“Naïve Bayes Classifier”

Papers We Love Bucharest

Stefan Alexandru Adam

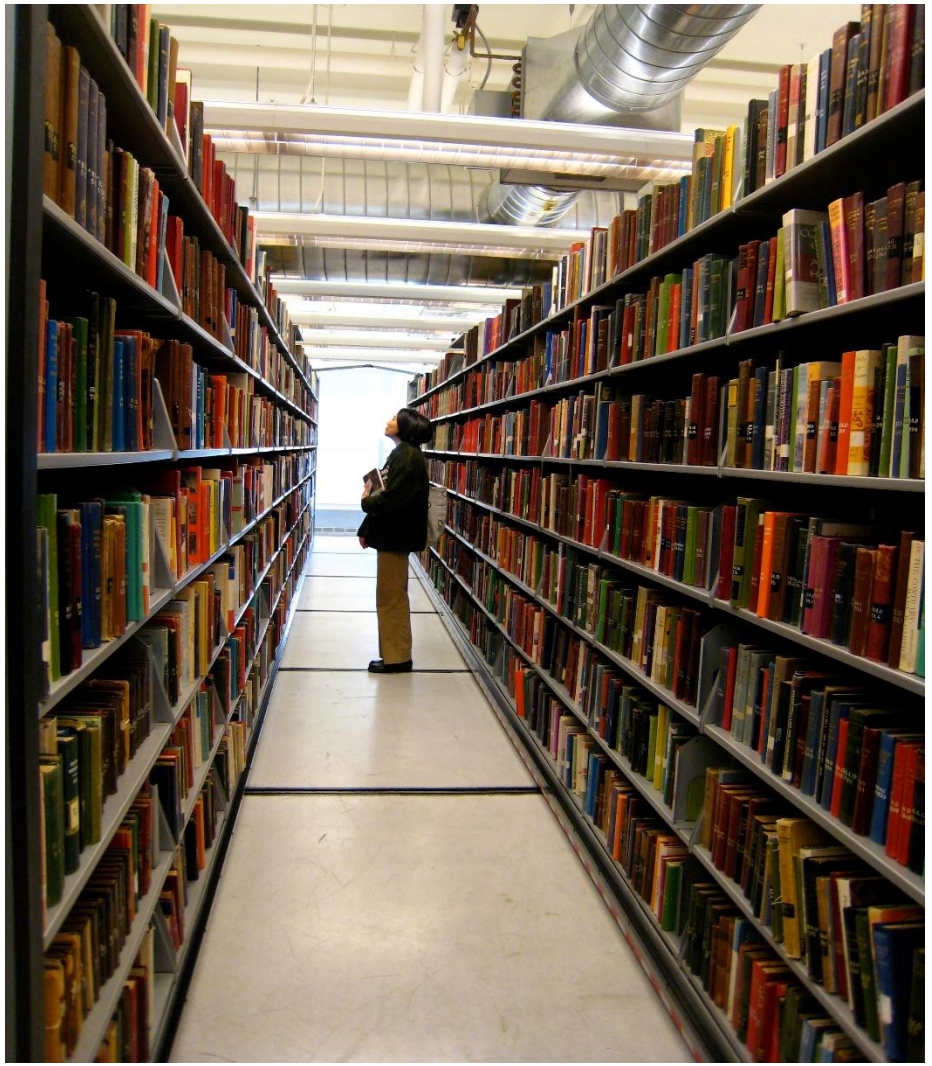
28th of March 2016

TechHub

“Library Problem”

The technique of searching was the following:

- All documents were read by a person called “Indexer” and then labeled with one or more keywords
- All keywords with their related addresses in the library were kept together forming the so called “Library vocabulary”. The keywords were stored in digital format
- Based on a query, an automatic device will search through keywords and returns the documents locations
  - *Documents returned were equally ranked, no distinction between them*
  - *There was no middle ground between tags*
  - *In 1960 it was observed that documents growing rate was exponential*

