

Bases de l'IA

Traitement automatique du langage naturel (cont.)

Elena CABRIO

elena.cabrio@univ-cotedazur.fr

Classification des textes

S'agit-il de spam?

Subject: Important notice!

From: Stanford University <newsforum@stanford.edu>

Date: October 28, 2011 12:34:16 PM PDT

To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.

Commentaires sur un film: positifs ou négatifs?



- Incroyablement décevant.



- Tendre, subtil, finement joué, il émeut aussi par sa justesse d'écriture.



- C'est la plus grande comédie screwball jamais filmée.



- C'était pathétique. Le pire, c'était les scènes de boxe.

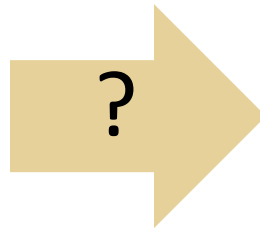
C'est quoi le sujet de l'article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...



Classification de texte

- Attribution de sujets, de thèmes ou de genres
- Détection du spam
- Identification de l'auteur
- Identification de l'âge/du sexe
- Identification de la langue
- Analyse des sentiments...

Classification de texte: définition

- *Entrée:*
 - un document d
 - un ensemble fixe de classes $C = \{c_1, c_2, \dots, c_J\}$
- *Sortie:* une classe $c \in C$

Méthodes de classification:

Règles codées à la main

- Règles basées sur des combinaisons de mots ou d'autres caractéristiques
 - spam: black-list-address OR (“dollars” AND “have been selected”)
- La précision peut être élevée, si des règles sont affinées par des experts
- Mais l'élaboration et le maintien de ces règles sont coûteux

Méthodes de classification :

Apprentissage supervisé

- *Entrée :*
 - Un document d
 - un ensemble fixe de classes $C = \{c_1, c_2, \dots, c_J\}$
 - Un corpus d'apprentissage de m documents étiquetés manuellement $(d_1, c_1), \dots, (d_m, c_m)$
- *Sortie:*
 - Un classifieur $\gamma: d \rightarrow c$

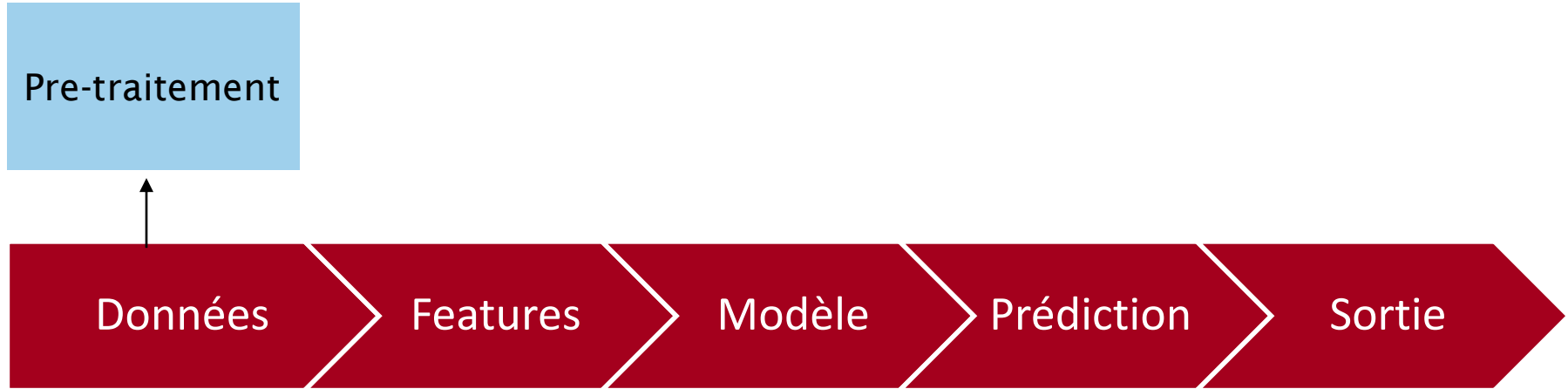
Méthodes de classification :

Apprentissage supervisé



Méthodes de classification :

Apprentissage supervisé



Dataset / Corpus

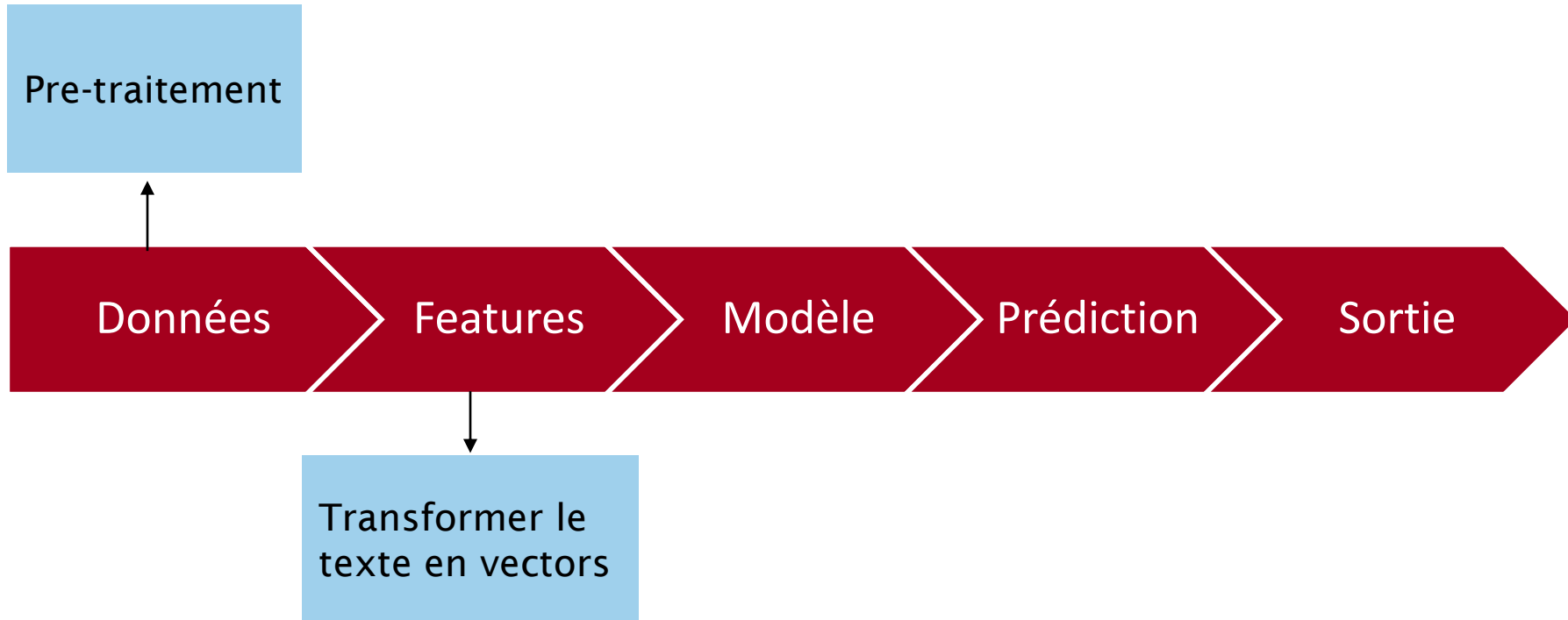
- Dictionnaire ou vocabulaire utilisé pour entraîner le modèle.
- Soit **étiqueté (pour l'apprentissage supervisé)**, soit **non étiqueté (pour l'apprentissage non supervisé)**.
- La taille dépend de l'algorithme utilisé.
- Doit être **prétraité** pour supprimer les caractères indésirables, pour le convertir au format souhaité, etc.

