabra:



word	Pos	Neg	ratio
I	0.19	0.20	
am	0.19	0.20	
happy	0.14	0.10	
because	0.10	0.05	
learning	0.10	0.10	
NLP	0.10	0.10	
sad	0.10	0.15	
not	0.10	0.15	

$$\text{ratio}(w_i) = \frac{P(w_i \mid \text{Pos})}{P(w_i \mid \text{Neg})}$$

$$\approx \frac{freq(w_i, 1) + 1}{freq(w_i, 0) + 1}$$

Logaritmo de la probabilidad (log likelihood)

Probabilidad previa (prior)

Likelihood

- Para hacer inferencia, se calcula:

$$\frac{P(pos)}{P(neg)} \prod_{i=1}^{m} \frac{P(w_i|pos)}{P(w_i|neg)} > 1$$

- A medida que ***m*** aumenta, podemos tener problemas de flujo numérico, por lo que introducimos el logaritmo:

$$\log \left(\frac{P(pos)}{P(neg)} \prod_{i=1}^{n} \frac{P(w_i|pos)}{P(w_i|neg)} \right) \Rightarrow \log \frac{P(pos)}{P(neg)} + \sum_{i=1}^{n} \log \frac{P(w_i|pos)}{P(w_i|neg)}$$

- **Recuerda**: log(a*b) = log(a) + log(b)

Logaritmo de la probabilidad (log likelihood)

- Además, introducimos λ de la siguiente manera:

doc: I am happy because I am learning.

$$\lambda(w) = log\frac{P(w|pos)}{P(w|neg)}$$

$$\lambda(\text{happy}) = log\frac{0.09}{0.01} \approx 2.2$$

word	Pos	Neg	λ
I	0.05	0.05	0
am	0.04	0.04	0
happy	0.09	0.01	
because	0.01	0.01	
learning	0.03	0.01	
NLP	0.02	0.02	
sad	0.01	0.09	
not	0.02	0.03	

- Tener el diccionario λ ayudará mucho a la hora de hacer inferencias.

Logaritmo de la probabilidad (log likelihood)

- Una vez que calculó el diccionario λ, resulta sencillo hacer inferencia:

doc: I am happy because I am learning.

$$\sum_{i=1}^{m} log \frac{P(w_i|pos)}{P(w_i|neg)} = \sum_{i=1}^{m} \lambda(w_i)$$

log likelihood = 0 + 0 + 2.2 + 0 + 0 + 0 + 1.1 = **3.3**

word	Pos	Neg	λ
I	0.05	0.05	0
am	0.04	0.04	0
happy	0.09	0.01	2.2
because	0.01	0.01	0
learning	0.03	0.01	1.1
NLP	0.02	0.02	0
sad	0.01	0.09	-2.2
not	0.02	0.03	-0.4

- Como puede ver arriba, como 3.3> 0, clasificaremos el *tweet* como positivo. Si obtuviéramos un número negativo, lo habríamos clasificado en la clase negativa.

Tipos de clasificadores Naive Bayes

- **Bernoulli**: características binarias (por ejemplo, presencia / ausencia de palabras)
- **Multinomial**: características discretas (por ejemplo, recuento de palabras)
- **Gaussiano**: características continuas / de valor real
  - Estadísticas calculadas para cada clase:
    - Para cada característica: media, desviación estándar

Ventajas y Desventajas

Ventajas

- Fácil de comprender
- Estimación de parámetros simple y eficiente
- Funciona bien con datos de alta dimensión
- Suele ser útil como comparación de referencia con métodos más sofisticados.

Desventajas

- La suposición de que las características son condicionalmente independientes dada la clase no es realista.
- Como resultado, otros tipos de clasificadores suelen tener un mejor rendimiento de generalización.
- Sus estimaciones de confianza para las predicciones