TEMA 9: MODELO BAG OF WORDS ALGORITMO NAÏVE BAYES

Modelo Bag of Words

- ¿Qué pasa si queremos aplicar técnicas de aprendizaje sobre documentos de texto?
  - Problema: los documentos de texto no tienen estructura
    - Número de palabras diferentes
    - Términos ambiguos
    - Falta de contexto
- Para realizar aprendizaje automático
  - Se necesita un conjunto de entrenamiento
    - Cada ejemplo es un documento
    - Cada documento estará caracterizado por sus palabras
- Necesidad de técnicas que extraigan información estructurada a partir de texto

Modelo Bag of Words

- El modelo Bag of Words es una técnica para extraer información a partir de texto
  - Considera cada documento como una bolsa (conjunto) de palabras
  - Realiza dos suposiciones irreales
    - Las palabras son independientes
    - El orden de las palabras es irrelevante
- A pesar de su simpleza y de sus suposiciones
  - Es una técnica muy eficaz y que ha demostrado funcionar bien para realizar aprendizaje automático

Modelo Bag of Words

- Al conjunto de documentos de texto disponibles para aprender el modelo Bag of Words se le llama corpus
- Cada documento del corpus se representa como un vector multi dimensional
  - Cada dimensión es un término único del corpus
    - Un término puede ser una palabra o un conjunto de palabras
  - El número de términos determina la dimensión del vector
    - El número de variables del problema
- Por tanto, cada documento es un vector con tantos elementos como términos del corpus
  - El valor de cada elemento representa el peso (relevancia) del término en el documento
  - Si un corpus tiene m términos $(t_i, i = \{1, \dots, m\})$ el documento d del corpus estará representado por el vector $d = \{w_1, \dots, w_m\}$
    - Donde w_i es el peso asociado al término t_i

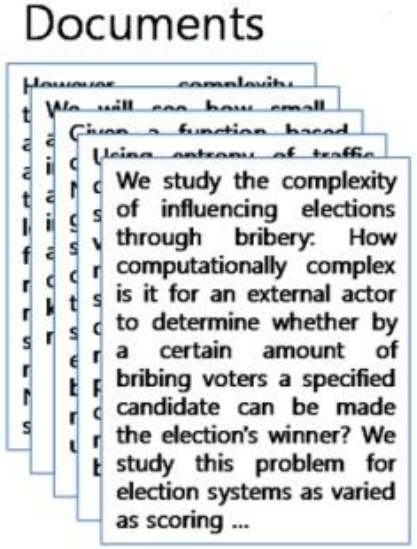
Modelo Bag of Words

- Ejemplo

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
...	
Zaire	0

Modelo Bag of Words

- El modelo bag of Words se representa con una matriz de dimensiones número de documentos (N) por número de términos (m)
  - Cada fila representa un documento
  - Cada columna representa un término
  - La celda $\{i, j\}$ representa el peso del término j en el contexto del documento i



	complexity	algorithm	entropy	traffic	network
D1	2	3	1	0	0
D2	0	0	0	2	1
D3	3	0	0	3	4
D4	2	4	2	0	0
D5	3	4	0	0	0

Modelo Bag of Words

- Antes de crear esta matriz se deben pre-procesar los documentos
  - Objetivo: reducir el número de términos del corpus
- **Normalización**: transformar varias formas del mismo término a un formato común
  - Ejemplo: Apple, apple, APPLE → apple
  - Ejemplo: Intelligent Systems, Intelligent-systems → intelligent systems
  - Proceso
    - Eliminación de signos de puntuación (puntos, guiones, comas, etc..)
    - Pasar el texto a minúsculas
    - Usar diccionarios de sinónimos (WordNet) para agrupar términos que sean sinónimos
      - Ejemplo: automobile, car → vehicle

Modelo Bag of Words

- **Eliminación de términos con frecuencias muy altas o muy bajas**
  - Los términos con frecuencias muy altas (aparecen muchas veces) componen una gran proporción del total de palabras pero no tienen mucha utilidad semántica
    - Ejemplos: the, a, an, we, do, to
  - Por contra, los términos con frecuencias muy bajas son ricos semánticamente pero son muy raros
    - Ejemplo: dextrosinistral
  - El resto de términos son los que mejor representan al corpus y, por tanto, deben ser incluidos en la matriz

Modelo Bag of Words

- **Eliminación de las llamadas stop-words**
  - Las stop-words son palabras que probablemente sean irrelevantes para el análisis del corpus
  - Aquellas que por sí mismas no poseen información
  - Se estima que componen en torno al 20-30% del corpus
  - No hay listas de stop-words únicas
    - Unas de las más comunes se pueden encontrar en http://www.ranks.nl/stopwords
- Posibles problemas de eliminar las stop-words
  - Pérdida del significado original y la estructura del texto
    - Ejemplo: this is not a good option → option
    - Ejemplo: to be or not to be → null

Modelo Bag of Words

- **Reducción de palabras a su raíz**
  - Objetivo: reducir la variabilidad de términos reduciéndolos a su forma básica (o raíz)
  - Técnicas:
    - Stemming: cortar las terminaciones de las palabras sin considerar las características lingüísticas de las palabras
      - Ejemplo: argue, argued, argues, arguing → argu
    - Lemmatization: reducir las palabras a su forma básica teniendo en cuenta el vocabulario y sus características morfológicas
      - Ejemplo: argue, argued, argues, arguing → argue

Modelo Bag of Words

- Una vez pre-procesado el corpus se conoce la dimensión de los vectores que representan a los documentos
  - Se debe calcular el peso de cada término
  - Técnicas
    - Pesos binarios (Binary weights)
    - FT: Frecuencia de los Términos (Term Frequency)
    - IFD: Inversa de la Frecuencia en los Documentos (Inverse Document Frequency)
    - FT-IFD: en inglés este proceso se conoce como TF-IDF

Modelo Bag of Words

- **Pesos binarios**
  - Los pesos toman el valor 0 o 1 y representan la presencia o ausencia del término en el documento
  - Ejemplo
    - D1: Text mining is to identify useful information in the text
    - D2: Useful information is mined from text
    - D3: Apple is delicious

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious	in	the
D1	1	1	1	1	0	1	1	1	0	0	0	1	1
D2	1	1	0	0	1	1	1	0	1	0	0	0	0
D3	0	0	0	0	0	1	0	0	0	1	1	0	0

Modelo Bag of Words

- **Frecuencia de los términos**
  - Representa la frecuencia de un término en un documento
    - Ocurrencias: Número de apariciones (conteo)
  - Idea: los términos con más apariciones serán más importantes en ese documento
  - Problema: los documentos más largos tendrán conteos más grandes
    - Frecuencia: ocurrencias / número de palabras de un documento
  - Ejemplo
    - D1: Text mining is to identify useful information in the text
    - D2: Useful information is mined from text
    - D3: Apple is delicious

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious	in	the
D1	2/10	1/10	1/10	1/10	0	1/10	1/10	1/10	0	0	0	1/10	1/10
D2	1/6	1/6	0	0	1/6	1/6	1/6	0	1/6	0	0	0	0
D3	0	0	0	0	0	1/3	0	0	0	1/3	1/3	0	0

Modelo Bag of Words

- **Inversa de la frecuencia en los documentos**
  - Idea: asignar pesos más grandes a términos no comunes en el corpus
    - Tienen más poder de diferenciación
  - Se calcula a partir de todo el corpus
    - Describe el corpus globalmente, no a los documentos individualmente
  - Ecuación
    - $IFD(t) = 1 + \log\left(\frac{N}{df(t)}\right)$
      - t es el término para el que se calcula el valor
      - N es el número de documentos del corpus
      - $df(t)$ es el número de documentos que contienen el término t

Modelo Bag of Words

- **Inversa de la frecuencia en los documentos**
  - Ejemplo
    - D1: Text mining is to identify useful information in the text
    - D2: Useful information is mined from text
    - D3: Apple is delicious

	text	Information	identify	mining	mined	is	useful	to	from	apple	delicious	in	the
N	3	3	3	3	3	3	3	3	3	3	3	3	3
df(t)	2	2	1	1	1	3	2	1	1	1	1	1	1
IFD	1.41	1.41	2.10	2.10	2.10	1	1.41	2.10	2.10	2.10	2.10	2.10	2.10

Modelo Bag of Words

- **FT-IFD**
  - Idea: valorar los términos que no son muy comunes en el corpus (IFD relativamente alto) pero que tienen un nivel de frecuencia razonable (FT relativamente alto)
- Es el método más habitual de asignar los pesos en el modelo Bag of Words
  - Fórmula general
    - $FT - IDF(t) = FT(t) * IDF(t)$
  - Fórmula comúnmente utilizada
    - $FT - IDF(t) = FT(t) * \log\left(\frac{N}{df(t)}\right)$

