

Analyse de données massives (ADM)

—

Classification de textes et images

DUT2 – Semestre 4 – IUT de La Rochelle

Introduction

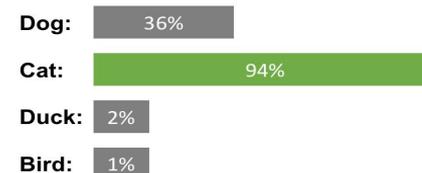
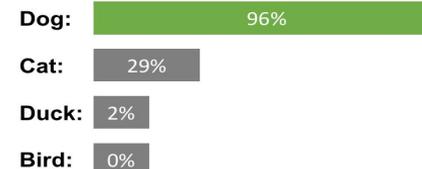
- Deuxième partie du cours ADM :
 - Classification de textes
 - Aujourd'hui
 - Classification des images
 - 9 mars

- Rendre le TP (20 mars):
 - 2 x Fichier Jupyter Notebook
 - Classif. Textes
 - Classif. Images
 - *clone gitlab project*
 - *commit + push dans github classroom*

"I love this movie.
I've seen it many times
and it's still awesome."



"This movie is bad.
I don't like it at all.
It's terrible."



Introduction

- Les **cerveaux humains** sont câblés pour **reconnaître des motifs** et **classer des objets** pour apprendre et prendre des décisions
- .. ils ne sont pas capables de traiter chaque objet comme unique
- .. nous n'avons pas beaucoup de ressources mémoire pour pouvoir traiter le monde qui nous entoure
- → nos cerveaux développent des «**concepts**» ou des représentations mentales de «**catégories d'objets**»
- La **classification** est fondamentale dans le **langage**, la prédiction, l'inférence, la prise de décision et toutes sortes d'interactions environnementales
- **Langage** : par exemple, comment le sens des mots d'une phrase peut être contextualisé par des mots ou des concepts antérieurs

Introduction

- La **classification des objets** consiste à attribuer une classe à un objet.
- Ce objet peut être du type :
 - Texte, image, audio, vidéo, etc.



- Nous faisons de la classification tout le temps :
 - On peut reconnaître le chemin chez nous du retour de l'université
 - On peut reconnaître un chat qui est noir même si on n'a vu que des chats blancs et oranges auparavant
 - On peut reconnaître l'ironie
 - On peut même faire la distinction entre un Chihuahua et un muffin

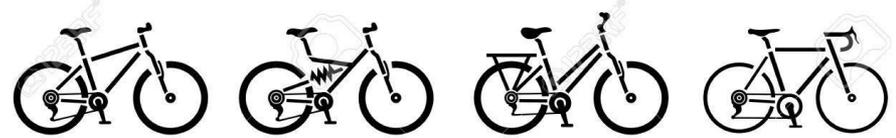


Introduction

- Pour savoir comment classer un objet, il est important de connaître les **caractéristiques** qui définissent une classe.
 - Si on considère les **caractéristiques** :
 - nombre de roues
 - selle
 - guidon
 - Un *vélo* est composé de 2 roues, des freins, une selle et un guidon.
 - Une *voiture* à 4 roues, des freins mais pas de selle ni guidon.
 - Une *moto* est aussi composée de 2 roues, une selle, des freins et un guidon.
- Avec ces caractéristiques, **la moto et le vélo ont la même représentation.**

Introduction

- Nous pourrions compliquer encore plus cette tâche et essayer de classer les vélos suivants :

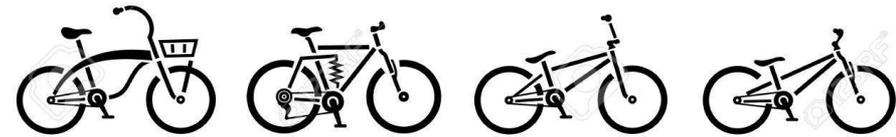


Hardtail

Downhill

Classic women

Road

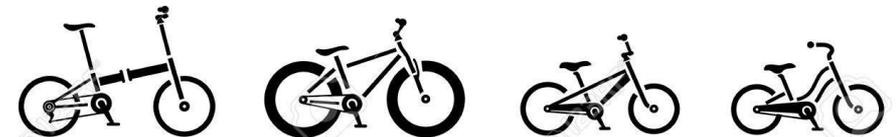


Vintage

Freeride

BMX

Dirt jump

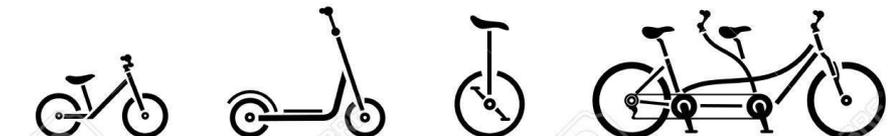


Folding

Fatbike

Kids' man

Kids' women



Kids' balance

Kick scooter

Unicycle

Tandem

Classification de textes

Introduction

Introduction

- Un texte (message, sms, livre, les paroles d'une musique, etc.) peut être classifié dans plusieurs types de classes.
- Le contenu d'un livre peut être considéré comme :
 - Romantique
 - Comédie
 - Suspense
 - Fantastique
 - Science-fiction
- Un commentaire peut être:
 - Positif
 - Négatif
 - Neutre
- Un mail/message peut être:
 - Spam
 - Pas spam

Classification de textes : filtrage du spam

- Première **application industrielle à grande échelle** du machine learning
- Problème de classification
- Succès du modèle Naïve Bayes



Dear Account Holder,

Due to suspicious activity, we have disabled your account. We highly recommend resetting your account password. You will no longer be able to use your card until doing so. We apologize for the inconvenience. Click the link below to reset your password:

-> <http://www.shelterplus.in/account-reactivation>

Classification de textes : filtrage du spam

- Comment définir si un message est un **spam** ou pas ? (spam or ham)
- Quelles caractéristiques sont importantes à analyser dans ce message ?



- La création de **règles** en écrivant des programmes informatiques est souvent **impossible**

beaucoup de données → beaucoup de cas → beaucoup de règles

Le **machine learning** permet d'**identifier des règles ou des patterns** dans les données

Classification de textes

Représentation de mots et de documents

Représentation des mots

- Après l'étape de prétraitement du texte, la représentation de mots et de documents est très pertinente car elle nous permet d'analyser des caractéristiques du texte.
- D'un côté, certains types de représentations sont très **simples** et rapide à calculer, par contre elles ne contiennent pas beaucoup d'informations sur les mots/documents.
- De l'autre côté, les représentations plus **riches** sont plus lentes à calculer mais elles contiennent plusieurs caractéristiques sur les mots/documents.
- Nous allons nous rappeler les **principaux types de représentations de mots et de documents**:
 - **One hot encoding**
 - **Bag-of-words (sac de mots)**
 - **TF-IDF**
 - **Word embeddings**

One hot encoding

- La représentation **one hot encoding** est une de représentations les plus simples car tous les mots sont indépendants entre eux.
- La taille de la représentation augmente avec le corpus.
- Chaque vecteur est équidistant de tous les autres vecteurs

“A friend in need is a friend indeed.”

