Etat de l'art en Intelligence Artificielle et Machine Learning appliqué à l'analyse statistique

12 FÉVRIER 2021

Université Bretagne Sud

Crée par : Groupe CMI (Licence, Master)

Introduction à l'IA et au ML

Ce rapport a pour but de faire la synthèse sur les pratiques et méthodes utilisées dans le milieu de l'intelligence artificielle et du machine learning appliqué à l'analyse statistique. Nous proposerons d'abord une définition de l'IA et ensuite du ML, nous étudierons les algorithmes les plus utilisés dans le milieu professionnel à l'aide des études de benchmark fournies par les entreprises, enfin nous parlerons des méthodes d'évaluation des modèles statistiques les plus courantes.

L'intelligence artificielle

L'Intelligence artificielle est définie comme étant « l'ensemble des techniques et théories mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence ». Elle correspond, entre autres, à un ensemble de concepts et de technologies simulant la cognition humaine plus qu'à une discipline autonome constituée.

Le machine learning

L'apprentissage automatique ou apprentissage statistique (machine learning en anglais), est un champ d'étude de l'intelligence artificielle qui concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus basé sur l'apprentissage sur un ou plusieurs types de données plutôt que par des algorithmes classiques.

Le machine learning peut être décliné en plusieurs types d'apprentissage :

1) L'apprentissage supervisé
2) L'apprentissage semi-supervisé
3) L'apprentissage renforcé

Chacun de ses types d'apprentissage ont évolué avec les besoins croissant en automatisation du traitement de données massives et des avancées technologiques (computer vision, natural language processing, emotion processing etc.)

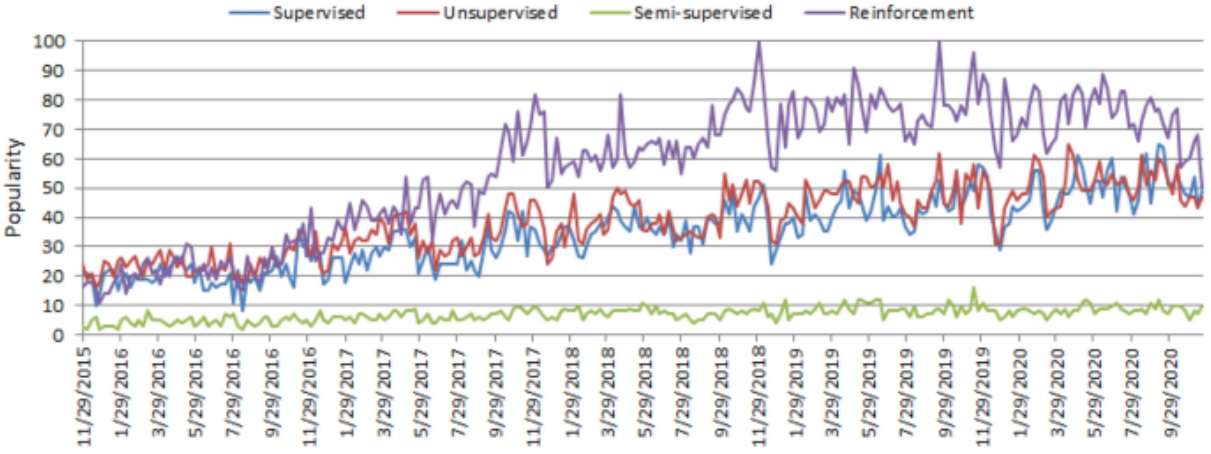


Figure 1 Evolution de l'utilisation des méthodes de ML au cours du temps

L'apprentissage supervisé consiste à développer une fonction qui associe des données d'entrée à des étiquettes (labels) cibles. Il est fourni avec un ensemble de données d'apprentissage labellisées par des méthodes de séparation des données (cross-validation) ainsi qu'un ensemble de données de test. Lorsque les variables cibles sont des valeurs réelles continues, les tâches d'apprentissage supervisé sont connues comme des problèmes de régression, et lorsque les variables cibles sont des variables catégoriques, les tâches sont connues comme des problèmes de classification.
Les algorithmes d'apprentissage supervisé courants comprennent :