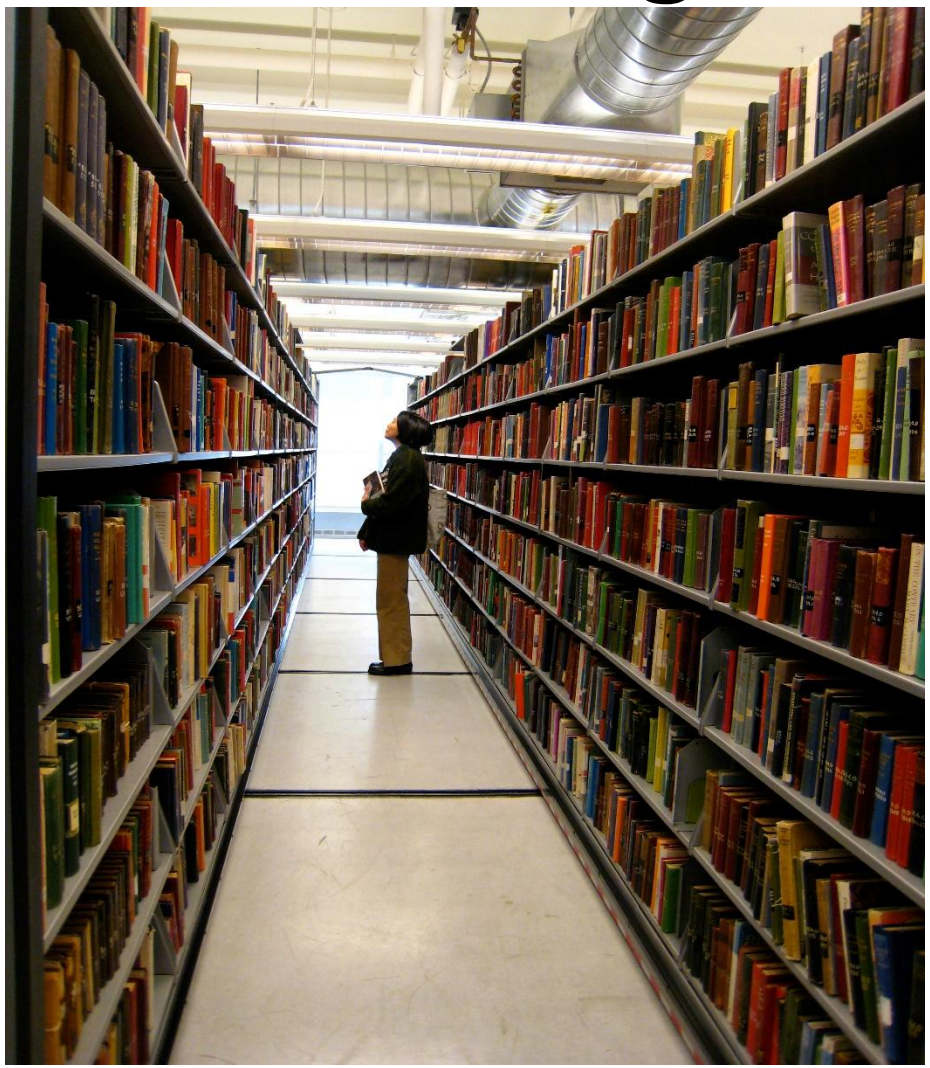


“Library Problem” – Probabilistic indexing

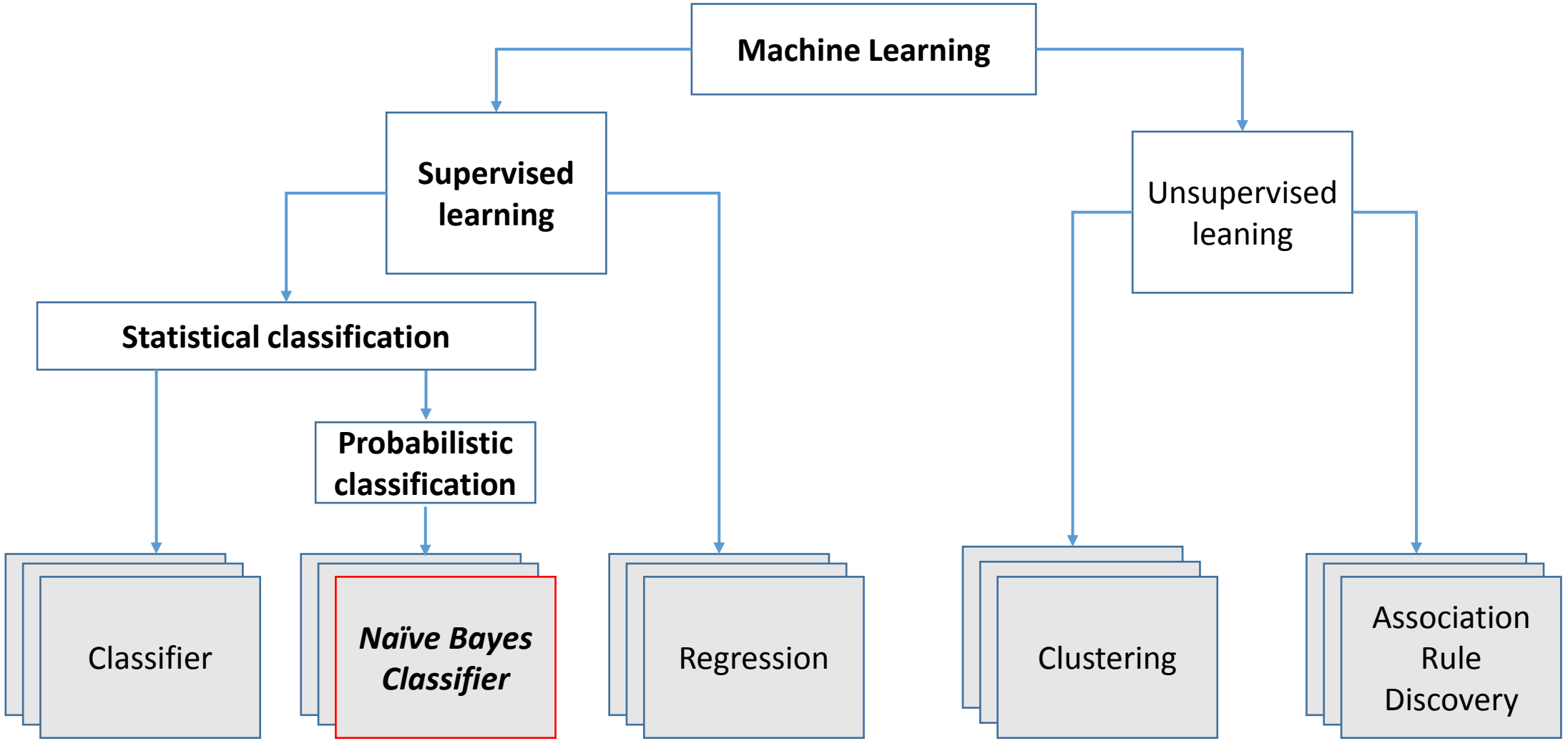
- Maron and Kuhns proposed a probabilistic approach:
  - Each tag should have weights associated with the corresponding documents
  - Based on Bayes Rules a “relevance number” was determined

$$P(D_i|I_j) - probability\ of\ document\ D_i\ given\ request\ I_j$$

It was the first ranking system and also put the basis of Naïve Bayes Classifier



What is Naïve Bayes Classifier



Statistical classification

- The problem of identifying to which set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known