$V = [a, \text{friend}, \text{in}, \text{need}, \text{is}, a, \text{indeed}], |V| = 7$

- Imaginez que nous ayons un vocabulaire de 50,000.
(Il y a environ un million de mots en anglais.)
- Chaque mot est représenté par 49,999 zéros et un seul 1
→ nous avons besoin de $50,000^2 = 2,5$ milliards d'unités d'espace mémoire.
- **Pas efficace** en termes de calcul.

a	friend	in	need	is	a	indeed
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

Bag-of-words (représentation de documents)

- La représentation **bag-of-words** est une représentation de documents en tenant compte de la fréquence des mots dans le document.
- La taille de la représentation augmente avec le corpus.

DOC1: "He is not a friend in need."

DOC2: "A friend in need is a friend indeed."

	he	friend	in	need	is	a	indeed
DOC1	1	1	1	1	1	1	0
DOC2	1	2	1	1	1	2	1

TF-IDF (représentation de documents)

- Term Frequency–Inverse Document Frequency
- Applications :
 - Recherche d'information (Information Extraction)
 - Fouille de textes (Text Mining)
- Cette mesure statistique permet d'évaluer **l'importance d'un terme** contenu dans un document, **relativement à une collection** de textes.
- 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.

TF-IDF : Term Frequency-Inverse Document Frequency

TF = $\frac{\text{Nombre de répétitions d'un mot dans un texte}}{\text{Nombre de mots dans un texte}}$

IDF = $\log \left[\frac{\text{Nombre de textes}}{\text{Nombre de textes contenant le mot}} \right]$

DOC1: "He is not a friend in need."

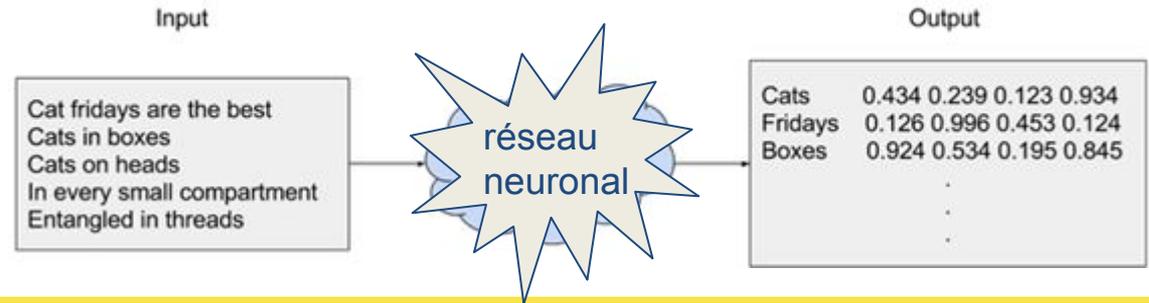
DOC2: "A friend in need is a friend indeed."

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

	TF		IDF	TF*IDF	
	DOC1	DOC2		DOC1	DOC2
he	1/7 = 0,14	0	log(2/1)=0,3	0,04	0
is	0,14	1/8=0,12	0	0	0
not	0,14	0	0,3	0,04	0
a	0,14	2/8 = 0,25	log(2/2)=0	0	0
friend	0,14	0,25	0	0	0
in	0,14	0,12	0,3	0,04	0,03
need	0,14	0,12	0,3	0,04	0,03
indeed	0	0,12	0,3	0	0,03

Word embeddings (« plongements de mots »)

- Cette représentation permet de représenter chaque mot d'un dictionnaire par un **vecteur de nombres réels**.
- Les mots avec des contextes similaires possèdent des vecteurs qui sont relativement proches.
- Cette technique est basée sur l'hypothèse qui veut que les mots apparaissant dans des contextes similaires ont des significations apparentées :
 - « chien » et « chat » (animaux domestiques)
 - « samedi » et « dimanche » (jours dans une semaine)
 - « vin rouge », « bière » (boissons alcoolisées)



Word embeddings (« plongements de mots »)

tous les mots du vocabulaire $|V|=50,000$

probabilités \implies .. $p(\text{autre mot}|\text{mat})$ $p(\text{the}|\text{mat})$ $p(\text{autre mot}|\text{mat})$..

fonction d'activation pour normaliser la sortie d'un réseau en une distribution de probabilité sur des classes de sortie prédites

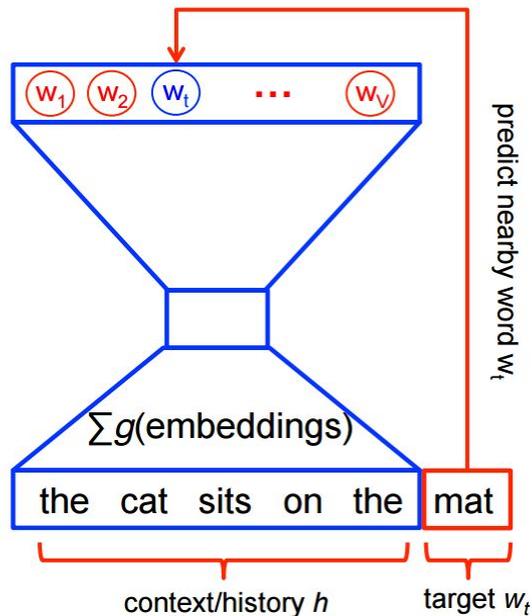
\implies Softmax classifier

la plupart des calculs sont ici
 $W \times V_{the} + b$

\implies Hidden layer

\implies Projection layer

0 4 9 7 867 67
The cat sits on the mat



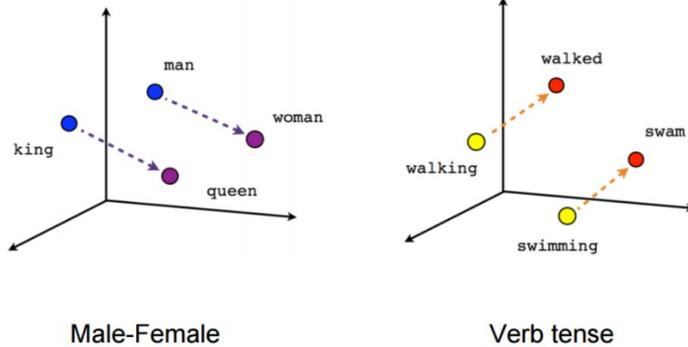
Vocabulaire $|V|=50,000$

0	1	2	3	4	5	..	50,000
the	a	in	with	cat	dog	..	