1) La régression linéaire

2) La méthode Naïve Bayes (NB)
3) La régression logistique (RL)
4) L'arbre de décision
5) LDA & QDA
6) La forêt aléatoire (RF)
7) La machine à vecteurs de support (SVM)
8) Les K-voisins les plus proches (KNN)
9) Le réseau neuronal artificiel (ANN)

Les algorithmes et modèles de machine learning

L'apprentissage automatique implique l'utilisation d'algorithmes et de modèles de machine learning. Une erreur commune à la plupart des personnes débutant dans ce domaine est le fait de confondre la notion « algorithme d'apprentissage automatique » et celle de « modèle d'apprentissage automatique ». Ce point permet donc de lever la différence entre un algorithme et un modèle de machine learning avant de les étudier.

Qu'est-ce qu'un algorithme de machine learning ?

Dans le domaine de l'apprentissage automatique, un « algorithme » est une procédure qui est exécutée sur des données pour créer un « modèle d'apprentissage automatique». Les algorithmes simulant les fonctions principales de la cognition humaine comme la reconnaissance de formes (*"pattern recognition"*), ils « apprennent » à partir de données, ou sont « adaptés » à un ensemble de données.

En tant que tels, les algorithmes de machine learning ont un certain nombre de propriétés qui permet de les aborder de différentes manières :

- Les algorithmes de ML peuvent être décrits à l'aide de mathématiques et de pseudo-code.
- L'efficacité des algorithmes de ML peut être analysée et décrite.
- Les algorithmes de ML peuvent être mis en œuvre avec la plupart des langages de programmation modernes.

Qu'est-ce qu'un modèle de machine learning ?

Un « modèle » en ML est le résultat d'un algorithme d'apprentissage automatique exécuté des données.
Un modèle représente ce qui à été appris par un algorithme d'apprentissage automatique. Le modèle est la « chose » qui est sauvegardée après l'exécution d'un algorithme de ML sur des données d'apprentissage et représente les règles, les chiffres et toute autre structure de données spécifiques à l'algorithme nécessaire pour faire des prédictions.

Voici quelques exemples :

- L'algorithme de régression linéaire aboutit à un modèle d'un vecteur de coefficients avec des valeurs spécifiques;
- L'algorithme de l'arbre de décision est composé d'un arbre d'instructions « si-alors » avec des valeurs menant à une décision;
- Les algorithmes de réseau neuronaux donnent un modèle composé d'une structure graphique avec des vecteurs ou des matrices de poids menant à une décision.

La meilleure approche pour comprendre un modèle de ML consiste à considérer le modèle comme un « programme ». Ce programme comprend à la fois des données et une procédure d'utilisation des données pour les classifier.

Analyse des algorithmes de ML pour la classification

Cette section a pour but d'analyser les utilisations et le fonctionnement des algorithmes de machine learning les plus utilisés dans l'analyse statistique en milieu professionnel.

Les réseaux de neurones artificiels

Concept des neurones artificiels

Les réseaux neuronaux reflètent le comportement du cerveau humain, permettant aux programmes informatiques de reconnaître des modèles et de résoudre des problèmes courants dans les domaines de l'IA, de l'apprentissage automatique (ML) et de l'apprentissage profond (DL).

Les réseaux neuronaux artificiels (ANN) sont constitués de couches de nœuds, contenant une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque nœud, ou neurone artificiel, se connecte à un autre et possède un poids et un seuil associés. Si la sortie d'un nœud individuel est supérieure à la valeur seuil spécifiée, ce nœud est activé, envoyant des données à la couche suivante du réseau. Dans le cas contraire, aucune donnée n'est transmise à la couche suivante du réseau.