Comment représenter les mots en tant qu'entrée pour les méthodes de classification ?

**Comment représenter la signification
des mots en tant qu'entrée pour les
méthodes de classification ?**

Méthodes de classification :

Apprentissage supervisé



Extraction de features

- Transformer le texte en vecteurs numeriques (modèle vectoriel)
- Choix:
 - One-hot encoding
 - Bag-of-words + $TF*IDF$
 - Word2vec
 - ...

One-hot encoding

- One hot encoding: **représentation de variables catégoriques sous forme de vecteurs binaires.**
- Il faut d'abord que les valeurs catégorielles soient converties en valeurs entières.
- Ensuite, chaque valeur entière est représentée sous la forme d'un **vecteur binaire composé de toutes les valeurs nulles, sauf l'indice de l'entier, qui est marqué d'un 1.**

One-hot encoding

[Text]

- text 1 "Python is a programming language."
- text 2 "I use python language for programming."

1 [Token Index]

token	index
python	1
is	2
a	3
programming	4
language	5
I	6
use	7
for	8

3 [One-hot encoded]

text	word	Token Index							
		1	2	3	4	5	6	7	8
1	1	1	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0
	4	0	0	0	1	0	0	0	0
	5	0	0	0	0	1	0	0	0
2	1	0	0	0	0	0	1	0	0
	2	0	0	0	0	0	0	1	0
	3	1	0	0	0	0	0	0	0
	4	0	0	0	0	1	0	0	0
	5	0	0	0	0	0	0	0	1
	6	0	0	0	1	0	0	0	0

2

Max length of word
per text = 5

<https://rfriend.tistory.com>

ignored

La représentation en sac de mots

- Conversion du le texte en une matrice où chaque ligne est une observation et chaque feature est un mot unique. **La valeur de chaque élément de la matrice est soit un indicateur binaire marquant la présence de ce mot, soit un nombre entier du nombre de fois où ce mot apparaît.**

	I	love	dogs	and	knitting	is	my	hobby	passion
Doc 1	1	1	1						
Doc 2	1		1	1	1				
Doc 3				1	1	1	2	1	1

TF*IDF

- TF-IDF permet d'évaluer l'importance d'un terme contenu dans un document, relativement à une collection ou un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.