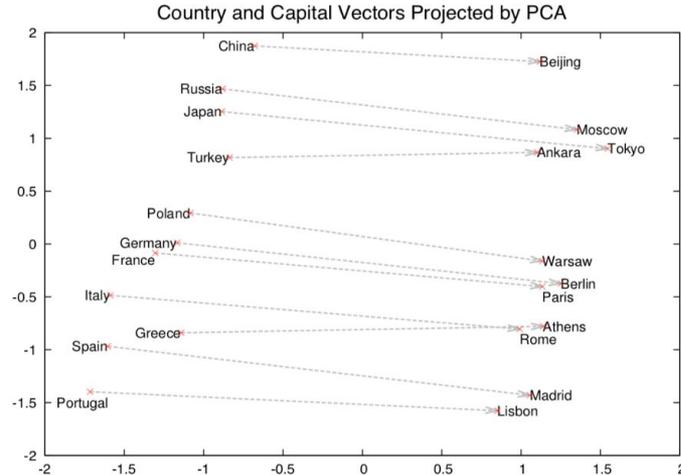
the	\rightarrow	0.34	0.4	0.11	0.5	0.89
cat	\rightarrow	0.6	0.23	0.8	0.87	0.21
dog	\rightarrow	0.77	0.21	0.09	0.29	0.05
...						

Word embeddings (« plongements de mots »)

- Relations sémantiques et géométriques



- Arithmétique vectorielle et sémantique



Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Classification de textes

Diviser les jeux de données

Diviser les jeux de données

- Afin d'entraîner les modèles et évaluer la performance de ses modèles avec chaque représentation de mots, nous allons diviser les jeux de données en :
 - entraînement
 - développement (validation)
 - teste



- Afin d'évaluer correctement la performance de chaque modèle, il est très important que les données d'entraînement et de teste soient différentes.

Classification de textes

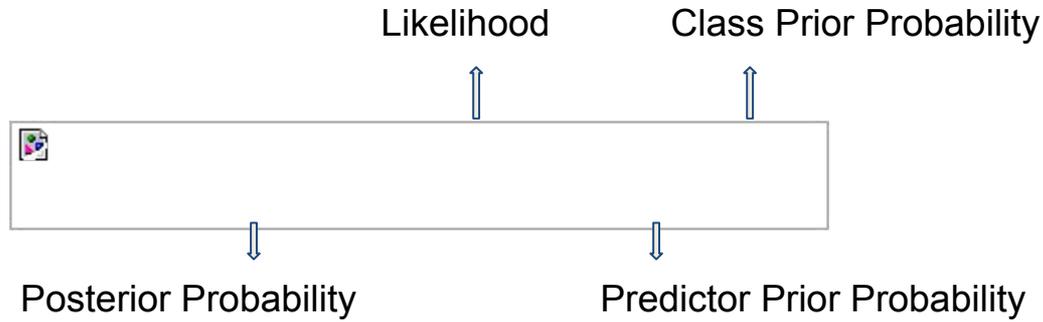
Approches de classification

Approches de classification

- Actuellement, il y a plusieurs approches disponibles pour la classification de données.
- Certaines approches sont plus adaptées pour certains types et quantité de données.
- Par exemple, les réseaux neuronaux sont très populaires actuellement car ils ont battu la plupart de systèmes.
- Dans notre cours, nous irons analyser trois approches :
 - **Naïve Bayes**
 - **SVM**
 - **Logistic Regression**

Naïve Bayes

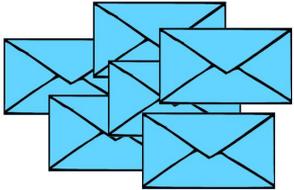
- La classification naïve bayésienne est un type de classification bayésienne probabiliste simple basée sur le théorème de Bayes avec une forte indépendance des hypothèses :



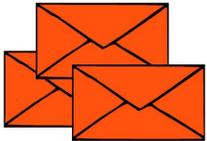
- Un classificateur bayésien naïf suppose que l'existence d'une caractéristique pour une classe, est indépendante de l'existence d'autres caractéristiques.

Naïve Bayes – exemple filtrage de spam

- Pré-traitement: on compte pas les mots vides ("a", "in", "the", etc.)
- Données d'entraînement 6 NORMAL mails, 3 SPAM mails
- $|V|=4$, $V=\{\text{"hello"}, \text{"friend"}, \text{"book"}, \text{"money"}\}$



6 NORMAL



3 SPAM

"hello" – 4 fois, "friend" – 3 fois, "book" – 2 fois, "money" – 1 fois, → 10 mots en total

$$P(\text{"hello"}|\text{NORMAL})=4/10=0.40$$

$$P(\text{"friend"}|\text{NORMAL})=3/10=0.30$$

$$P(\text{"book"}|\text{NORMAL})=2/10=0.20;$$

$$P(\text{"money"}|\text{NORMAL})=1/10=0.10$$

"hello" – 3 fois, "friend" – 2 fois, "book" – 0 fois, "money" – 4 fois, → 9 mots en total

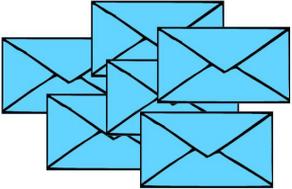
$$P(\text{"hello"}|\text{SPAM})=3/9=0.33$$

$$P(\text{"friend"}|\text{SPAM})=2/9=0.22$$

$$P(\text{"book"}|\text{SPAM})=0$$

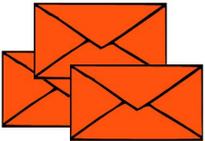
$$P(\text{"money"}|\text{SPAM})=4/9=0.44$$

Naïve Bayes – example Spam Filtering



6 NORMAL

$P(\text{"hello"}|N)=0,40$
 $P(\text{"friend"}|N)=0,30$
 $P(\text{"book"}|N)=0,20$
 $P(\text{"money"}|N)=0,10$



3 SPAM

$P(\text{"hello"}|S)=0,33$
 $P(\text{"friend"}|S)=0,22$
 $P(\text{"book"}|S)=0$
 $P(\text{"money"}|S)=0,44$

Naïve = Nous supposons que chaque mot d'une phrase est indépendant des autres.

$$P(N|\text{"hello friend"}) = \frac{P(\text{"hello friend"}|N) \times P(N)}{P(\text{"hello friend"})}$$

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

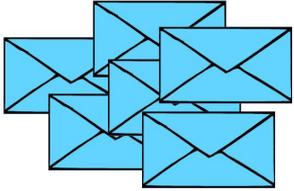
$$P(\text{"hello friend"}|N) = P(\text{"hello"}|N) \times P(\text{"friend"}|N) = 0,40 \times 0,30 = 0,12$$

$$P(N) = \frac{\text{\#Nombre N}}{\text{\#Nombre Total}} = \frac{6}{6+3} = 0,66$$

