L'apprentissage automatique avec Python

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle (IA). Mais qu'est-ce que l'intelligence artificielle ?

Andrew Moore, ancien doyen de l'école d'informatique de l'université Carnegie Mellon, l'a définie comme suit : "L'intelligence artificielle est la science et l'ingénierie qui consiste à faire en sorte que les ordinateurs se comportent d'une manière dont, jusqu'à récemment, nous pensions qu'elle nécessitait une intelligence humaine."

La question "Qu'est-ce que l'intelligence artificielle ?" dépend de la réponse à une question plus générale : "Qu'est-ce que l'intelligence ?"

Il est extrêmement difficile de répondre à la question précédente.

Pour se rapprocher des réponses, nous pouvons diviser l'IA en deux parties :

l'IA faible et l'IA forte

IA faible :

- traite de problèmes d'application spécifiques
- soutient la pensée humaine dans certains domaines
- capable d'apprendre dans des sous-domaines
- aucune conscience

IA forte :

- "intelligence générale" (raisonnement, pensée logique, utilisation de stratégies, résolution d'énigmes et jugement dans l'incertitude)
- Comparable à l'intelligence humaine, mais n'est pas nécessairement la même, peut être différente
- faire des plans
- généralement capable d'apprendre
- Aptitudes à la communication, langage naturel
- Conscience ?
- sensibilité, émotions ?
- perception de soi ?

Nous connaissons maintenant l'intelligence artificielle, l'IA faible et l'IA forte, mais qu'en est-il de l'apprentissage automatique ?

Commençons par une très "ancienne" tentative de définition par Arthur Samuek, un pionnier d'IBM :

"Apprentissage automatique : Domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés."

Une bonne tentative, mais de nombreuses questions restent sans réponse. Près de 40 ans plus tard, en 1998, Tom Mitchell a façonné un "problème d'apprentissage bien posé" comme suit :

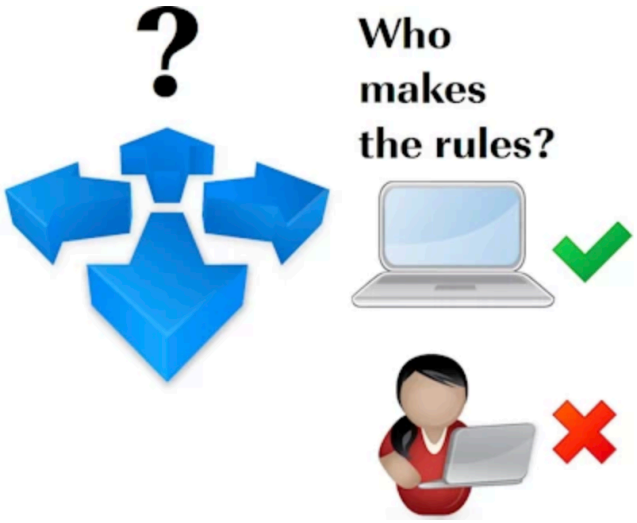
"Problème d'apprentissage bien posé : on dit qu'un programme informatique apprend de l'expérience E en ce qui concerne une certaine tâche T et une certaine mesure de performance P, si sa performance sur T, telle que mesurée par P, s'améliore avec l'expérience E."

Annotation : Un problème mathématique est dit correctement (aussi bien posé, bien posé ou correctement posé) si les conditions suivantes sont réunies :

- Le problème a une solution (existence).
- Cette solution est clairement définie (unicité).
- Le comportement de la solution change continuellement avec les données d'entrée initiales (stabilité).

Apprentissage automatique :

L'apprentissage automatique signifie qu'un algorithme (la machine) apprend automatiquement. Cela signifie qu'il est capable d'extraire automatiquement les connaissances nécessaires à partir de données données données. L'objectif est de faire des prédictions sur des données nouvelles, non vues. Il existe une autre façon de présenter les choses : Dans les algorithmes traditionnels de prise de décision heuristique, les programmeurs fixent les règles selon lesquelles les décisions sont prises. Avec l'apprentissage automatique, le programme le fait de manière indépendante, sans intervention humaine !