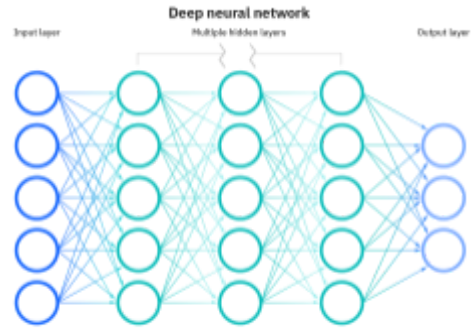


*Figure SEQ Figure * ARABIC 2: Illustration d'un réseau de neurones artificiel*

Types de réseaux de neurones

Les réseaux neuronaux peuvent être classés en différents types, qui sont utilisés à des fins différentes. Bien qu'il ne s'agisse pas d'une liste exhaustive, les types ci-dessous sont représentatifs des types les plus courants de réseaux neuronaux que vous rencontrerez dans les cas d'utilisation les plus courants :
Le perceptron est le plus ancien réseau neuronal, créé par Frank Rosenblatt en 1958. Il possède un seul neurone et constitue la forme la plus simple de réseau neuronal.

Les réseaux neuronaux à action directe, ou perceptrons multicouches (MLP), sont ceux sur lesquels nous nous sommes principalement concentrés. Ils se composent d'une couche d'entrée, d'une ou plusieurs couches cachées et d'une couche de sortie.

Les réseaux neuronaux convolutifs (CNN) sont similaires aux réseaux à anticipation, mais ils sont généralement utilisés pour la reconnaissance des images, la reconnaissance des formes et/ou la vision par ordinateur. Ces réseaux exploitent les principes de l'algèbre linéaire, en particulier la multiplication des matrices, pour identifier des motifs dans une image.

Les réseaux neuronaux récurrents (RNN) sont identifiés par leurs boucles de rétroaction. Ces algorithmes d'apprentissage sont principalement exploités lors de l'utilisation de données de séries temporelles pour faire des prédictions sur des résultats futurs, comme les prédictions boursières ou les prévisions de ventes.

Cas d'utilisation en entreprise

Depuis des décennies, IBM est un pionnier dans le développement des technologies d'IA et des réseaux neuronaux, mis en évidence par le développement et l'évolution d'IBM Watson. Watson est désormais une solution de confiance pour les entreprises qui cherchent à appliquer des techniques avancées de traitement du langage naturel et d'apprentissage profond à leurs systèmes en utilisant une approche par paliers éprouvée pour l'adoption et la mise en œuvre de l'IA.

Watson utilise le cadre Apache Unstructured Information Management Architecture (UIMA) et le logiciel DeepQA d'IBM pour mettre à la disposition des applications de puissantes capacités d'apprentissage profond. Grâce à des outils comme IBM Watson Studio, votre entreprise peut mettre en production des projets d'IA open source de manière transparente, tout en déployant et en exécutant des modèles sur n'importe quel cloud.

La méthode Naïve Bayes

Concept de la méthode

La méthode de Naïve Bayes est une méthode utilisant un classificateur Bayésien naïf. Comme son nom l'indique, ce classificateur repose sur le théorème suivant :

$$P(VC \mid VE) = \frac{P(VE \mid VC) * P(VC)}{P(VE)}$$

Avec VC étant la variable cible et VE étant les variables explicatives

Ce classificateur fait appel à une hypothèse naïve étant que chacune des variables explicatives sont supposées indépendantes. Bien que cette hypothèse soit rarement vérifiée, les estimations obtenues grâce au classificateur Bayésien naïf n'en reste pas moins très bonnes. Pour l'utiliser, il suffit de connaître l'estimation des probabilités conditionnelles et les probabilités à posteriori.

Avantage de la méthode

Le principal avantage de ce classificateur est sa vitesse d'apprentissage et ses prédictions. En effet, admettre que les variables explicatives sont toutes indépendantes engendre de forte réduction des calculs à réaliser. Pour des jeux de données de faible taille, ce classificateur se montre très efficace Cette qualité amène logiquement à l'utilisation de ce classificateur en combinaison avec d'autres algorithmes (Arbres de décision).
Une autre qualité de ce classificateur est que celui-ci est facilement incrémentale. Dans son cas, il peut être rapidement mis à jour sans nécessiter de refaire chaque calcul. Il suffit de mettre à jour les probabilités conditionnelles univariées des variables.