Modelo Bag of Words

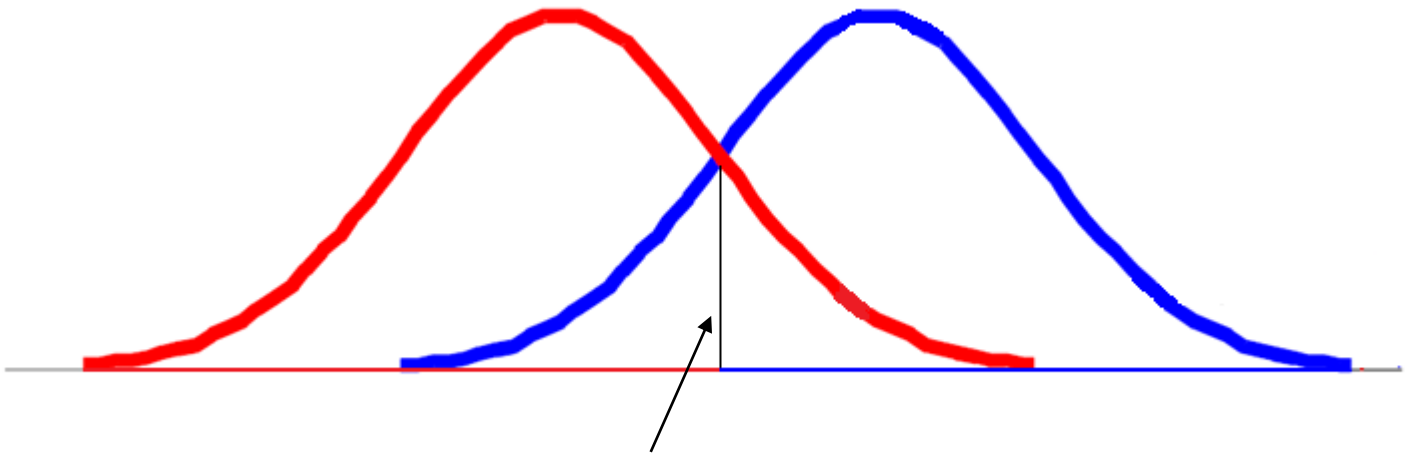
- Ventajas
  - Intuitivo
  - Fácil de implementar
  - Ha probado empíricamente ser muy eficaz
- Inconvenientes
  - Está basado en suposiciones no realistas
  - Ajustar los parámetros del modelo es costoso (validación cruzada)
    - Stop-words a utilizar
    - Umbrales de eliminación de términos en base a frecuencias
    - Método de cálculo de pesos
    - Método de obtención de los términos
      - Palabra simple (unigrama)
      - Conjuntos de dos palabras (bigramas)
      - Etc…

Clasificador Naïve Bayes

- Un clasificador que se aplica habitualmente para resolver problemas de clasificación a partir de texto
- Clasificador Naïve Bayes
  - Basado en el teorema de Bayes (estadístico)
  - Clasifica el ejemplo en la clase asociada a la mayor probabilidad
  - Soporte multi-clase nativo

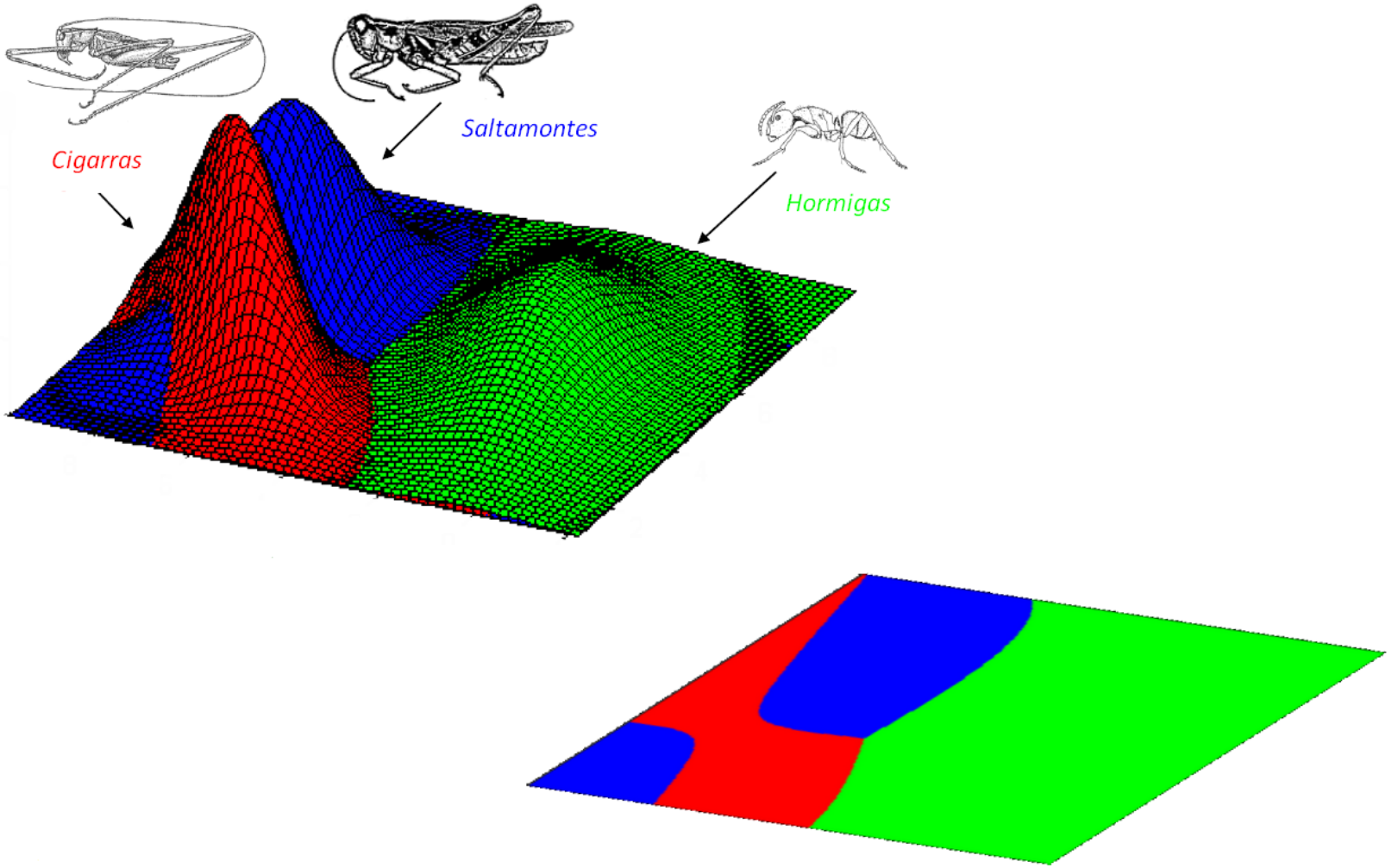
Thomas Bayes
1702 - 1761

Frontera de decisión: idea intuitiva



- La frontera de decisión se traza donde las probabilidades a posteriori de cada clase $p(c_j|x)$ son iguales

Frontera de decisión: idea intuitiva



Clasificador Naïve Bayes

- Método de clasificación MAP: Mayor probabilidad A Posteriori

$$c_{NB} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{n} p(x_i|c_j)$$

  - C es el conjunto de clases del problema
  - n es el número de atributos del problema
  - x_i es el ejemplo a clasificar
  - c_j es la clase j-ésima
  - $p(c_j)$ es la probabilidad de tener ejemplo de la clase j a priori
    - Número de ejemplos de la clase j entre en número total de ejemplos
  - $p(x_i|c_j)$ es la probabilidad de que el atributo i tenga el valor x_i para la clase c_j

Clasificador Naïve Bayes

- Cálculo de $p(x_i|c_j)$
  - Si el atributo es categórico:
    - Frecuencia relativa: $\frac{n_{x_i \wedge c_j}}{n_{c_j}}$
      - Porcentaje de ejemplos cuya clase es c_j que tienen el valor x_i en el i-ésimo atributo
      - Problema: puede dar ceros
        - Ningún ejemplo de la clase C_j con valor x_i
      - Solución

$$\frac{n_{x_i \wedge c_j} + mp}{n_{c_j} + m}$$

$p = \frac{1}{k}$ siendo k el número de valores del atributo

m: constante, normalmente se asigna a 1

  - Si el atributo es continuo:
    - Estimación asumiendo una distribución Gaussiana para lo que se necesita la media (μ_{i,c_j}) y la desviación estándar (σ_{i,c_j}) de cada clase:

$$p(x_i|c_j) = \frac{1}{\sqrt{2 * \pi} * \sigma_{i,c_j}} * e^{-\frac{(x_i - \mu_{i,c_j})^2}{2 * \sigma_{i,c_j}^2}}$$

Clasificador Naïve Bayes

- Algoritmo de aprendizaje
  - Dado un conjunto de ejemplos S:
    - Para cada clase c_j
      - Calcular su probabilidad a priori en S → $p(c_j)$
      - Para cada atributo *k*
        - Si es numérico: calcular media y desviación de los valores: μ_{k,c_j}, σ_{k,c_j}
        - Si es categórico: para cada valor *i* calcular su probabilidad → $p(x_{ik}|c_j)$
- Algoritmo de clasificación
  - Aplicar $c_{NB} = \arg\max_{c_j \in C} p(c_j) \prod_{i=1}^{n} p(x_i|c_j)$

Clasificador Naïve Bayes

- Muy rápido de entrenar
- Muy rápido al clasificar
- Rendimiento competitivo
- Aplicaciones exitosas: filtros spam