Taxonomie de l'apprentissage automatique

Il existe deux approches différentes de l'apprentissage automatique :

- l'apprentissage non supervisé
- L'apprentissage supervisé

Nous ne traiterons que de l'apprentissage supervisé dans ce tutoriel.

Exemples pour l'apprentissage automatique :

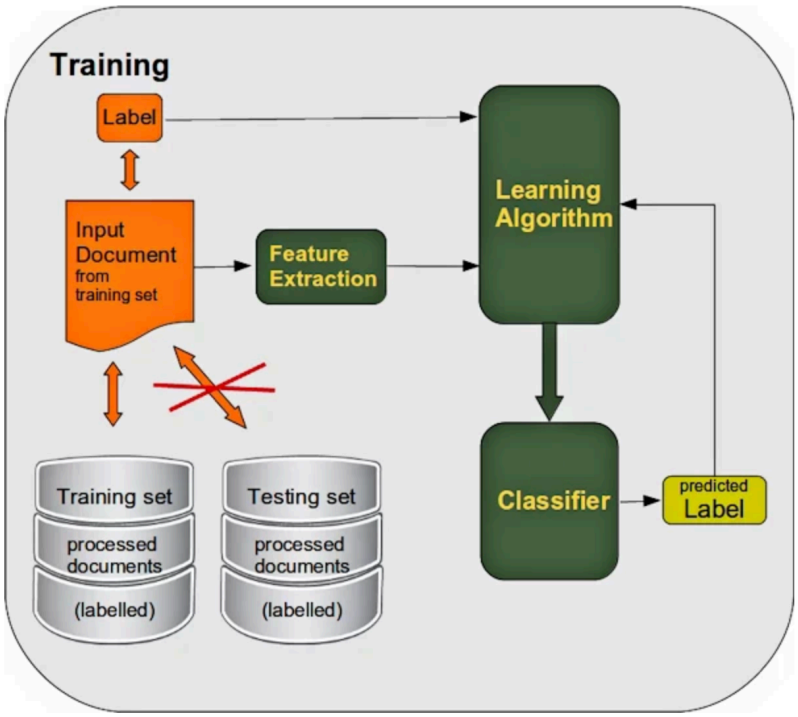
- filtre anti-spam : l'algorithme apprend un modèle prédictif à partir de données étiquetées comme spam et "non spam" (ham). Après l'apprentissage, il peut prédire pour les nouveaux e-mails s'ils sont des spams ou non.
- reconnaissance de caractères
- reconnaissance d'objets dans les images
- et bien d'autres encore

Comme nous l'avons déjà mentionné, un filtre anti-spam peut être mis en œuvre à l'aide d'un classificateur basé sur l'apprentissage automatique.