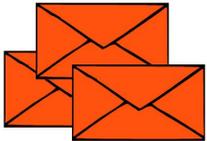
$$P(N|\text{"hello friend"}) = P(\text{"hello friend"}|N) \times P(N) = 0,12 \times 0,66 = \mathbf{0,07}$$

$$P(S|\text{"hello friend"}) = 0,33 \times 0,22 \times 0,33 = \mathbf{0,02}$$

Naïve Bayes – example Spam Filtering



- $P(\text{"hello"}|N)=0.40$
- $P(\text{"friend"}|N)=0.30$
- $P(\text{"book"}|N)=0.20;$
- $P(\text{"money"}|N)=0.10$



- $P(\text{"hello"}|S)=0.33$
- $P(\text{"friend"}|S)=0.22$
- $P(\text{"book"}|S)=0$
- $P(\text{"money"}|S)=0.44$

- Nouveau mail: "Book Money Money Money Money":

- $P(N)=6/(6+3)=0.66$
- $P(S)=6/(6+3)=0.33$

- $P(N) \times P(\text{"book"}|N) \times P(\text{"money"}|N)^4$
- $= 0.66 \times 0.20 \times 0.10^4 = \mathbf{0.0000132}$

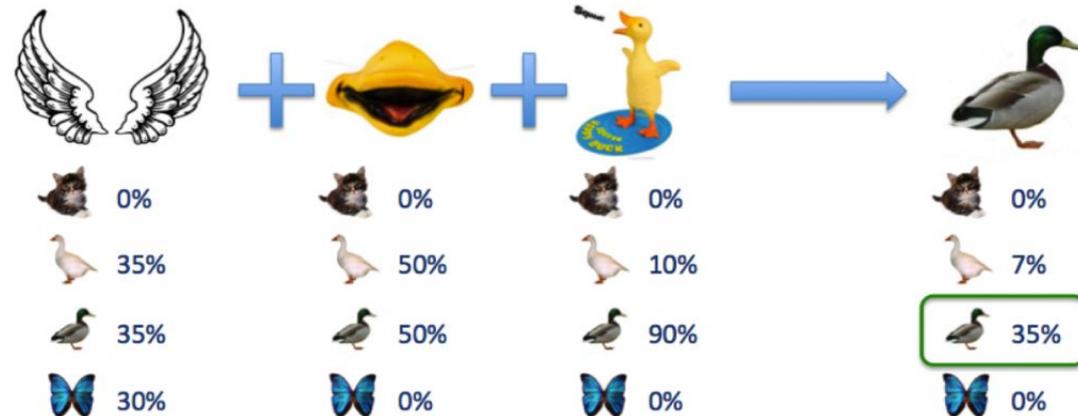
- $P(S) \times P(\text{"book"}|S) \times P(\text{"money"}|S)$
- $= 0.33 \times \mathbf{0} \times 0.44^4 = \mathbf{0}$

peu importe la valeur de la supposition initiale qu'il s'agissait de spam $P(S)$, peu importe que nous voyions le mot «money», tous les e-mails contenant le mot «book» seront toujours classés comme **NORMAL**.

N=NORMAL
S=SPAM

Naïve Bayes

- Un fruit peut être considéré comme une pomme s'il est rouge, arrondi, et fait une dizaine de centimètres.
- Même si ces caractéristiques sont liées dans la réalité, un classificateur bayésien naïf déterminera que le fruit est une pomme en considérant indépendamment ces caractéristiques de couleur, de forme et de taille.

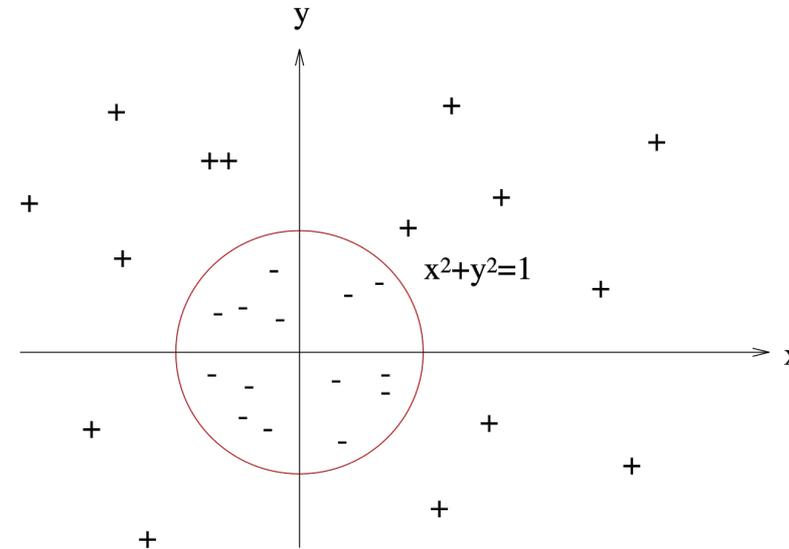


SVM : support vector machine

- Les SVMs sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de classification et de régression. Le but est de trouver un «hyperplan» qui pourrait séparer les données avec précision. Il pourrait y avoir de nombreux hyperplans de ce type → «hyperplan optimal» → **problème d'optimisation**
- Ces techniques reposent sur deux idées clés :
 - la notion de «marge» maximale : la distance entre la frontière de séparation («hyperplan») et les échantillons les plus proches («vecteurs support»)
 - la notion de fonction noyau («kernel trick») : les données qui ne sont pas linéairement séparables sont transformées pour être mappées dans un nouvel espace ($2d \rightarrow \varphi((a, b)) = (a, b, a^2 + b^2) \leftarrow 3d$, polynomial $d=2$)
- Les SVM maximisent la **marge** autour de l'**hyperplan de séparation**
- **Algorithme déterministe !**

SVM : support vector machine

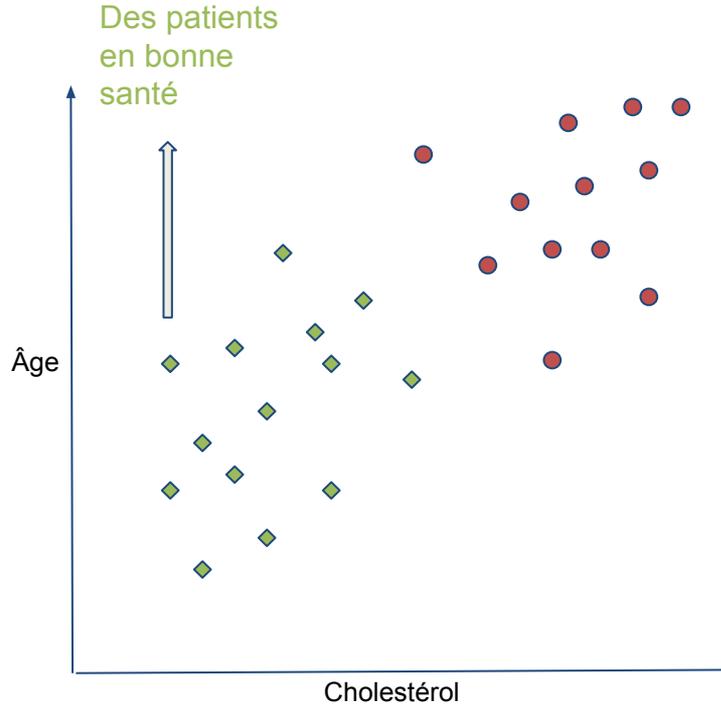
- La deuxième idée est de transformer l'espace de représentation des données d'entrées en un espace de plus grande dimension
- Il est probable qu'il existe une séparation plus simple entre les classes
- Les fonctions noyau permettent de transformer un produit scalaire dans un espace de grande dimension :
 - le noyau polynomial
 - le noyau gaussien