$$w_{x,y} = \text{tf}_{x,y} \times \log \left(\frac{N}{\text{df}_x} \right)$$

TF-IDF

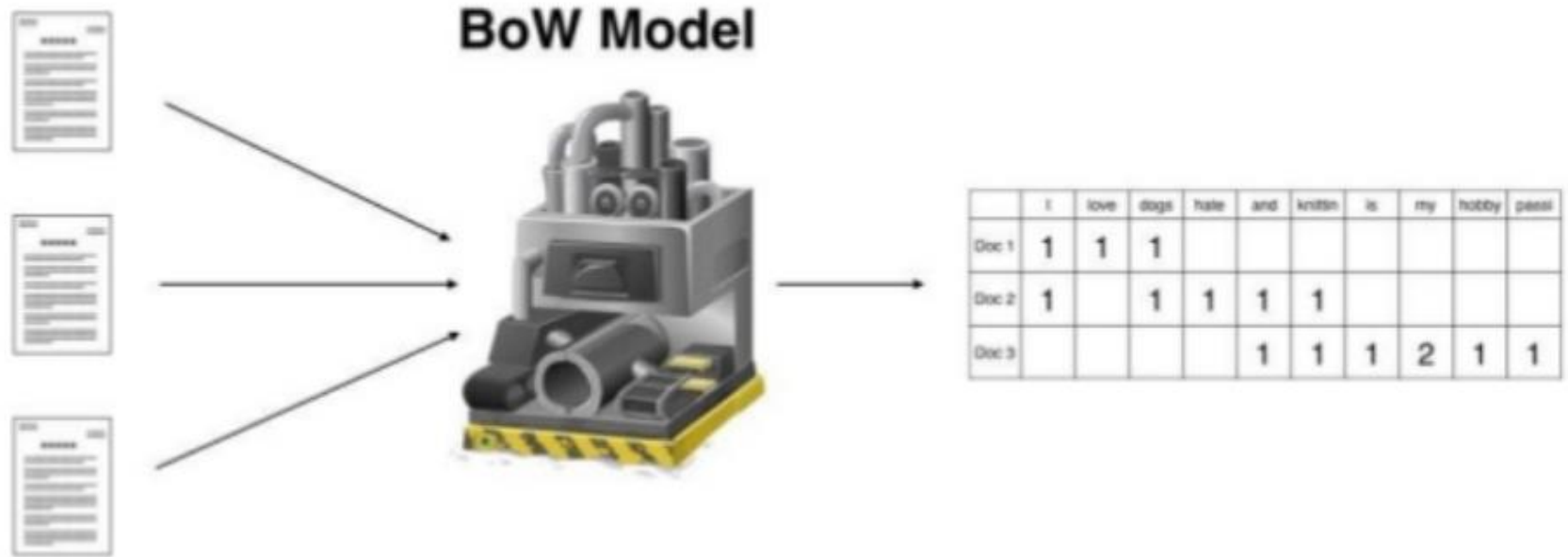
Term x within document y

$\text{tf}_{x,y}$ = frequency of x in y

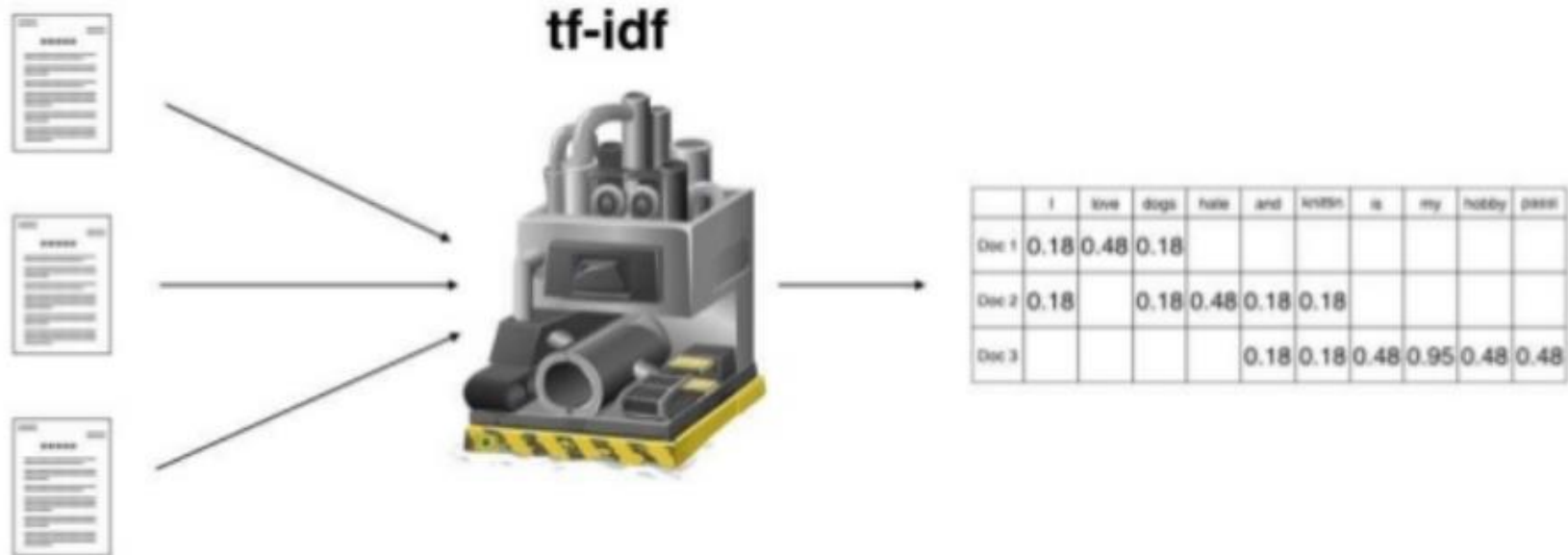
df_x = number of documents containing x

N = total number of documents

BOW+TF*IDF



BOW+TF*IDF





BOW+TF*IDF

- L'ordre des mots est sans importance
- Le document "Jean est plus rapide que Marie" ne peut être distingué du document "Marie est plus rapide que Jean".

Hypothèse distributionnelle

- Le degré de similarité sémantique entre deux expressions linguistiques est une fonction de similarité de leurs contextes linguistiques.
- Similitude de sens = Similitude de contexte
- Définition simple : **contexte = mots environnants**

Interprétation géométrique:

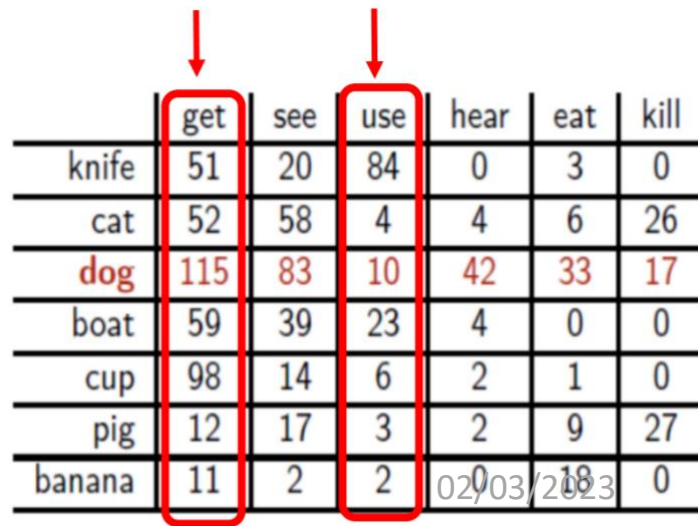
Co-occurrence comme feature

- Rappelons la matrice **terme-document**
 - Les rangées sont des termes, les colonnes sont des documents, les cellules représentent le nombre de fois qu'un terme apparaît dans un document.
- Ici, nous créons une **matrice de cooccurrence mot-mot**
 - Les rangées et les colonnes sont des mots
 - Cellule (R,C) signifie "combien de fois le mot C apparaît dans le voisinage du mot R".
- **Voisinage** = une fenêtre de taille fixe autour du mot

Matrice de co-occurrence

- Chaque vecteur décrit l'utilisation du mot dans le corpus/document.
- Les vecteurs peuvent être vus comme les coordonnées du point dans un espace euclidien à n dimensions
 - Exemple: $n = 2$
 - Dimensions = 'get' et 'use'

Matrice de co-occurrence

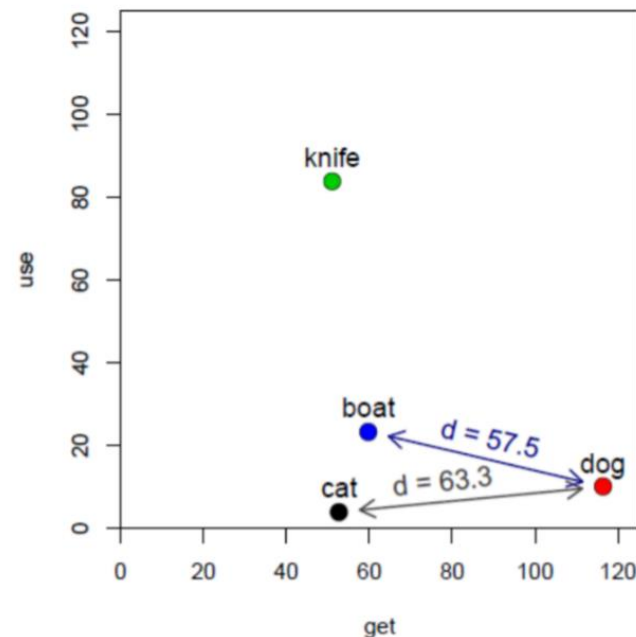


	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Distance et similarité

- Deux dimensions sélectionnées 'get' et 'use'
- Similitude entre les mots = proximité spatiale dans l'espace dimensionnel
- Mesurée par la distance euclidienne

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

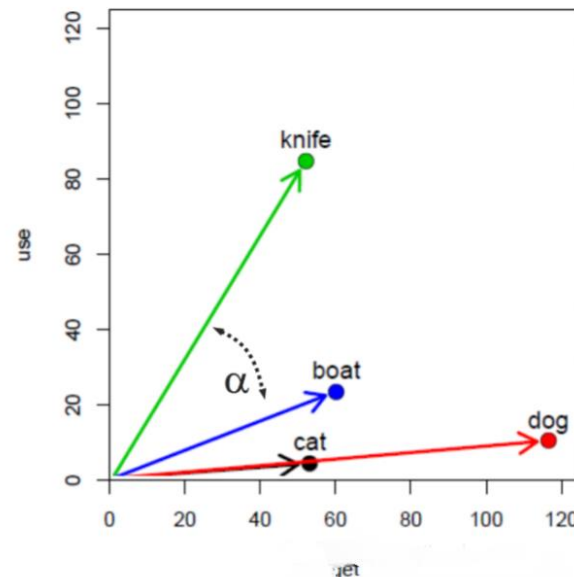


Distance et similarité

- La position exacte dans l'espace dépend de la fréquence du mot.
- Les mots les plus fréquents apparaissent plus loin de l'origine.
 - Par exemple, si 'dog' est plus fréquent que 'cat'
 - Cela ne signifie pas qu'il est plus important
- **Solution : Ignorez la longueur et regardez uniquement la direction**

Angle et similarité

- L'angle ne tient pas compte de l'emplacement exact du point
- **Méthode** : Normaliser par la longueur des vecteurs ou utiliser uniquement l'angle comme mesure de distance.
- **Métrique standard** : Similitude cosinus entre vecteurs



Problématiques relatives à la matrice de cooccurrence

- Problème de l'utilisation directe de la cooccurrence :
 - Les vecteurs résultants sont de très haute dimension
 - Taille de la dimension = Nombre de mots dans le corpus
 - Milliards!
 - Le sous-échantillonnage des dimensions n'est pas simple
 - Combien de colonnes sélectionner ?
 - Quelles colonnes sélectionner ?
- Solution: **Compression or Dimensionality Reduction Techniques**

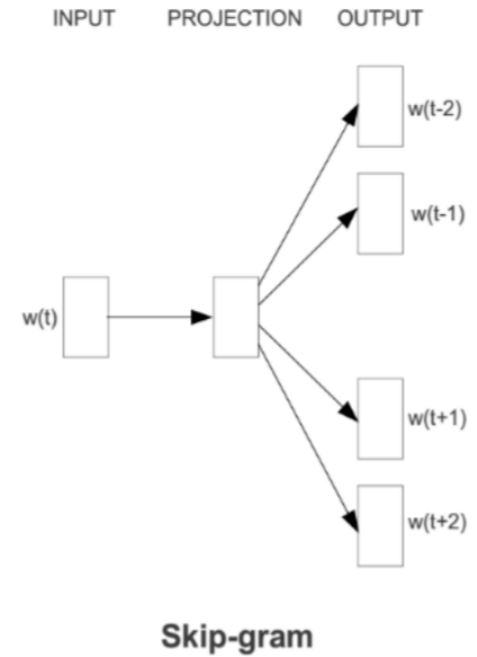
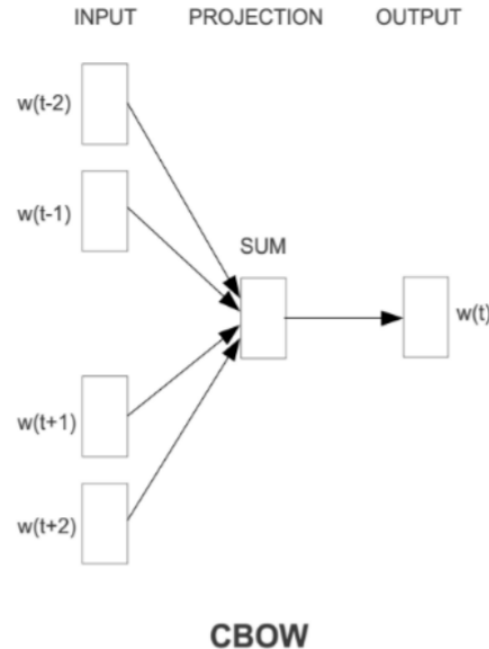
Word2Vec: Représentation du sens des mots

- **Idée clé :** **Prédire les mots environnants de chaque mot**
- Donne de meilleures relations sémantiques/syntaxiques des mots grâce aux vecteurs.
- Avantages :
 - Plus rapide
 - Plus facile d'intégrer de nouveaux mots et documents.

Deux méthodes d'apprentissage Word2Vec



- Continuous Bag of Words (CBOW): utilise les mots du contexte dans une fenêtre pour prédire le mot du milieu.
- Skip-gram: utilise le mot du milieu pour prédire les mots du contexte dans une fenêtre.



L'incroyable pouvoir des vecteurs de mots



Word
Vectors



Vector
Composition

<https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/>

- $\text{vecteur}(\text{king}) - \text{vecteur}(\text{man}) + \text{vecteur}(\text{woman}) = \text{vecteur}(\text{queen})$

Analogies

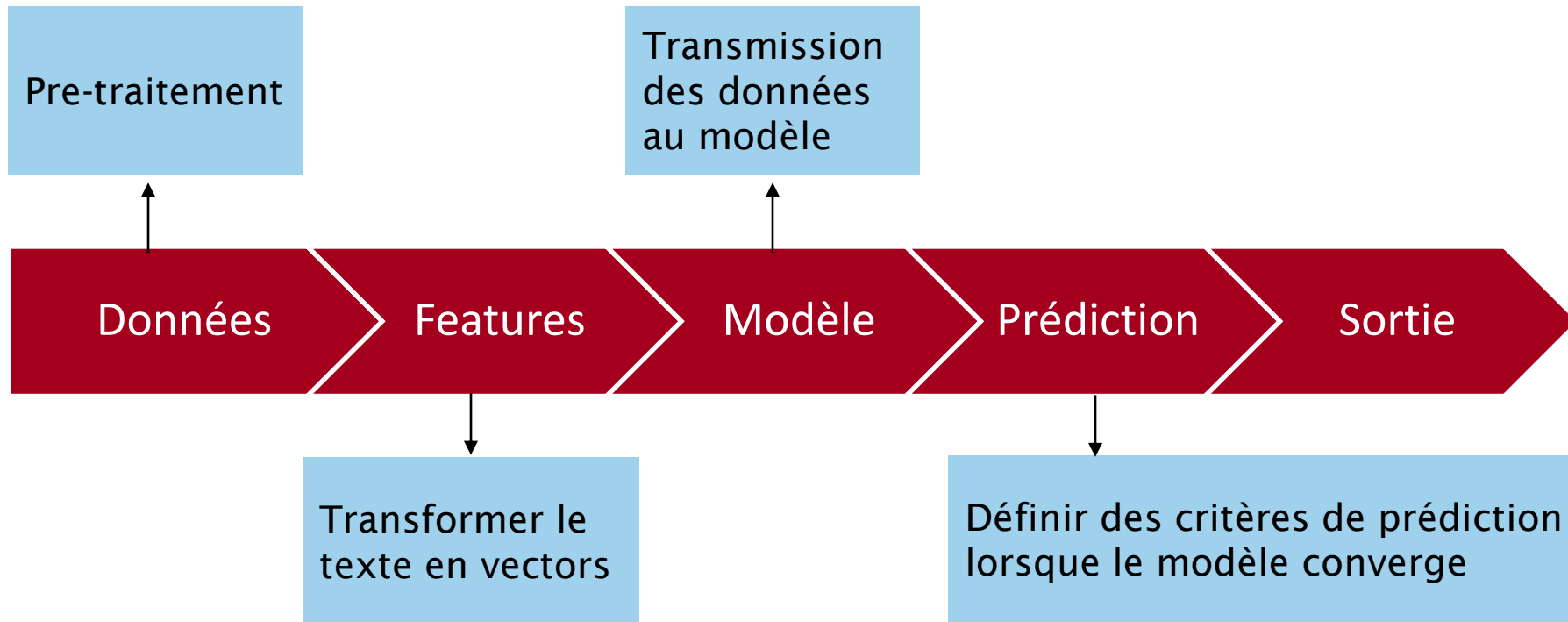
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Mikolov et al., 2013

Méthodes de classification :

Apprentissage supervisé



Méthodes de classification :

Apprentissage supervisé

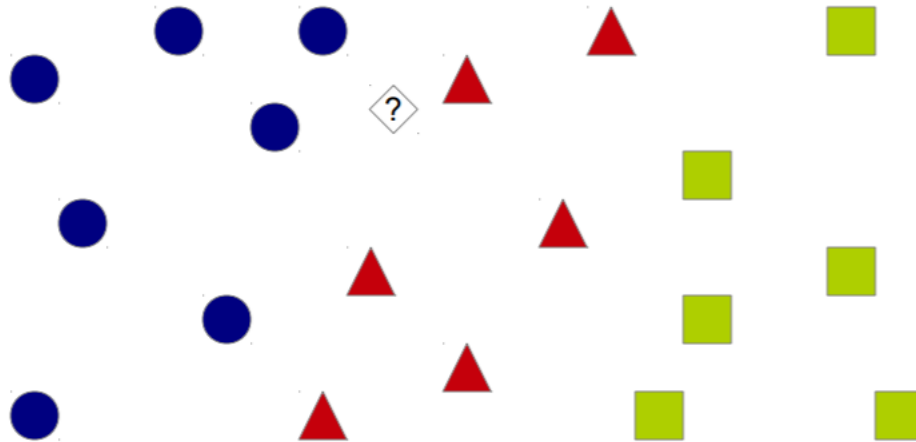
- Tout type de classificateur
 - k plus proches voisins
 - Machine à vecteurs de support
 - Naïve Bayes
 - Régression logistique
 - Réseaux de neurones...

Algorithme k plus proches voisins

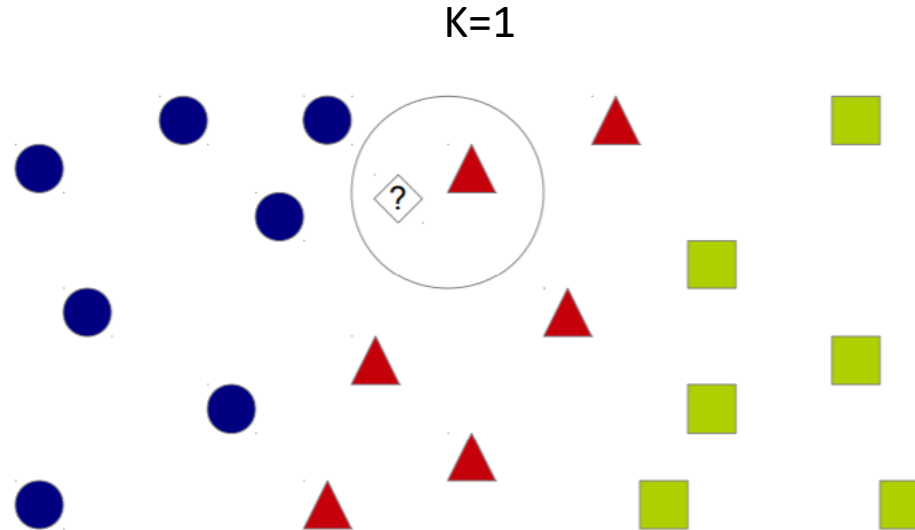
- **Entrée:** les k exemples d'apprentissage les plus proches dans l'espace des features.
- **Sortie:** l'appartenance à une classe. Un objet est classé par un vote de ses voisins, l'objet étant affecté à la classe la plus courante parmi ses k voisins les plus proches (k est un nombre entier positif, généralement petit).
 - Si $k = 1$, alors l'objet est simplement affecté à la classe de ce seul voisin le plus proche



