



Procesamiento de Lenguaje Natural

Naive Bayes

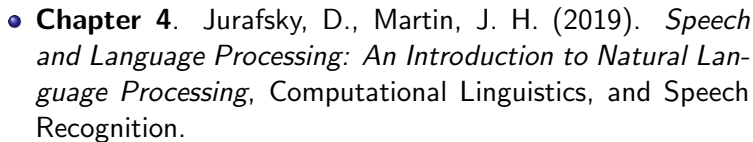
Mauricio Toledo-Acosta
mauricio.toledo@unison.mx

Departamento de Matemáticas
Universidad de Sonora



Section 1

Introducción





Section 2

Clasificación en el PLN



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, un libro o un producto	↔	El sentimiento del autor hacia el producto
Editorial o un texto político	↔	El sentimiento hacia un candidato o una acción política.
Textos en redes sociales	↔	El estado de ánimo.



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



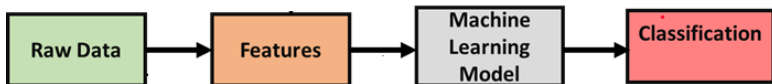
Clasificación en el Machine Learning

Procesamiento
de Lenguaje
Natural

Introducción

Clasificación
en el PLN

Clasificador
Naïve Bayes



¿Quiénes son las *features* en el NLP?

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



Section 3

Clasificador Naive Bayes



Clasificación

Procesamiento
de Lenguaje
Natural

Introducción

Clasificación
en el PLN

Clasificador
Naive Bayes

Clasificador Naive Bayes

Naive Bayes es un algoritmo 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.



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

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

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



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

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

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



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

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

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



Enfoque probabilístico

Procesamiento
de Lenguaje
Natural

Introducción

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

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

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 $C = \{c_0, \dots, c_m\}$, la clase predicha es

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



El objetivo de la clasificación Bayesiana es determinar

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

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

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

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

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



$$\hat{c} = \operatorname{argmax}_{c \in C} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}.$$

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

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

Usando la suposición *naive* de que los eventos w_1, \dots, w_n son independientes:

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



Formulación

Procesamiento
de Lenguaje
Natural

Introducción

Clasificación
en el PLN

Clasificador
Naive Bayes

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

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



Aspectos técnicos

Procesamiento
de Lenguaje
Natural

Introducción

Clasificación
en el PLN

Clasificador
Naive Bayes

Para evitar el underflowing, usando logaritmos:

$$\hat{c} = \operatorname{argmax}_{c \in C} (\log P(w_1|c) + \log P(w_2|c) + \dots + \log P(w_n|c) + P(c))$$



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

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

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





Ejemplo

Procesamiento de Lenguaje Natural

Introducción

Clasificación en el PLN

Clasificador Naïve 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

Introducción

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:

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

para enteros no negativos x_1, \dots, x_k .



- 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.



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