SVM : support vector machine

SVM linéaire → Trouver une surface de décision linéaire («hyperplan») qui peut séparer les classes de patients et qui a la plus grande distance (c'est-à-dire le plus grand «écart» ou «marge») entre les patients à la frontière (c'est-à-dire les «vecteurs support»)

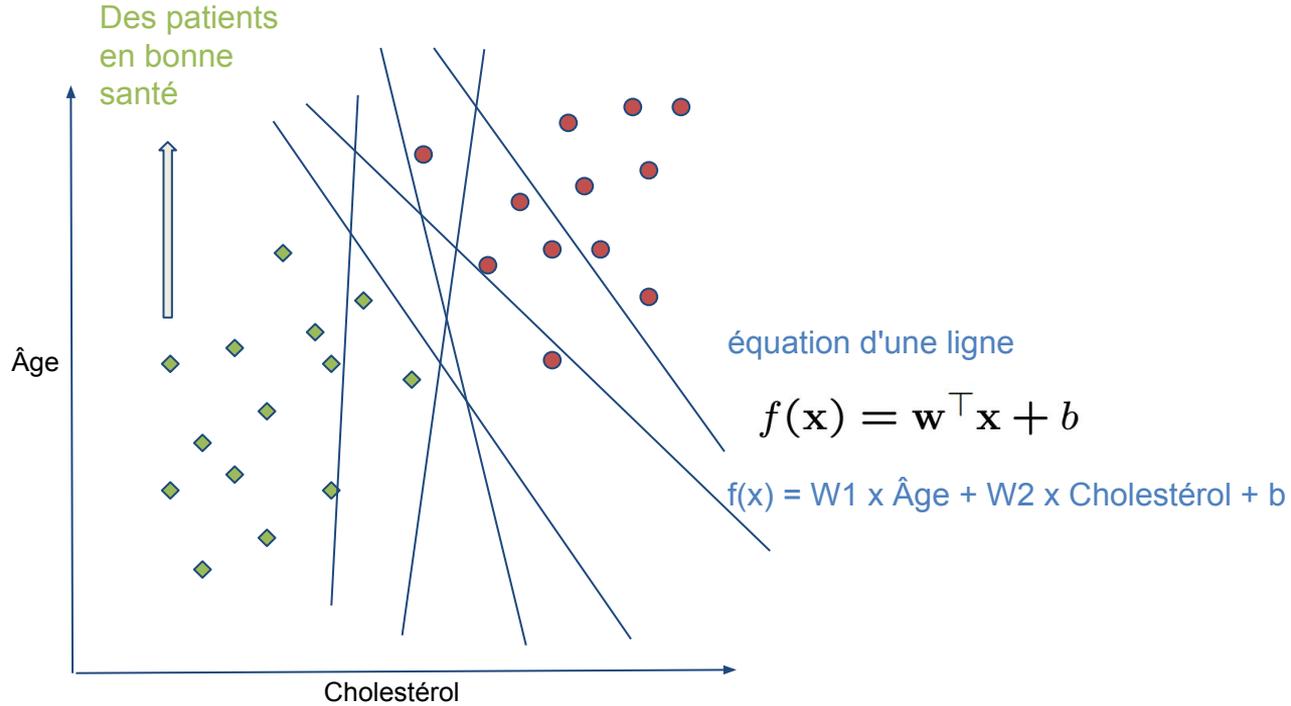
	Cholestérol	Âge	Classe
Patient 1	150	25	Pas malade
Patient 2	250	30	Malade
Patient 3	130	65	Malade
Patient 4	350	45	Malade
...			



SVM : support vector machine

SVM linéaire → Trouver une surface de décision linéaire («hyperplan») qui peut séparer les classes de patients et qui a la plus grande distance (c'est-à-dire le plus grand «écart» ou «marge») entre les patients à la frontière (c'est-à-dire les «vecteurs support»)

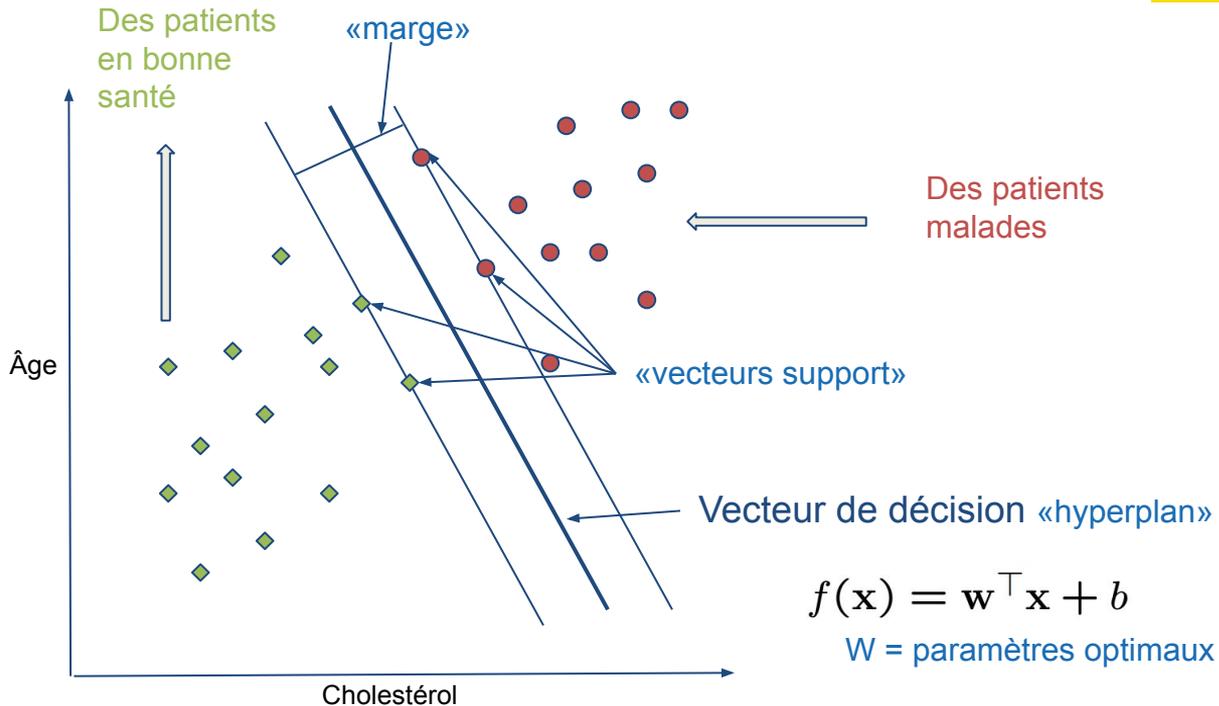
	Cholestérol	Âge	Classe
Patient 1	150	25	Pas malade
Patient 2	250	30	Malade
Patient 3	130	65	Malade
Patient 4	350	45	Malade
...			