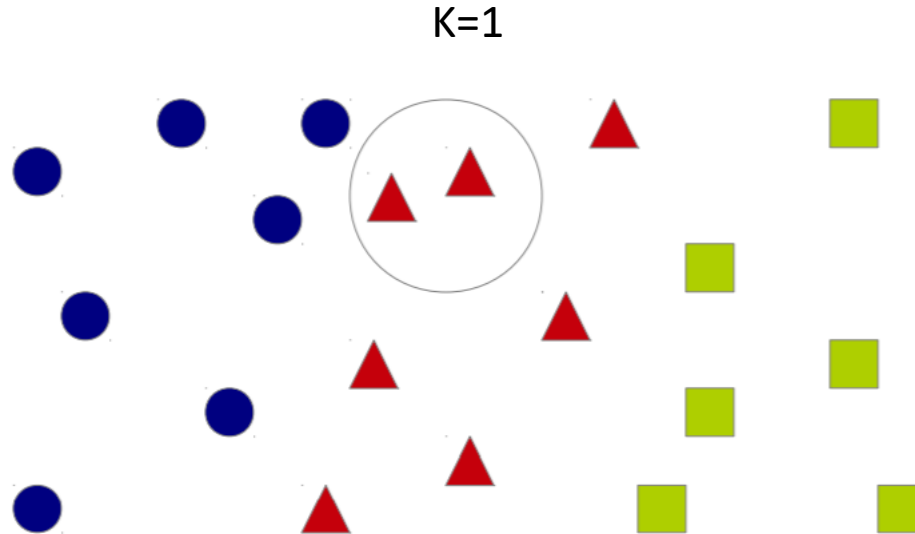
Algorithme k plus proches voisins



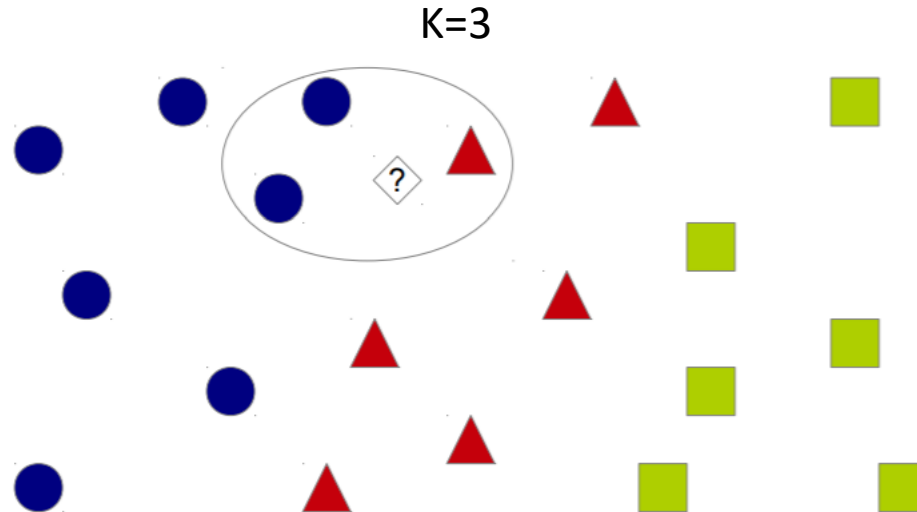
Algorithme k plus proches voisins



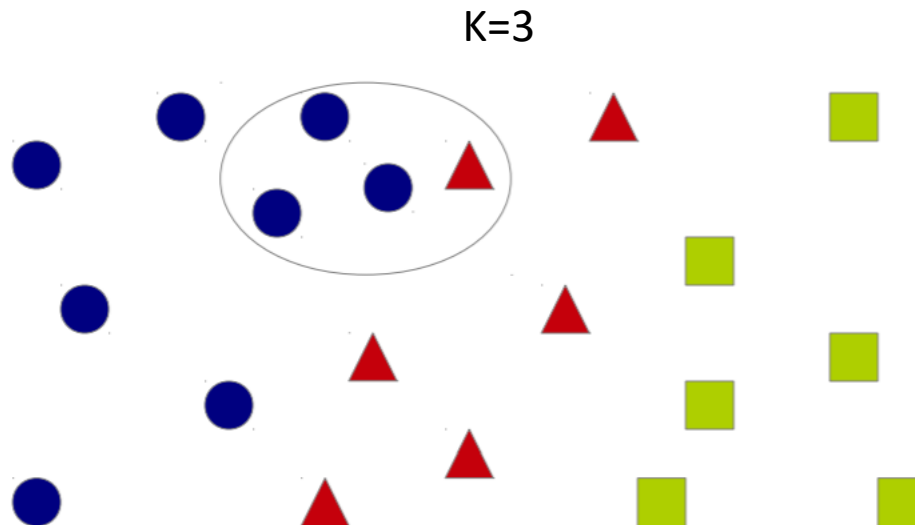
Algorithme k plus proches voisins



Algorithme k plus proches voisins



Algorithme k plus proches voisins



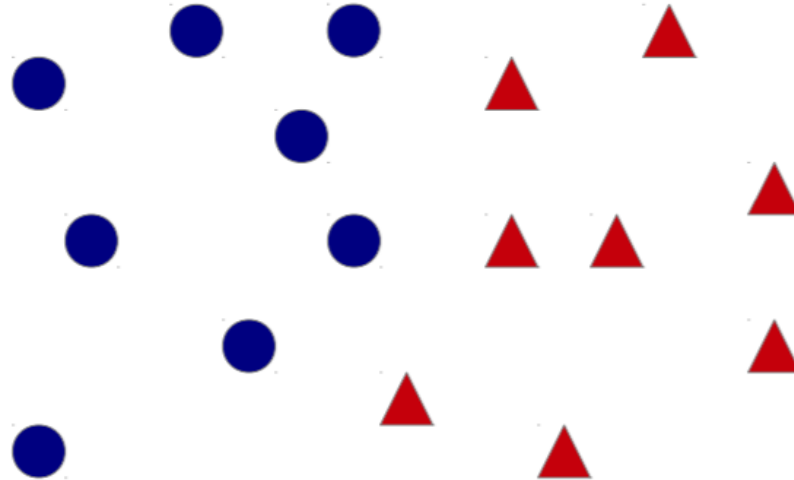
- Approche simple
- No black-box
- Sélectionner
 - Features
 - Métriques de distance
 - Valeur de k (majority voting)

Machine à vecteurs de support

- *Marge maximale.* La marge est la distance entre la frontière de séparation et les échantillons les plus proches. Ces derniers sont appelés *vecteurs supports*. Dans les SVM, la frontière de séparation est choisie comme celle qui maximise la marge
- Pour traiter des cas où les données ne sont pas linéairement séparables, l'espace de représentation des données d'entrées est transformé en un espace de plus grande dimension dans lequel il est probable qu'il existe une séparation linéaire. Ceci est réalisé grâce à une fonction noyau.

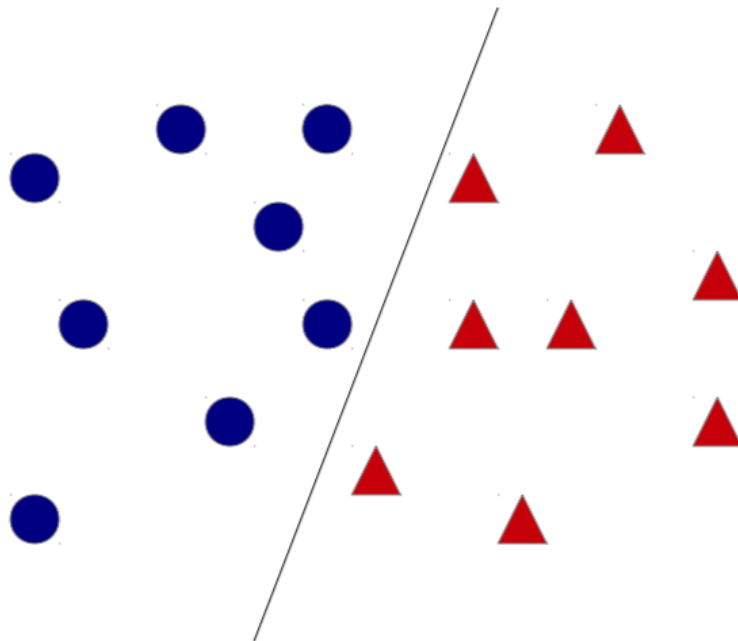


Machine à vecteurs de support



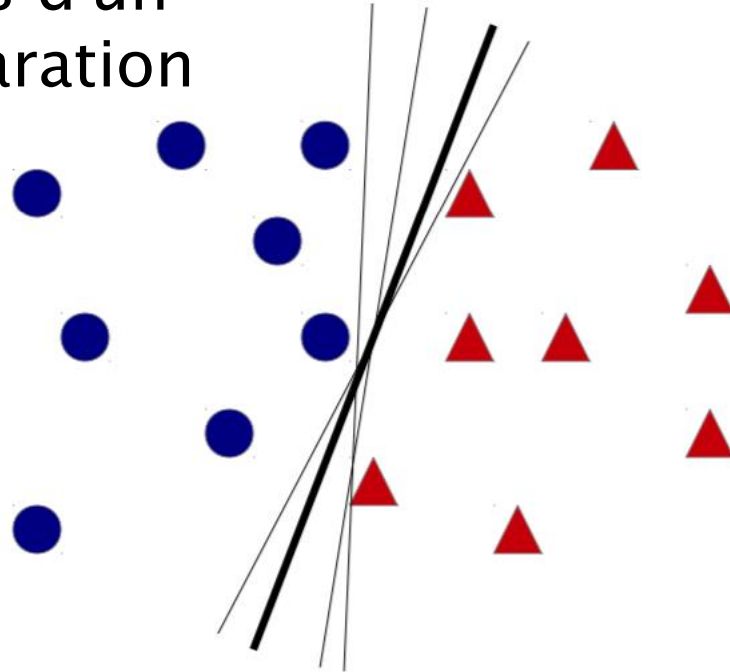
Machine à vecteurs de support

Trouver un hyperplan dans l'espace vectoriel qui sépare les éléments des deux catégories.



Machine à vecteurs de support

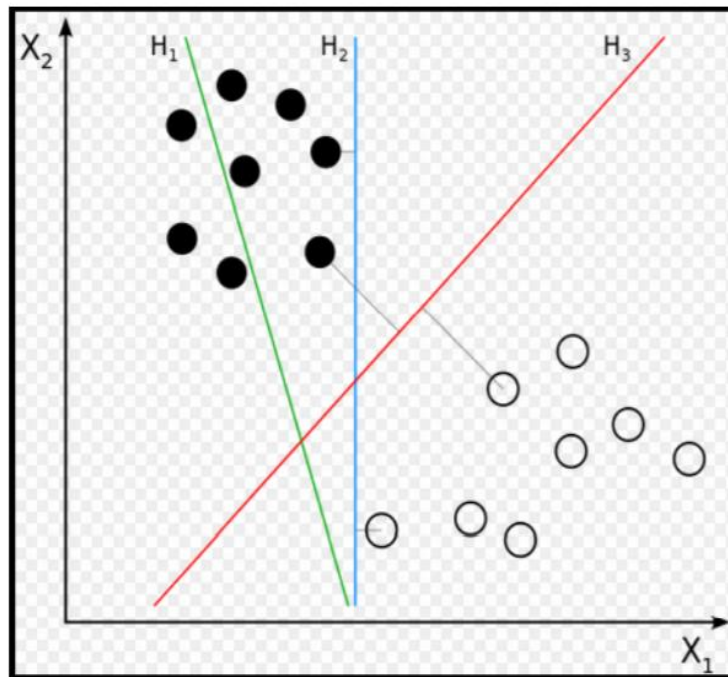
Il peut y avoir plus d'un hyperplan de séparation possible.



Machine à vecteurs de support

Trouvez l'hyperplan
avec la marge
maximale

Les vecteurs aux
marges sont appelés
vecteurs de support



Machine à vecteurs de support

- Fournit généralement de bons résultats
- Il existe des bibliothèques disponibles pour de nombreux langages de programmation
- Classification linéaire et non linéaire
- Boîte noire
- Les problèmes multi-classes sont généralement modélisés par de nombreuses classifications binaires.

Classification naïve bayésienne

- La **classification naïve bayésienne** est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance (dite naïve) des hypothèses
- Il suffit d'un petit nombre de données d'apprentissage pour estimer les paramètres nécessaires à la classification

Bayes' theorem

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

In here, A and B are two events, $P(A)$ and $P(B)$ are the two probabilities of A and B if treated as independent events, and $P(A|B)$ and $P(B|A)$ is the compound probability of A given B and B given A , respectively.

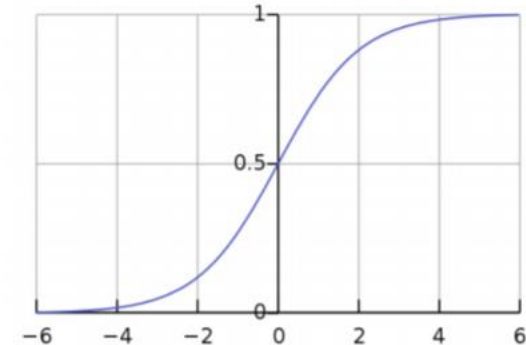
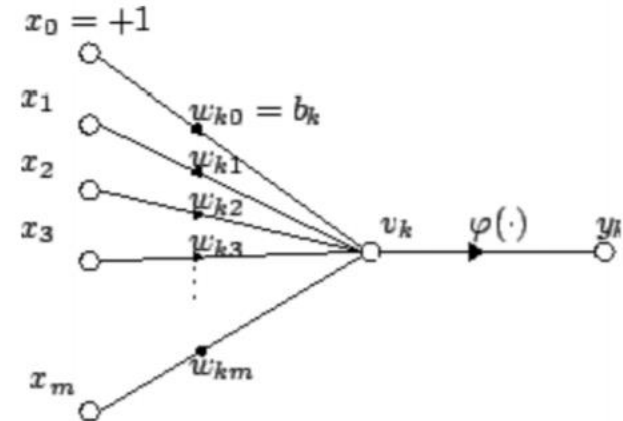
$P(A)$ and $P(B)$ are the crucial points here, and **they refer to the so-called *a priori* probabilities of A and B .**

Réseaux de neurones

- Les éléments de base des réseaux neuror
- L'entrée est un vecteur: $x = [x_1, \dots, x_m]$
- Poids et biais :
 - Le neurone a des poids $w = [w_1, w_2, \dots, w_m]$
- Bias term = b (or w_0)
- Fonction d'activation ψ
 - Transforme l'agrégat
 - e.g., sigmoid, ReLU
- Calcul de la sortie :