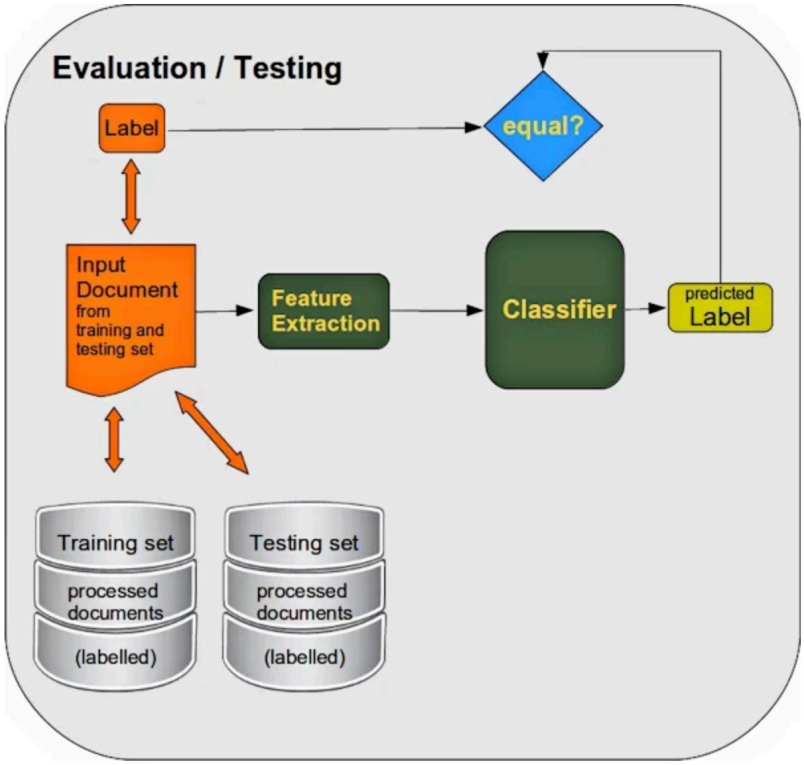
Au cœur de l'apprentissage automatique se trouve le concept d'automatisation de la prise de décision à partir de données sans que l'utilisateur ne spécifie de règles explicites sur la manière de prendre cette décision. Dans le cas des courriels, l'utilisateur ne fournit pas une liste de mots ou de caractéristiques qui constituent un spam. Au lieu de cela, l'utilisateur fournit des exemples d'e-mails spam et non spam qui sont marqués comme tels. C'est ce qu'on appelle l'ensemble d'apprentissage.

L'objectif d'un modèle d'apprentissage automatique est de prédire de nouvelles données, auparavant invisibles. Dans une application réelle, nous ne sommes pas intéressés par le fait de marquer un e-mail déjà marqué comme spam ou non. Nous voulons plutôt faciliter la vie des utilisateurs en classant automatiquement les nouveaux courriels entrants.

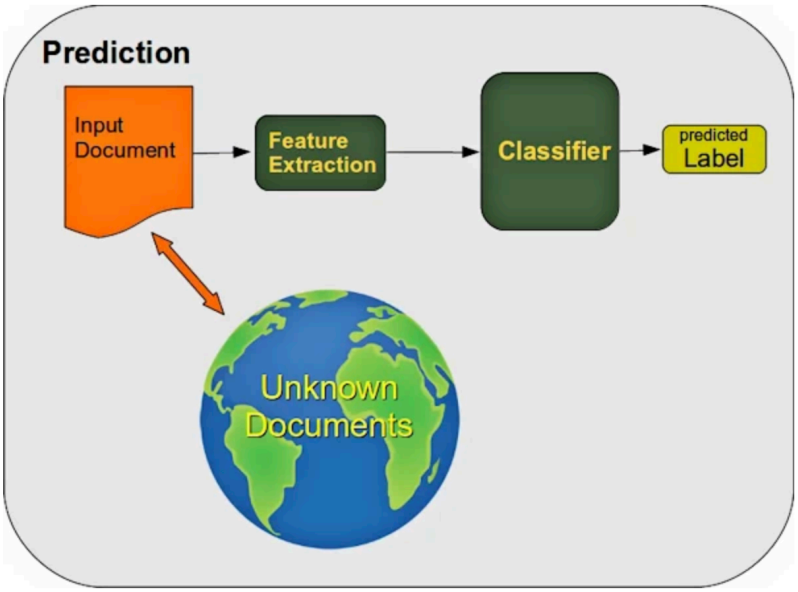
Ces exemples sont ensuite appris ou entraînés par l'algorithme :



Après la phase d'apprentissage, nous devons évaluer le classificateur. Nous testons à la fois sur des données d'apprentissage étiquetées et sur des données de test étiquetées non apprises :



Si nous sommes satisfaits des résultats, le classificateur est prêt à classer des documents complètement nouveaux :



Les données sont présentées à l'algorithme généralement sous la forme d'un tableau bidimensionnel (ou matrice) de nombres. Chaque point de données (également connu sous le nom d'échantillon ou d'instance de formation) à partir duquel nous voulons apprendre ou prendre une décision est représenté par une liste de nombres, appelée vecteur de caractéristiques, et les caractéristiques qu'il contient représentent les propriétés de ce point.

Entrée [ ]: