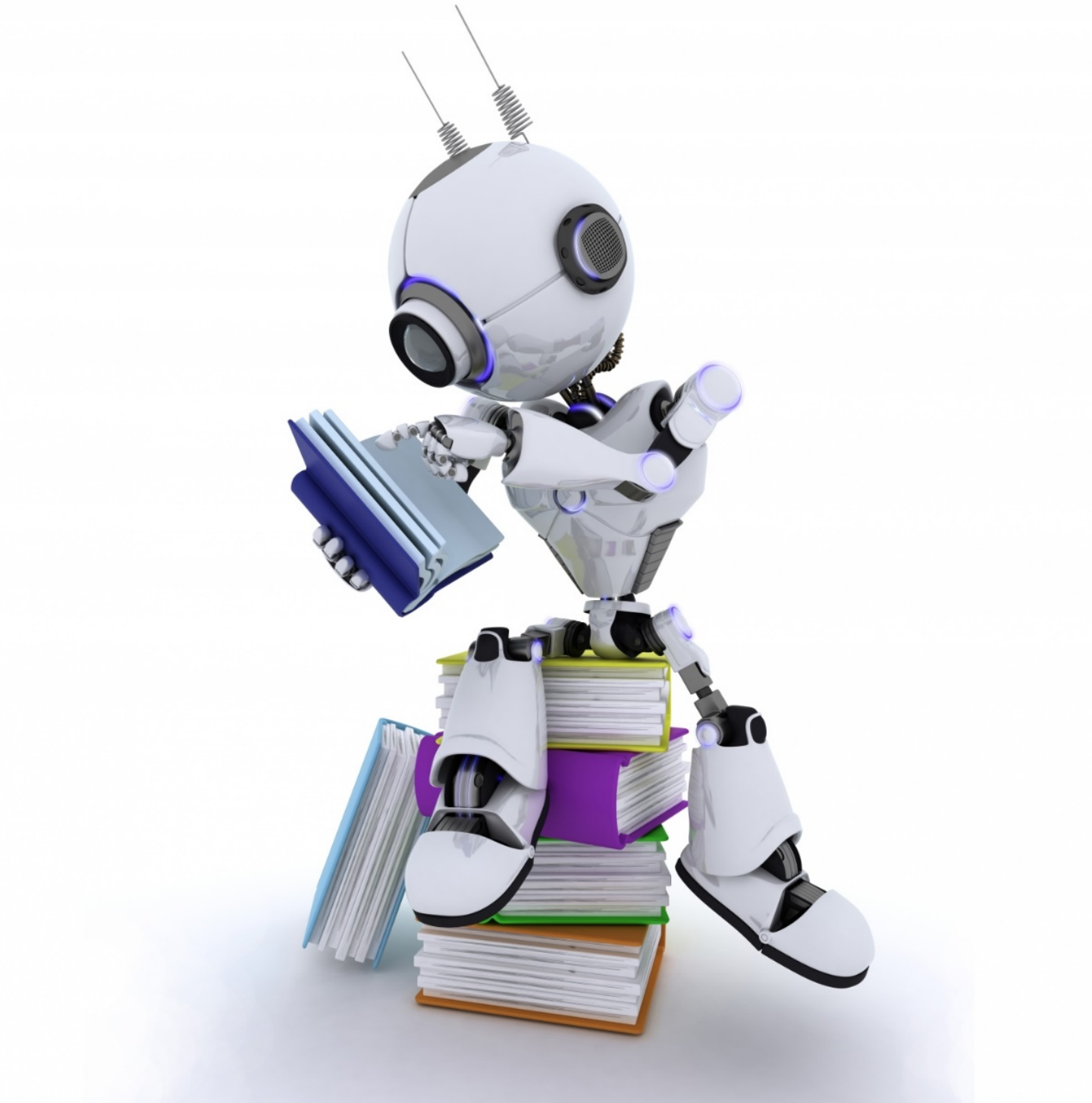
COMSOFT direct

Bechtle's Software Specialist

inside loop

Machine Learning & Big Data, pourquoi pas moi ?

François Fayard - Fondateur & Consultant, Inside Loop
Spécialiste du Calcul Haute Performance



Machine Learning & Big Data, pourquoi pas moi ?