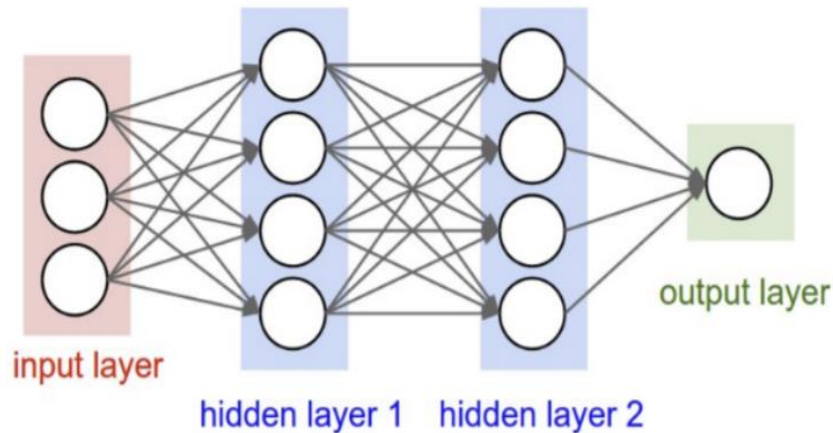
$$y = \psi(\sum_{j=1}^m w_j x_j + b)$$

Bases de l'IA



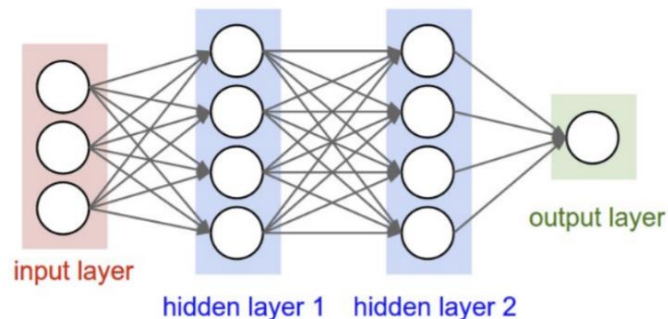
Réseaux de neurones: Couches entièrement connectées

- Une couche dont les neurones sont connectés à tous les neurones de la couche précédente.
- Chaque neurone prend en entrée toutes les sorties de la couche précédente.
- Plusieurs couches peuvent être empilées ensemble
- Exemple : 3 couches entièrement connectées



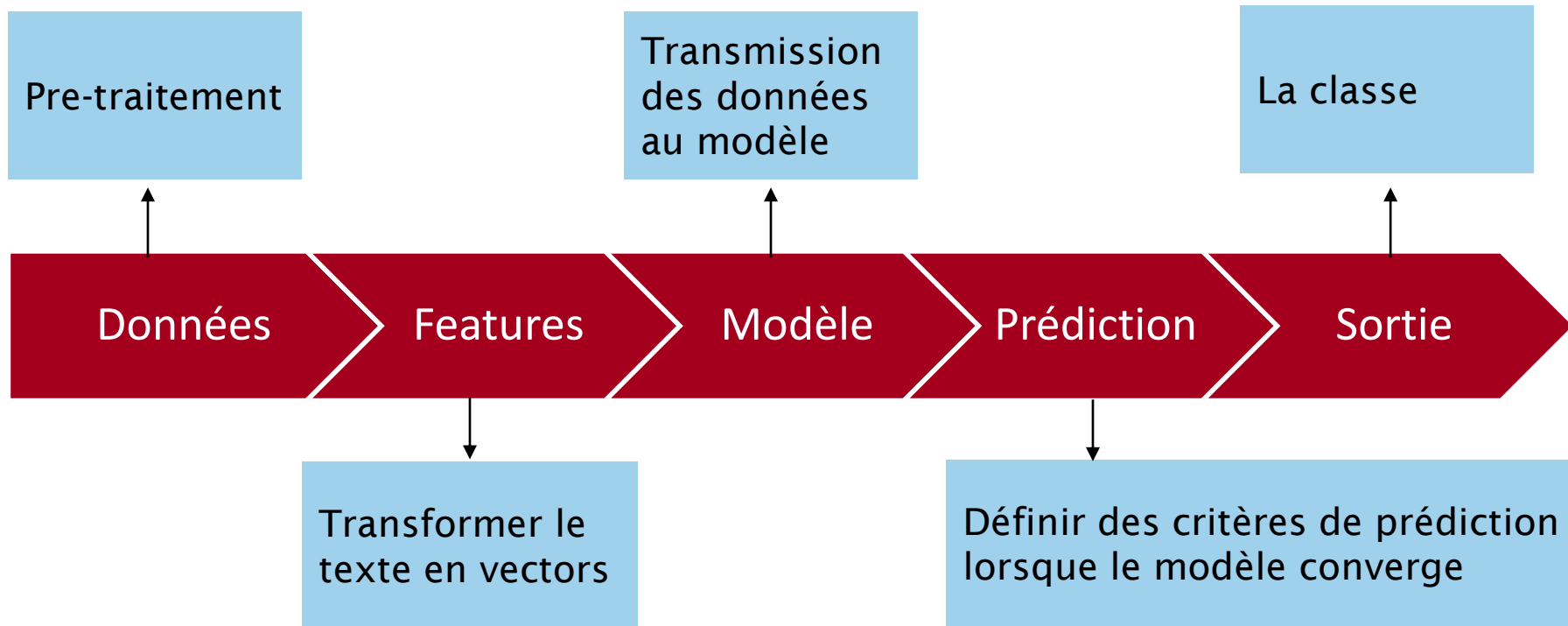
Réseaux de neurones:

- **Input layer:** les vecteurs d'entrée sont donnés comme entrées ici
- **Hidden layer:** Représentation intermédiaire des entrées
 - Multiple hidden layers can be stacked together
- **Output layer:** résultat final
 - Peut avoir un ou plusieurs neurones dans la couche de sortie.
- Notez que l'information circule dans une seule direction



Méthodes de classification :

Apprentissage supervisé



Per class evaluation measures

Rappel:

détermine la proportion des valeurs positives qui ont été prédites avec précision.

Precision:

La métrique de précision mesure le nombre de classes correctement prédites par le modèle - les vrais positifs et les vrais négatifs

Accuracy: $(1 - \text{error rate})$

Micro- vs. Macro-Averaging

- Si nous avons plus d'une classe, comment pouvons-nous combiner plusieurs mesures de performance en une seule quantité ?
- **Macroaveraging:** Calculer les performances pour chaque classe, puis faire la moyenne.
- **Microaveraging:** Recueillir les décisions pour toutes les classes, calculer le tableau de contingence, évaluer.

Development Test Sets and Cross-validation

Training set

Development Test Set

Test Set

- Metric: P/R/F1 or Accuracy
- Nouveau test set
 - éviter le surapprentissage
 - une estimation plus prudente de la performance
- Validation croisée sur des fractionnements multiples
 - Gérer les erreurs d'échantillonnage de différents ensembles de données

Training Set

Dev Test

Training Set

Dev Test

Dev Test

Training Set

Test Set