SVM : support vector machine

SVM linéaire → Trouver une surface de décision linéaire («hyperplan») qui peut séparer les classes de patients et qui a la plus grande distance (c'est-à-dire le plus grand «écart» ou «marge») entre les patients à la frontière (c'est-à-dire les «vecteurs support»)

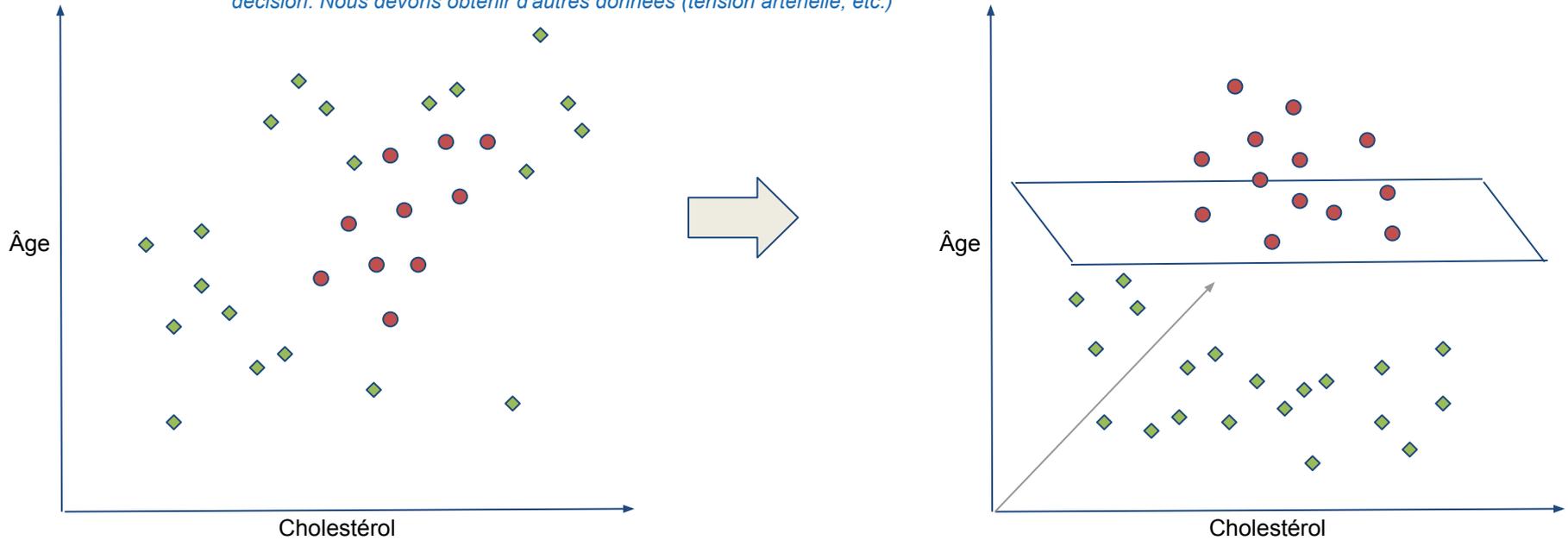
	Cholestérol	Âge	Classe
Patient 1	150	25	Pas malade
Patient 2	250	30	Malade
Patient 3	130	65	Malade
Patient 4	350	45	Malade
...			



SVM : support vector machine

Si une telle surface de **décision linéaire n'existe pas**, les données sont mappées dans un espace dimensionnel beaucoup plus élevé («espace de caractéristiques») où se trouve la surface de décision de séparation. L'espace des caractéristiques est construit via une projection mathématique intelligente («kernel trick», kernel polynomial ou gaussien).

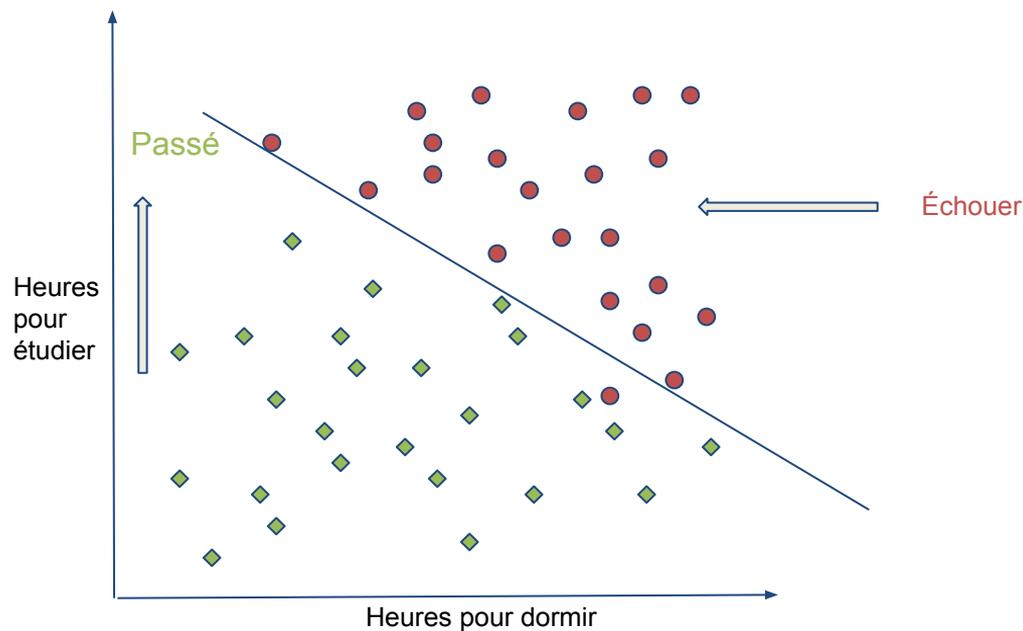
Cela peut également signifier que les caractéristiques ne sont pas suffisantes. L'âge et le cholestérol sont corrélés mais pas suffisants pour prendre une décision. Nous devons obtenir d'autres données (tension artérielle, etc.)