Classificateur Bayésien Naïf

Comme vu précédemment, l'idée de départ de ce classificateur est de calculer la probabilité conditionnelle :

$$P(VC \mid VE) = \frac{P(VE \mid VC) * P(VC)}{P(VE)}$$

Cependant, la probabilité conditionnelle P(VE | VC) n'est pas facilement estimable, il faudra faire alors appel à la version naïve de ce classificateur :

$$PNB(VC \mid VE) = P(VC) * \sum P(VE \mid VC) / P(VE)$$

Pour calculer ce classificateur, il suffit d'avoir en paramètre P(VC). Par contre, la probabilité P(VE | VC) est difficile à calculer car il faudra sauvegarder chaque instance.

Classificateur Bayésien Naïf Moyenné

L'amélioration du classificateur bayésien peut avoir lieu de deux principales méthodes. La première consiste à sélectionner des variables et la seconde consiste à pondérer les variables. La sélection des variables consiste à ne sélectionner que certaines variables. On utilisera alors le terme SNB *« Selective Naïve Bayes »*. Mais il serait trop simple de sélectionner aléatoirement certaines variables. La solution serait de supprimer celles qui ne sont pas informatives (dont la loi qui, a priori, ne serait pas informative). Pondérer chaque variable permet également d'améliorer les prédictions du classificateur. Cette approche nous amène donc à moyenner chaque variable.
Le moyennage consiste à combiner la prédiction de différents classificateurs de façon à améliorer les capacités prédictives.

Le classificateur Bayésien naïf moyenné procède de la même manière que le classificateur Bayésien naïf à la différence près qu'il ajoute une pondération par variable. Cette pondération a pour but de limiter le biais enclenché par l'hypothèse initiale du classificateur qui consiste à admettre que chaque variable explicative est indépendante aux autres.
On peut logiquement deviner que le classificateur Bayésien Naïf moyenné sera plus précis dans ses prédictions que le classificateur initial. Cette différence sera de plus en plus évidente lorsque le jeu de données sera grand.

Exemple d'utilisation

Il y a de nombreuses utilisations de la méthode de Naïve Bayes afin de réaliser des prédictions sur un échantillon donné. On peut notamment l'utiliser dans plusieurs domaines comme les finances, la médecine, le sport, les prédictions météorologiques, …
Dans le cas d'une étude réalisée par un étudiant en Master 1 Mathématiques et Applications Spécialité Statistique de l'université de Strasbourg, L'algorithme Naïve Bayes est utilisé pour prédire l'apparition de séisme en fonction de plusieurs paramètres comme la localisation, la magnitude etc. Cette étude fait également appel aux algorithmes SVM et k plus proches voisins.

Les Forêts d'arbres décisionnel

Concept de l'arbre de décision

Un arbre de décision est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Il s'agit d'un algorithme classique d'apprentissage supervisé. L'objectif d'un arbre de décision est de construire pas à pas des « segments » de population les plus « pures » possible. Cet algorithme peut être utilisé pour des problèmes de régression et de classification grâce à des variables explicatives qui peuvent être quantitatives et/ou qualitatives.

Choix de la variable de segmentation

Afin de décider comment segmenter la population initiale, on teste toutes les variables et on choisit la variable X qui présente la plus forte liaison avec Y. Afin de quantifier cette liaison on utilise la quantité du χ^2 calculée sur le tableau de contingence (croisement de Y avec X_i). Les prochaines divisions tiennent compte de la nouvelle population afin de renouveler la segmentation. Cette segmentation en cascade forme un arbre de décision avec des segments de plus en plus pure.
Dans le cas d'une variable quantitative, toutes les valeurs seront testées et la segmentation sera effectuée pour la valeur séparant au mieux la population identifiée précédemment.

Arrêt de la segmentation

L'objectif de l'arrêt de la segmentation est de conserver une capacité de généralisation du modèle. En effet, si l'on poursuit la segmentation même sur de très faibles effectifs, on risque que notre modèle soit sur-ajusté par rapport aux données. Il est possible de définir des règles d'élagage sur les effectifs, sur la significativité des segmentations ou sur l'homogénéité des segments.

Règles de décision

Un segment terminal est affecté à la classe à (k de la variable Y) la plus représentée. Mais cette règle est bonne si la variable à expliquer Y présente des modalités équilibrées en proportion. Dans le cas contraire, si les modalités sont déséquilibrées, il est plus judicieux d'affecter le segment terminal à la modalité sur-représentée par rapport à la distribution d'origine

Exemple d'utilisation

Cet exemple illustre le cas où l'on cherche à prédire si des sportifs vont disputer ou non un match en fonction de données météorologiques.

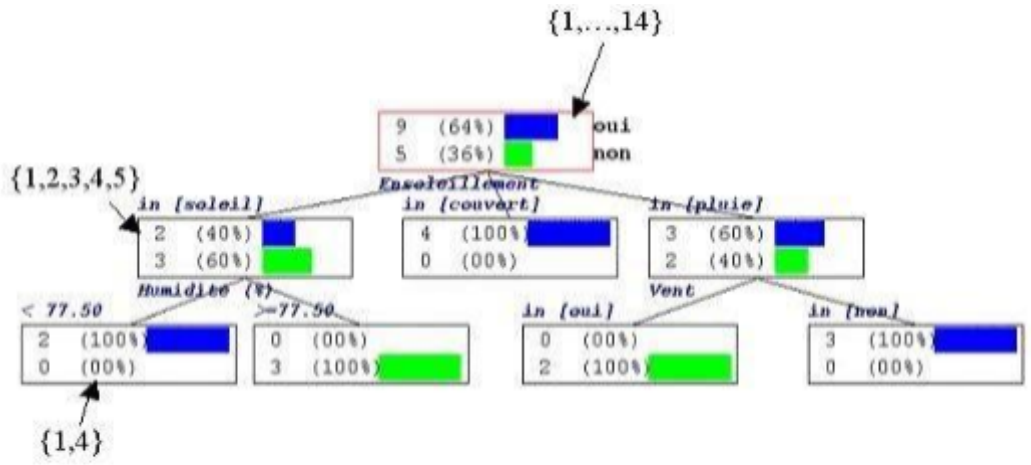


*Figure SEQ Figure * ARABIC 3 : Illustration d'un arbre de décision*

Avantages et inconvénients

L'arbre de décision possède de nombreux points fort :

- Il ne nécessite pas d'hypothèse sur les données ;
- Les variables explicatives peuvent être de nature qualitative et quantitative ;
- Il est idéal pour trouver les seuils de coupure optimaux pour les variables continues explicatives et il est robuste aux données aberrantes;
- De plus ce modèle est de type « boîte blanche » en effet il est simple d'expliquer les sorties du modèle, notamment grâce au côté visuel de la méthode.

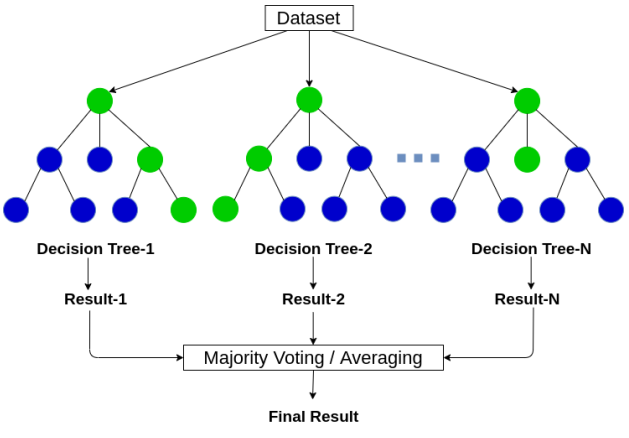
Également il est facile à implémenter grâce à de simple *« if » « else if »*. A l'inverse, ce modèle a certains défauts :

- Il est nécessaire d'avoir un effectif important car sur un petit échantillon le modèle peut s'avérer instable;
- Si une variable à plus de 2 groupes la classification peut être difficile;
- Il faut aussi faire attention à ce qu'une variable explicative n'en cache pas une autre.

