

Procesamiento de Lenguaje Natural

Introducció

en el PLN

Clasificador Naive Bayes

## Procesamiento de Lenguaje Natural Naive Bayes

Mauricio Toledo-Acosta mauricio.toledo@unison.mx

Departamento de Matemáticas Universidad de Sonora



Procesamiento de Lenguaje Natural

Introducción

Clasificaciór en el PLN

Clasificador Naive Bayes

# Section 1 Introducción



#### Referencias

Procesamiento de Lenguaje Natural

Introducció

Ciasificación en el PLN

Clasificador Naive Bayes



 Chapter 4. Jurafsky, D., Martin, J. H. (2019). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.



Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes

# Section 2 Clasificación en el PLN



#### Clasificación en PLN

Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

Naive Bayes

Consideremos la tarea de clasificación en PLN, es decir, asignar categorías a textos.

 Análisis de sentimientos: La extracción del sentimiento, es decir, la orientación positiva o negativa que el escritor expresa hacia algún objeto.

Reseña de una película, $\ \leftrightarrow$ un libro o un producto		El sentimiento del au- tor hacia el producto	
Editorial o un texto político	$\leftrightarrow$	El sentimiento hacia un candidato o una acción política.	
Textos en redes so- ciales	$\leftrightarrow$	El estado de animo.	



#### Clasificación en PLN

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes

- Detección de SPAM
- Identificación de idioma
- Atribución de autoría
- Detección de tópicos o temática
- Predicción de la siguiente palabra



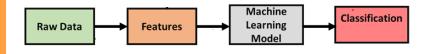
## Clasificación en el Machine Learning

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Naive Bayes



¿Quienes son las features en el NLP?

- Métodos clásicos: BOW, TF-IDF, ...
- Embeddings: Redes Neuronales, LLMs.



Procesamiento de Lenguaje Natural

Introducció

en el PLN

Clasificador Naive Bayes

## Section 3

Clasificador Naive Bayes



#### Clasificación

Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

Clasificador Naive Bayes

#### 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.



Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

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

$$P(\operatorname{clase}_j|d)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase j si observamos el documento d?



Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

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

$$P(\operatorname{clase}_j|d)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase j si observamos el documento d?

En un problema de clasificación binaria, predecimos que un documento *d* pertenece a la clase 0 si

$$P(\mathsf{clase}_0|d) > P(\mathsf{clase}_1|d).$$



Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

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

$$P(\operatorname{clase}_j|d)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase j si observamos el documento d?

En un problema de clasificación binaria, predecimos que un documento *d* pertenece a la clase 1 si

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



Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

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

$$P(\operatorname{clase}_j|d)$$

Es decir, ¿cuál es la probabilidad de que estemos en la clase j si observamos el documento d?

En general, en un problema de clasificación multiclase con clases  $C = \{c_0, ..., c_m\}$ , la clase predicha es

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d).$$



#### Formulación

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes El objetivo de la clasificación Bayesiana es determinar

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d)$$

transformado esta expresión en algo más fácil de calcular. Para esto, usamos la regla de Bayes:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)}.$$

En todas las clases  $c \in C$ , el término P(d) no cambia, entonces:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c).$$



## Formulación

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes  $\hat{c} = \operatorname{argmax} \underbrace{P(d|c)}^{\text{likelihood prior}} \underbrace{P(c)}.$ 

Representamos cada documento d con las features  $w_1, ..., w_n$  (típicamente, las palabras del documento).

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1, w_2, ..., w_n | c) P(c).$$

Usando la suposición *naive* de que los eventos  $w_1$ , ...,  $w_n$  son independientes:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1|c)P(w_2|c) \cdots P(w_n|c)P(c).$$



#### Formulación

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes También estamos haciendo la suposición que la posición de la palabra no importa, sólo su aparición.

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(w_1|c)P(w_2|c) \cdots P(w_n|c)P(c).$$



# Aspectos técnicos

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Baye Para evitar el underflowing, usando logaritmos:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \left( \log P(w_1|c) + \log P(w_2|c) + \cdots + \log P(w_n|c) + P(c) \right)$$



#### Entrenamiento

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes Entrenar el modelo Naive Bayes quiere decir estimar las probabilidades P(c) y  $P(w_j|c)$ . Esto lo hacemos contando:

$$P(c) = rac{N_c}{N_{ ext{total}}}$$
 $P(w_i|c) = rac{ ext{count}(w_i, c) + 1}{\sum_{w \in V} ( ext{count}(w, c) + 1)}$ 

V es el vocabulario del corpus: el conjunto de todas las palabras que aparecen en el corpus de entrenamiento (en todas las clases).



#### Consideraciones adicionales

Procesamiento de Lenguaje Natural

Introducciói

Clasificación en el PLN

Clasificador Naive Bayes

- ¿Qué hacemos con las palabras desconocidas en el conjunto de prueba?
- ¿Qué hacemos con las stopwords?
- ¿Cómo manejamos las negaciones? didn't like this movie ,  $\rightarrow$  didn't NOT\_like but I ... NOT\_this NOT\_movie , but I ...
- Podemos incluir n-gramas, palabras de acuerdo al dominio de la tarea.
- Podemos incluir conteos binarios (presencia o ausencia de palabras).



# Ejemplo

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun



# ¿Por qué Naive Bayes Multinomial?

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes 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:

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

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



#### La distribución multinomial

Procesamiento de Lenguaje Natural

Introducció

Clasificación en el PLN

Clasificador Naive Bayes

#### 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.



# Ejercicio

Procesamiento de Lenguaje Natural

Introducció

Clasificaciór en el PLN

Clasificador Naive Bayes Considera los siguientes reviews cortos de películas:

fun, copuple, love, love	comedy
fast, furious, shoot	action
copuple, fly, fast, fun, fun	comedy
furious, shoot, shoot, fun	action
fly, fast, shoot, love	action

Usando Naive Bayes y un suavizado *add-1*, predice el genero de la pelicula del siguiente review:

fast, couple, shoot, fly, live