Logistic Regression

- La **régression logistique** est un algorithme de classification qui transforme sa sortie à l'aide de la fonction **sigmoïde** logistique pour donner une valeur de probabilité pour les classes de sortie.
- ~ réseau neuronal à 1 couche
- **Algorithme statistique !**

	Heures pour dormir	Heures pour étudier	Classe
Student 1	8	7	Passé
Student 2	12	5	Échouer
Student 3	10	3	Échouer
Student 4	9	8	Passé
...			



Logistic Regression

- La **régression logistique** à trouver les coefficients optimaux, de sorte que l'erreur globale (fonction de coût) soit minimisée, par descente de gradient. LR transforme sa sortie en utilisant la fonction sigmoïde logistique $\sigma = \frac{1}{1 + e^{-z}}$ pour renvoyer une valeur de probabilité. $p \geq 0.5, class = 1$
 $p < 0.5, class = 0$

Attribuer une probabilité à chaque résultat :

$$P(y = 1|x) = \sigma(w^T x + b)$$

S'entraîner pour maximiser les probabilités. LR une fonction de coût (perte) appelée Cross-Entropy

y = classe correcte, p = probabilité de la classe prédite $-(y \log(p) + (1 - y) \log(1 - p))$

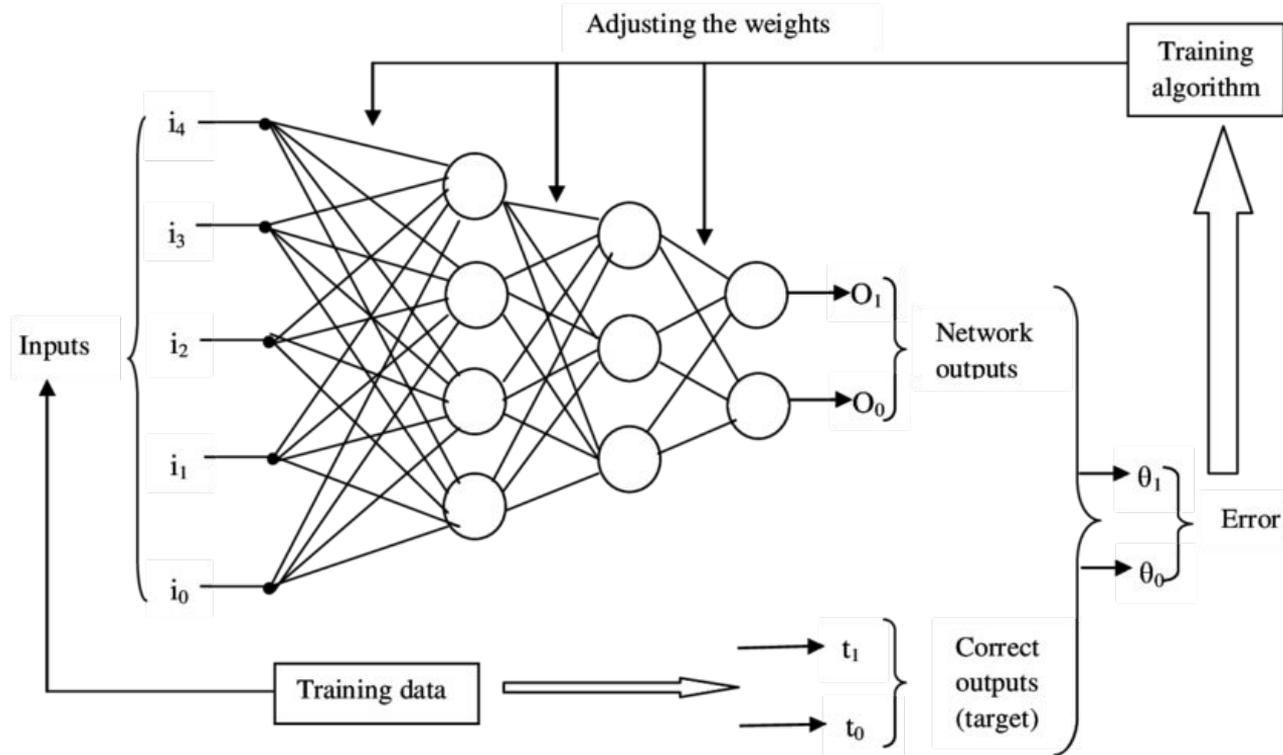
$$l(w) = - \sum_{n=1}^N \sigma(w^T x_n + b)^{y_n} (1 - \sigma(w^T x_n + b))^{(1-y_n)}$$

Pour minimiser les coûts (pertes ~ erreurs), nous utilisons la descente de gradient (Gradient Descent)

Descente de gradient (stochastique) connaît un très grand intérêt aujourd'hui, en particulier pour l'entraînement des réseaux de neurones profonds (deep learning).

Logistic Regression

(descente de gradient)



Quand utiliser ces modèles

En fonction du nombre d'ensembles d'entraînement (données) / caractéristiques dont vous disposez, vous pouvez choisir d'utiliser la régression logistique, SVM ou Naïve Bayes.

Généralement, Naive Bayes est bon mais trop naïf et a de faibles performances.

Pour les autres algorithmes, considérons :

- n = nombre de caractéristiques
- m = nombre d'exemples (textes/images)

1. Si n est grand (1 à 10 000) et m est petit (10 à 1 000): utilisez la régression logistique ou SVM linéaire
2. Si n est petit (1 - 10 00) et m est intermédiaire (10 - 10 000): utilisez SVM avec noyau (gaussien, polynomial, etc.)
3. Si n est petit (1 - 10 00), m est grand (50 000 - 1 000 000 +): tout d'abord, ajoutez manuellement plus de caractéristiques, puis utilisez la régression logistique ou SVM linéaire
4. Si n est grand (1 à 10 000), m est grand (50 000 - 1 000 000 +): utiliser des réseaux de neurones

Il est généralement conseillé d'essayer d'abord d'utiliser la régression logistique pour voir comment fonctionne le modèle. S'il échoue, vous pouvez essayer d'utiliser SVM sans noyau (autrement appelé SVM avec un noyau linéaire). La régression logistique et SVM avec un noyau linéaire ont des performances similaires mais en fonction de vos fonctionnalités, l'une peut être plus efficace que l'autre.

Classification de textes

Analyse de performance, analyse des erreurs

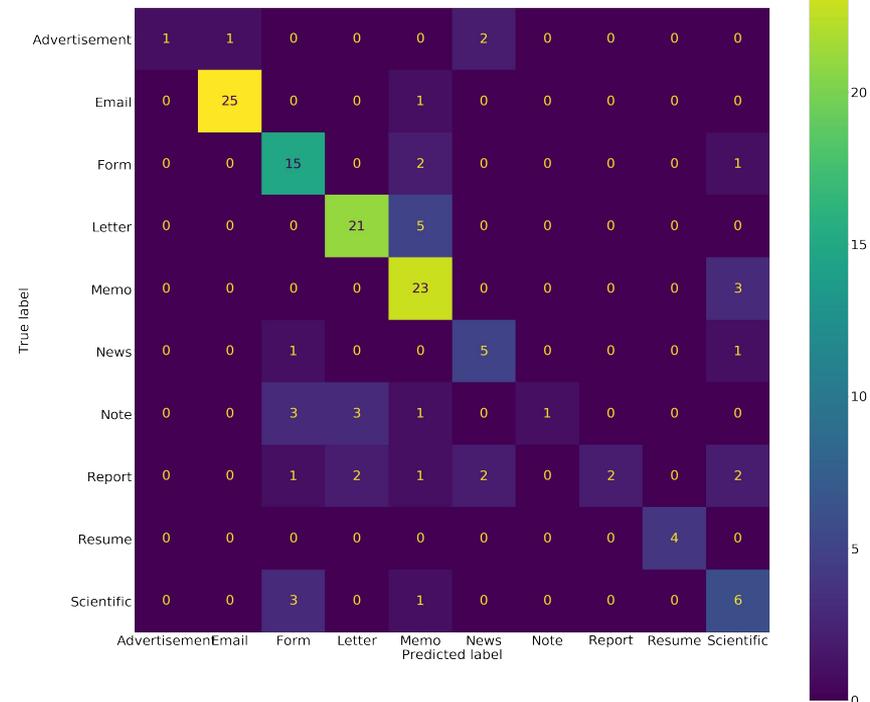
Analyse de la performance

- Après avoir entraîné notre modèle, nous pouvons faire la prédiction des classes des textes du jeux de données de test.
- La fonction f1-score calcule la performance d'une méthode à partir de l'analyse de la quantité de données qui ont été prédit correctement.
 - précision = la proportion des prédictions positifs était effectivement correcte
 - rappel = la proportion de résultats positifs réels a été identifiée correctement
 - F1 = la moyenne pondérée de la précision et du rappel (score final)

Matrice de confusion

- La matrice de confusion nous permettent de visualiser les résultats de la prédiction :

- Vrai positifs
- Faux positifs
- Vrai négatifs
- Faux négatifs



Jeux de données : Tobacco3482

Image

Texte

THE TOBACCO INSTITUTE

1875 J STREET, NORTHWEST
WASHINGTON, DC 20006
202/457-4800 • 800/898-4433

SAMUEL D. CHILCOTE, JR.
President

-- VIA FACSIMILE --

September 21, 1994

MEMORANDUM

TO: The Members of the Executive Committee

FROM: Samuel D. Chilcote, Jr. 

Administrative Law Judge Vitonne yesterday presided over the opening day of the Occupational Safety and Health Administration's (OSHA) scheduled 11-week hearings on the proposed indoor air quality standard. OSHA staff testified in the morning; the balance of the day was consumed by a cross-examination period.

Dr. Michael Silverstein, OSHA's director of policy, expressed the Agency's intent to use the information obtained at the hearing to "assure that the final rule is effective." Noting that ETS issues had received a disproportionate amount of pre-hearing attention, Silverstein stressed that tobacco is only one of many airborne "hazards" OSHA would address in the final rule. He denied any political motivation for the rulemaking.

John Martonik, acting director of health standards programs for OSHA, stressed that the regulation would not ban smoking, but only restrict smoking to separately exhausted smoking areas under negative pressure so that nonsmoking coworkers are "protected ... from the adverse health effects of ETS." Martonik estimated that 4.4 million buildings and 70.7 million workers would be affected by the rule. Compliance with the IAQ standard is estimated at \$8.1 billion.

Following the opening statements, nearly a dozen individuals acknowledged their intent to cross-examine OSHA staffers. The questions, primarily from tobacco industry representatives, sought to clarify matters ranging from the scope of the rule to the research the Agency cited to support the proposed action. Following one lengthy exchange, Agency staffers stated that the rule as written does not preempt state and local laws banning/restricting smoking.

CONFIDENTIAL:
TOBACCO LITIGATION

TICT 0008012

THE TOBACCO INSTITUTE

1875 J STREET, NORTHWEST SAMUEL D. CHILCOTE, JR.
WASHINGTON, DC 20006 President
202/457-4800 • 800/898-4433

@- VIA FACSIMILE -@

September 21, 1994

TO: The Members of the Executive Committee

FROM: Samuel D. Chilcote, Jr.

Administrative Law Judge Vitonne yesterday presided over the opening day of the Occupational Safety and Health Administration's (OSHA) scheduled 11-week hearings on the proposed indoor air quality standard. OSHA staff testified in the morning; the balance of the day was consumed by a cross-examination period.

Dr. Michael Silverstein, OSHA's director of policy, expressed the Agency's intent to use the information obtained at the hearing to "assure that the final rule is effective." Noting that ETS issues had received a disproportionate amount of pre-hearing attention, Silverstein stressed that tobacco is only one of many airborne "hazards" OSHA would address in the final rule. He denied any political motivation for the rulemaking.

John Martonik, acting director of health standards programs for OSHA, stressed that the regulation would not ban smoking, but only restrict smoking to separately exhausted smoking areas under negative pressure so that nonsmoking coworkers are "protected ... from the adverse health effects of ETS." Martonik estimated that 4.4 million buildings and 70.7 million workers would be affected by the rule. Compliance with the IAQ standard is estimated at \$8.1 billion.

Following the opening statements, nearly a dozen individuals acknowledged their intent to cross-examine OSHA staffers. The questions, primarily from tobacco industry representatives, sought to clarify matters ranging from the scope of the rule to the research the Agency cited to support the proposed action. Following one lengthy exchange, Agency staffers stated that the

rule as written does not preempt state and local laws banning/restricting smoking.

CONFIDENTIAL:
TOBACCO LITIGATION

TICT 0008012

Le gouvernement américain a attaqué en justice cinq grands groupes américains du tabac pour avoir amassé d'importants bénéfices en mentant sur les dangers de la cigarette.

Dans ce procès 6 910 192 de documents ont été collectés et numérisés. Afin de faciliter l'exploitation de ces documents par les avocats, vous êtes en charge de mettre en place une classification automatique des types de documents: **Advertisement, Email, Form, Letter, Memo, News, Note, Report, Resume, Scientific.**

Merci beaucoup pour votre attention!

Si vous avez de questions, vous pouvez nous contacter sur Discord

Aller plus loin:

<https://www.kaggle.com/competitions>

Machine learning Coursera famous courses, Andrew Ng,
<https://www.coursera.org/learn/machine-learning>

Machine learning Coursera (on youtube), Andrew Ng, <https://www.youtube.com/watch?v=PPLop4L2eGk>

The most famous book on deep learning: <https://www.deeplearningbook.org/> (Ian Goodfellow, Yoshua Bengio and Aaron Courville)