La mode est aujourd'hui au Machine Learning et au Big Data, ces techniques dont tout le monde parle et qu'il ne faudrait surtout pas laisser passer. Les géants du Web, Google, Apple, Facebook et Amazon l'utilisent depuis bientôt 10 ans pour que leurs serveurs puissent détecter les images inappropriées ou effectuer de la reconnaissance vocale. Ces GAFA se permettent même de venir chasser sur les terres des constructeurs automobiles et développer la voiture autonome qui pourrait être à l'origine de la prochaine révolution des transports. Bien que ces technologies ne soient pas réservées aux géants du Web, il est souvent difficile de savoir ce qu'elles peuvent nous apporter. Entre l'attitude qui consiste à ne rien faire et celle qui consiste à accumuler des montagnes de données dans l'espoir vain de les utiliser un jour, il y a un juste milieu qu'il est difficile d'appréhender. Dans cet article, nous commencerons par vous donner quelques succès du Machine Learning dans des domaines aussi divers que la santé ou la production d'énergie. Sur la base de ces exemples, nous définirons ce qu'on appelle le Machine Learning. Puis nous rappellerons son histoire, des premiers perceptrons aux réseaux de neurones multicouches en passant par les arbres de décision et autres "forêts aléatoires" . Cela nous permettra de comprendre pourquoi ces technologies qui sont nées dans les années 1950s deviennent si importantes aujourd'hui. On introduira aussi les différents types d'apprentissages : supervisé, non supervisé et par renforcement. Nous vous expliquerons ensuite comment entrer dans la danse, notamment grâce aux outils open source disponibles dans la « Intel Python Distribution » qui a été optimisée pour les plateformes Intel. Ensuite, nous mettrons en avant les conséquences de la progression du Machine Learning dans votre écosystème. Le rapport entre les hommes et les données ne sera plus direct, mais se fera par l'intermédiaire de ces nouveaux algorithmes. Données dont la qualité est beaucoup plus importante que la quantité. Données qu'il faudra entretenir et enrichir au fil du temps plutôt que d'accumuler sans but précis. Enfin, nous présenterons les challenges à venir pour les utilisateurs du Machine Learning et quelles sont les solutions sur lesquelles travaillent les chercheurs.

Le Machine Learning et ses succès

C'est souvent lors d'une visite chez un dermatologue que sont détectés les mélanomes à l'origine d'un cancer de la peau. Bien que ce type de cancer puisse être traité de manière efficace s'il est détecté rapidement, une détection tardive peut faire passer une chance de survie à 5 ans de 97% à 14%. Une détection rapide est donc essentielle pour le patient qui ne peut pas toujours voir un dermatologue au moindre doute.

Des chercheurs de l'université californienne de Stanford ont récolté une base de 130 000 images de maladies de peau, chaque image étant classée selon le type de maladie diagnostiquée. Cette base de données a été utilisée pour entraîner un réseau de neurones convolutif, un algorithme de Machine Learning très souvent utilisé dans la reconnaissance d'image.

Le but est de discerner un mélanome d'une maladie de peau bénigne. En adaptant ce réseau de neurones initialement développé par Google pour faire la différence entre des chats et des chiens, les chercheurs ont obtenu un taux d'erreur similaire à un dermatologue et pensent qu'il est possible d'implémenter cet algorithme dans un smartphone. D'autres tests sont bien sûr nécessaires, mais la possibilité d'avoir un diagnostic fiable pour des personnes situées dans des déserts médicaux est désormais toute proche. La prévision de la demande d'électricité représente un enjeu majeur pour EDF. En effet, contrairement aux hydrocarbures, l'électricité ne se stocke pas et la production doit être constamment en équilibre avec la consommation. Afin d'optimiser ses moyens de production, l'entreprise doit bénéficier de bonnes prédictions sur des échelles de temps allant du jour à plusieurs années.

En fonction de l'historique de la consommation électrique, des prévisions météo et du prix de l'électricité, EDF utilise le Machine Learning pour prévoir chaque jour la consommation électrique du lendemain. Ce sont ces prévisions qui permettent à EDF de prendre des décisions quant à la mise en marche de ses différents moyens de production, de l'éolienne à la centrale nucléaire.