- An algorithm which implements a classification is called *classifier*

Probabilistic basis

Notations :

- Prior probability : P(A) – probability of A
- Conditional probability: P(A|B) - probability of A given B
- Joint probability: P(A,B) – probability of A and B
  - $P(A, B) = P(A) \cdot P(B|A)$

Bayes Rule :

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

Thomas Bayes – English statistician

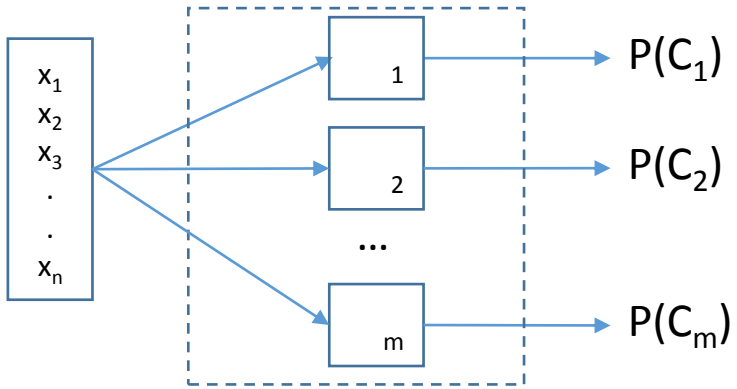
Probabilistic Classification

- Given input $X = [x_1, x_2, \ldots, x_n]$ and classes $C_1, C_2, \ldots, C_m$

1. Discriminative classifier

2. Generative classifier

Ex: Naïve Bayes



Bayes classifier

Given:

$X = [x_1, x_2, \dots, x_n]$, Classes $C_1, C_2, \dots, C_m$ and a training set T

Define an algorithm which computes:

$P(C_i|X)$ *using Bayes Rule:*

$$P(C_i|X) = \frac{\boldsymbol{P(X|C_i)} \cdot P(C_i)}{P(X)}$$

$\boldsymbol{P(X|C_i)} = P(x_1, x_2, \dots, x_n|C_i)$ ***is hard to compute in general***

$$= P(\boldsymbol{x_1}|C_i)P(\boldsymbol{x_2}|C_i, x_1)P(\boldsymbol{x_3}|C_i, x_1, x_2) \dots P(\boldsymbol{x_n}|C_i, x_1, x_2, \dots, x_{n-1})$$

Return $C_{map} = argmax_{C_i} P(X|C_i)$

Naïve Bayes Classifier

Compute:

$P(X|C_i) = P(\boldsymbol{x_1}|C_i)P(\boldsymbol{x_2}|C_i, x_1) \dots P(\boldsymbol{x_n}|C_i, x_1, x_2, \dots, x_{n-1})$ which is known as a difficult problem

Solution:

Naïve (Idiot) assumption – consider that $x_1, x_2, \dots, x_n$ attributes are independent so that

$$P(X|C_i) = P(x_1|C_i)P(x_2|C_i) \dots P(x_n|C_i)$$

Naïve Bayes – Discrete value features

- Categorical distribution CD=$(w_1, w_2, .., w_n)$ where n is the number of categories and $\sum w_i = 1$
- Problem – Find $\{w_j\}$ e. g. compute $P(x_j|C_i)$ when x_j is a discrete value attribute $x_j \in \{v1, v2, v3, \dots\}$

$$P(x_j|C_i) = \frac{Count(T, C_i, x_j)}{Count(T, C_i)}$$

Example:

$$P(Wind = weak \mid No) = \frac{2}{5}$$

$$P(Wind = strong \mid No) = \frac{3}{5}$$

$$CD = (\frac{2}{5}, \frac{3}{5})$$

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Naïve Bayes – continuous values features

- Usually use Normal (Gaussian) distribution N (μ, σ)
  - $P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{(-\frac{x-\mu}{2\sigma^2})}$ where μ is the mean and σ is the standard deviation

    $\mu = \frac{1}{N}\sum x_i, \sigma = \frac{1}{N}\sum (x_i - \mu)^2$
- Suppose the temperature is continuous

  Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

  No: 27.3, 30.1, 17.4, 29.5, 15.1

  Compute μ and σ for each class

  $\mu_{Yes} = 21.64, \quad \sigma_{Yes} = 2.39$

  $\mu_{No} = 23.88, \quad \sigma_{No} = 7.09$

Naïve Bayes Algorithm

- Learning phase

  For each class value C_i
  Compute $P(C_i)$
  For each attribute X_j
  //Compute Distribution D_{ij}
  If attribute X_j is discrete
  D_{ij} = Categorical distribution for C_i and X_j
  Else
  D_{ij} = Normal distribution for C_i and X_j

- Testing Phase – Given unknown $X' = [x'_1, x'_2, \ldots, x'_n]$
  estimate class = $argmax_{C_i} P(C_i) \cdot \prod_j D_{ij}$

Example on “Play Tennis” Data

What if X’=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

P(Yes)=9/14

P(No)=5/14

Example on “Play Tennis” Data

What if X’=(Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong)

Outlook	Play =Yes	Play = no
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play =Yes	Play = no
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play =Yes	Play = no
High	3/9	4/5
Normal	6/9	1/5

Wind	Play =Yes	Play = no
Strong	3/9	3/5
Weak	6/9	2/5

P(Yes)=9/14
P(No)=5/14
P(Yes|X’) = 2/9*3/9*3/9*3/9*9/14=486/91854 = 0.0052
P(No|X’) = 3/5*1/5*4/5*3/5*5/14=180/8750 = 0.02
Because P(No|X’)> P(Yes|X’) the answer is No

Important Issues

- Violation of *Independence Assumption*
  - Even if in many real cases $P(x_1, x_2, \dots, x_n|C_i) \neq P(x_1|C_i)P(x_2|C_i) \dots P(x_n|C_i)$ Naïve Bayes works very well
- Zero conditional probability handling
  - In case $P(x_j|C_i) = 0$ then $P(x_1, x_2, \dots, x_n|C_i) = 0$
  - To fix this issue consider applying Laplace Smoothing

$$P(x_j|C_i) = \frac{Count(T, C_i, x_j)+1}{Count(T, C_i)+k}, where\ x_j = \{v_1, v_2, \dots, v_k\}$$

Important notes

- Very competitive (proved success in spam filtering)
- Fast and easy to implement
- A good candidate of a base learner in ensemble learning
- Very popular

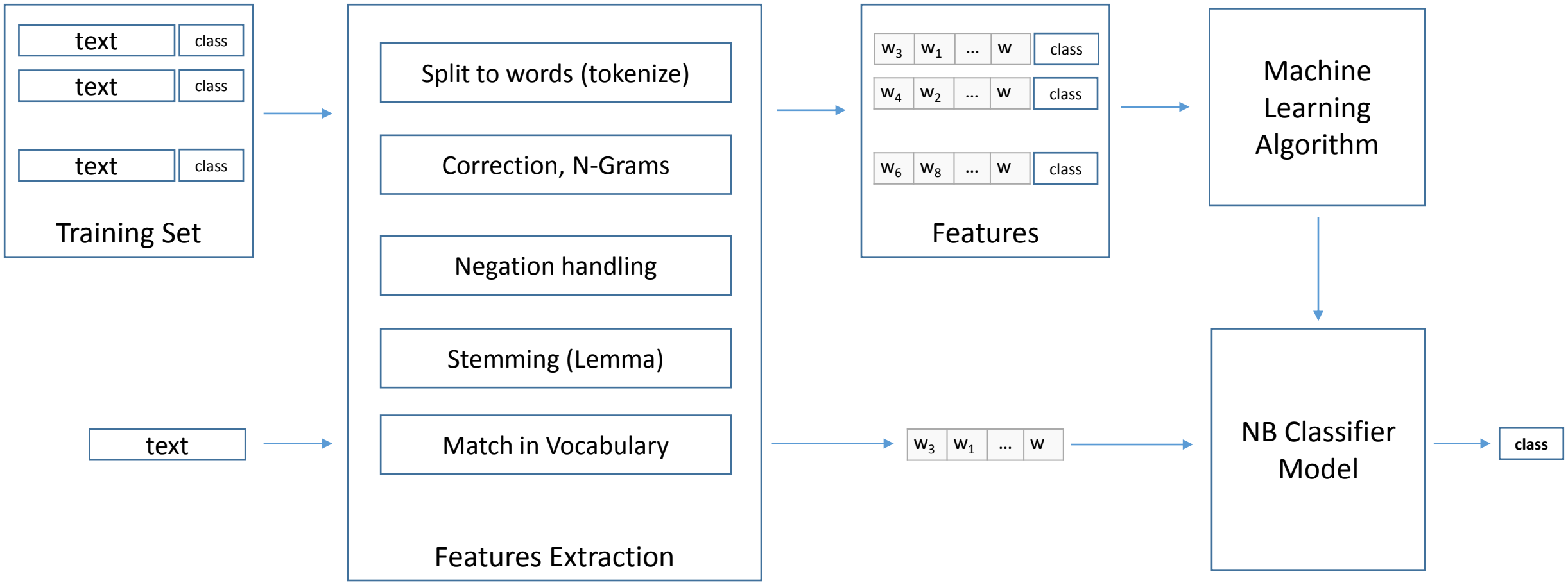
Naïve Bayes Applications

- Text classifications. Widely used in:
  - Spam filtering
  - Classify documents based on topics (technology, politics, science, etc.)
  - Sentiment analysis
  - Information retrieval
- Image classifications (e.g. Face detection)
- Medical field
  - Disease detection (Alzheimer's based on genome wide data)
  - Treatment detection

Example On Sentiment Analysis

- Tag a text as being positive or negative
- Usual algorithms on Sentiment Analysis
  - Naïve Bayes > 70-90 percent efficient
  - Entropy Maximization
  - Support Vector Machines (one of the most efficient)

Sentiment Analysis Engine



Demo