La généralisation des compteurs intelligents Linky va permettre à EDF de récolter un grand nombre de données et affiner ses prédictions. Ces problèmes de prévision sont communs à de nombreuses industries et les mêmes outils peuvent être utilisés pour la gestion des stocks des grands distributeurs.

Le Machine Learning consiste donc à apprendre à un ordinateur à prendre des décisions à partir d'exemples. Dans ce qu'on appelle l'apprentissage supervisé, on associe à une donnée d'entrée une information simple qui permettra ensuite de prendre la décision. Dans le cas de la détection des mélanomes, la donnée d'entrée est une photo d'un grain de beauté et l'information est de savoir si un mélanome est probable ou non. La décision qui doit être prise est d'effectuer ou non une biopsie. Les données d'entrées peuvent être des données structurées comme des résultats d'analyse de sang ou des données non structurées comme des images, des vidéos, du texte, des graphes, ou différents signaux. L'information de sortie peut être un choix binaire (malin/bénin), une classification (type de maladie de peau) ou un compte rendu si l'algorithme sait générer du texte en langage naturel.

Déjà 60 ans d'histoire

Plusieurs vagues ont façonné le Machine Learning d'aujourd'hui. La première vague date de 1957 où le premier réseau de neurones est entraîné afin de faire de la reconnaissance sur des images de 20 pixels par 20 pixels. L'entraînement du réseau se fait alors par des moteurs électriques. Cependant, bien que prometteur, on se rend compte rapidement que ces réseaux à une couche ne permettent pas de différencier des structures complexes. C'est durant les années 1980s que Yann Le Cun, alors chercheur aux Bell Labs utilise avec succès l'algorithme de backpropagation sur un réseau de neurones à plusieurs couches afin de faire de la reconnaissance de chiffres. En 1996, le Crédit Mutuel de Bretagne est la première banque à équiper ses distributeurs automatiques d'un mécanisme permettant de déposer et traiter automatiquement des chèques grâce au réseau de neurones LeNet. Parallèlement, les machines à vecteur de support se développent durant les années 1990s.

Puis en 2001, Leo Breinman propose les forêts aléatoires qui sont toujours très utilisées comme le prouve le nombre de victoires aux compétitions de Machine Learning Kaggle. Que ce soit pour prédire la qualité de l'air, ou pour prédire une activité biologique de composants en fonction de leur structure moléculaire, les succès de ces algorithmes sont nombreux.

Mais pourquoi le Machine Learning qui prend ses racines dans des travaux qui ont maintenant 60 ans a connu une progression fulgurante ces 10 dernières années ? Une des raisons est que ces techniques sont désormais disponibles à tout le monde, notamment grâce au développement des bibliothèques open source accessibles en Python comme Scikit-Learn, Caffe, TensorFlow ou Keras. Intel qui ne se définit plus seulement comme un fabricant de processeurs, mais comme une entreprise de la donnée, a d'ailleurs optimisé toutes ces bibliothèques afin qu'elles tirent le meilleur de ses processeurs. On les retrouve dans la « Intel Python Distribution » disponible gratuitement sous Linux. Des versions Windows et macOS de cette distribution sont aussi disponibles mais nous vous conseillons fortement d'utiliser Linux pour avoir l'ensemble des outils. En quelques lignes de code, il est désormais possible de détecter des commentaires négatifs parmi des commentaires clients faits sur un site internet. Une autre révolution qui a permis l'émergence du Machine Learning est la révolution matérielle. À la fin des années 2000s, les GPU utilisés sur nos ordinateurs pour accélérer les calculs d'images en 3D sont détournés de leur utilisation pour entraîner des réseaux de neurones. Leur capacité à multiplier des matrices denses très rapidement permet de faire passer le temps d'entraînement des réseaux convolutifs d'une semaine de calcul à moins d'une journée. Pendant les années 2010s, les processeurs, plus versatiles, gagnent eux aussi des cœurs ainsi que des instructions comme les instructions vectorielles sur 256 puis 512 bits qui permettent d'accélérer ces calculs. En 2016, Intel rachète Nervana et prépare la sortie de Lake Crest, un accélérateur conçu spécialement pour les réseaux de neurones et dont la conception pourrait permettre de surclasser les GPUs à la fois en termes de puissance de calcul et en consommation énergétique.

On distingue aujourd'hui trois types de Machine Learning : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

Dans l'apprentissage supervisé, un humain ou un système a par exemple annoté une collection d'images afin de savoir si elles représentent un chat ou un chien. La machine va apprendre à partir de cette collection afin de prédire sur de nouvelles images si elle a affaire à un chat ou un chien.

L'apprentissage non supervisé est quant à lui utilisé par exemple dans la détection d'anomalies. Supposons que vous ayez une grande base de données de détections bancaires et que vous souhaitez faire de la détection de fraudes. Bien entendu, vous ne savez pas quelles sont les transactions frauduleuses et celles qui ne le sont pas. C'est ce qu'on appelle un problème non supervisé. De même, si vous avez collecté des données en très grande dimension, par exemple pour les personnes travaillant avec des données génétiques, et que vous voulez en comprendre la structure en les visualisant par exemple, vous devez faire ce qu'on appelle une réduction de dimensionnalité, problème qui entre dans l'apprentissage non supervisé. Si vous souhaitez trouver dans votre base de clients, les groupes de clients qui ont des comportements similaires, vous utiliserez des méthodes dites de clustering encore un algorithme d'apprentissage non supervisé.

L'apprentissage par renforcement permet quant à lui une prise de décision séquentielle dans un environnement certain comme le jeu de go où on voit le plateau et donc les mouvements de l'adversaire ou incertain comme le poker où on ne voit pas les cartes de l'adversaire.

Pour entrer dans la danse du Machine Learning, il ne faudra jamais oublier que cette science est avant tout expérimentale. Inutile de chercher des grands théorèmes mathématiques pour savoir si telle ou telle idée va fonctionner. Il vous faudra essayer, échouer et réessayer à nouveau avant de connaître vos premiers succès. Après avoir installé « Intel Python Distribution » sur votre ordinateur, commencez à jouer avec Scikit-Learn pour le Machine Learning classique ainsi que Keras pour le Deep Learning. Prenez d'abord des données déjà utilisées. Vous pourrez ensuite essayer les mêmes méthodes sur vos propres données. Mais l'apprentissage ne sera possible que par l'exemple. Pour cela, Intel propose des workshops gratuits d'une journée dans toutes les villes d'Europe. Ces workshops passeront notamment par Paris et Toulouse. Pour approfondir vos connaissances, de nombreuses formations payantes sont proposées par différents organismes comme l'Ecole Polytechnique ou des sociétés indépendantes.

"La distribution Python optimisée par Intel est disponible gratuitement sur https://software.intel.com/en-us/distribution-for-python"

Ses liens avec le Big Data

Le Machine Learning change radicalement le lien que nous entretenons avec les données. Avant, pour connecter les données à l'ordinateur, l'humain était indispensable. Aujourd'hui c'est l'ordinateur qui va valoriser, exploiter et mettre en forme ces données pour que l'humain puisse les visualiser de manière efficace afin de prendre les meilleures décisions. Aujourd'hui, l'humain voit donc les données à travers le travail de l'ordinateur.

Le Machine Learning est donc intimement lié au Big Data. Cependant, il ne faudrait pas croire qu'amasser un maximum de données sans se soucier de leur qualité vous aidera à faire des choix judicieux. La question centrale est de savoir quand on a assez de données pour faire quelque chose d'utile et de pertinent d'un point de vue business. Et quelquefois les résultats sont surprenants. Par exemple, en 2012, lors de la seconde élection de Barack Obama, un data analyst, Nate Silver, a prédit le résultat des élections américaines dans chacun des états. Les médias se sont empressés de relayer l'information et souligné que Nate Silver avait utilisé le « Big Data » pour prédire ces résultats. La vérité est que les données utilisées par Nate Silver ne prennent que 188 kB soit à peine 15% de la capacité d'une bonne vieille disquette. On peut donc faire des systèmes de prédiction fiables avec des données très petites.

Bien sûr, si nous disposons d'une très grande base de données de très bonne qualité, ce sont les systèmes simples qui auront la meilleure performance. Quand ce nombre de données diminue, il faudra alors utiliser des méthodes d'apprentissage plus complexes afin de mieux exploiter ces données. Quand a-t-on suffisamment de données ? Il n'y a pas de réponse toute faite à cette question, et tant que vous n'avez pas essayé, vous ne pourrez pas le savoir. En pratique, il vous faudra donc faire beaucoup d'expérimentation avec vos données avant de connaître leur vraie valeur.

L'essentiel est de réaliser que la qualité de vos données est plus importante que leur quantité. On dit souvent qu'il y a un océan de données, mais que la plupart de ces données sont imbuvables. Une question majeure dans votre business est de savoir comment les données que vous allez collecter vivent et s'améliorent de jour en jour. Les données ne sont pas statiques et doivent passer des étapes de nettoyage, de curation et d'annotation avant d'être utilisées. Elles doivent être préparées, modélisées, observées, puis repréparées dans un cycle dynamique de vie des données. En général, peu de données de qualité permettent de savoir si une idée va marcher raisonnablement. Ce peu de données permet souvent de savoir si elles contiennent du signal et permettent d'ajuster le type de données à collecter avant de se lancer dans une récolte de plus grande ampleur.

Les challenges pour les prochaines années

Le Machine Learning est une perpétuelle évolution et de nombreux challenges se présentent à nous.
Le premier challenge est l'hypothèse de stationnarité. Le Machine Learning est utilisé pour généraliser à de nouvelles données ce que vous avez appris sur votre base d'apprentissage. Cependant, il est tout à fait possible que la distribution de données sur laquelle vous allez utiliser votre algorithme n'ait pas du tout la même distribution que votre base d'apprentissage. Cela peut être le cas par exemple si vous entraînez votre algorithme avec les résultats venant d'une machine médicale fournie par un constructeur et que vous essayez de faire de la prédiction avec des données venant d'une machine venant d'un autre constructeur. Si vous faites cela, vous violez ce qu'on appelle la « stationnarité » de vos données. Un des défis actuels des chercheurs en Machine Learning est de développer des algorithmes robustes à des changements de distribution de vos données.
Un second problème est celui de l'interprétabilité. Plus on a de données, plus on construit des systèmes dont l'interprétabilité est difficile. Autrement dit, plus les systèmes sont complexes, moins on est capable d'expliquer pourquoi un système prend une certaine décision. C'est pourtant fondamental lorsqu'on prédit ce qu'une personne va devoir payer pour une assurance. Si le prix est élevé, on va devoir expliquer à cette personne les raisons de ce prix. Un autre problème est de savoir comment apprendre dans des contextes où on a très peu de données. Autant les données du Web comme les images, les photos sont surabondantes, autant des données médicales peuvent être rares car très chères à obtenir. On va donc devoir construire des méthodes d'apprentissage efficaces avec peu de données. Une des approches pour résoudre ce problème est ce qu'on appelle le “transfer learning”.

À vous de jouer

En conclusion, il faut se souvenir que la science des données est une science expérimentale. Personne ne pourra prévoir si une idée va marcher ou non avant d'essayer. Il vous faudra donc essayer sur vos propres données, pourquoi pas avec un petit jeu de données de qualité. Une fois ces expérimentations faites, vous pourrez collecter un plus grand jeu de données. Mais ces données devront être enrichies au fur et à mesure de leur vie dans votre base de données. Pour les traiter, de nombreux outils open source sont disponibles comme la « Intel Python Distribution » qui regroupe les bibliothèques de Machine Learning développées par Google, Facebook ou l'Inria et optimisées par Intel pour marcher au mieux sur vos processeurs Intel. Pour progresser, rien ne remplacera l'expérimentation, mais n'oubliez pas que de nombreux organismes de formations sont disponibles en France pour vous faire progresser sur cette voie.

François Fayard (fayard@insideloop.io), dirigeant d'Inside Loop, Data & Computational Experts.

Merci à Alexandre Gramfort et Pierre Gaillard, chercheurs en Machine Learning à l'INRIA, pour leur aide précieuse lors de la rédaction de cet article