Random Forest

La Random Forest est une méthode de Machine Learning qui repose sur l'utilisation des arbres de décisions. La méthode consiste à créer un grand nombre d'échantillons d'apprentissage N grâce à une méthode de tirage avec remise (*"Bootstrap"*). Ensuite, sur ces N échantillons on va construire N arbres de décisions. Chaque arbre affecte une réponse et par un système de « vote », la réponse est définie grâce à la majorité des arbres. Il est également possible de pondérer le vote d'un arbre en fonction des performances de ces prédictions individuelles.

Le principal défaut de cette méthode est qu'il est de type « boîte noire ». En effet, il n'est plus possible d'expliquer aussi facilement les résultats de ce modèle qu'avec un arbre de décision seule. De plus, entraîner un modèle de random forest est bien plus exigeant en termes de puissance de calcul.



LDA & QDA

Analyse Discriminante linéaire (LDA)

Concept

L'analyse discriminante linéaire (*"Linear discriminant analysis"*) est une généralisation du discriminant de Fisher. Le but de cette analyse est de trouver une ou plusieurs combinaisons linéaires à partir des variables explicatives séparant deux classes distinctes. Pour cela, il faut réussir à trouver la classe qui maximise la fonction discriminante linéaire. Les combinaisons linéaires pourront alors être utilisées comme classificateur linéaire.

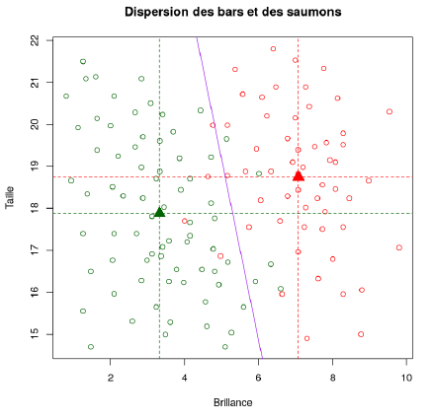
L'analyse discriminante linéaire se base sur l'hypothèse suivante qui suppose que les variables indépendantes sont normalement distribuées. Dans les cas où cette hypothèse n'est pas cohérente, il faudra alors se tourner vers d'autres méthodes telle que la régression logistique qui reste une méthode assez similaire à LDA.

Principe de l'analyse discriminante linéaire

L'analyse discriminante linéaire suppose que toutes les classes sont linéairement séparables par des hyperplans que l'on pourra créer à partir des variables explicatives. De ce fait, on peut représenter cette méthode géométriquement.

Pour cela, il suffit d'estimer le point de coordonnées moyennes correspondant aux deux classes que l'on cherche à séparer avec un nouvel hyperplan. Une fois que ces deux points seront trouvés, il sera alors possible de tracer cet hyperplan de manière à maximiser la séparation des populations des deux classes. Pour mener à bien cette séparation, il est nécessaire de respecter deux règles :

- ➔ Maximiser la distance entre les moyennes des deux classes
- ➔ Minimiser la variation entre chaque classe



Ci-dessus, on observe deux classes (bars en vert, saumons en rouge) réparties sur un plan en deux dimensions (Brillance,Taille) correspondant aux variables explicatives. Chaque coordonnées moyennes de ces deux classes sont représentées par un triangle de la couleur de la classe. A partir de ces deux triangles, il a donc été possible de tracer l'hyperplan qui séparera linéairement les deux classes. De cette manière, on estime que chaque individu dont les coordonnées se situent à la gauche de l'hyperplan seront de la classe bars et chaque individu dont les coordonnées se situent à la gauche du plan seront de la classe saumons. Pour n classes, il y aura logiquement n-1 hyperplan à tracer afin de distinguer chaque classe.

On remarque qu'il est nécessaire d'avoir un éparpillement inter-classes élevé, c'est-à-dire que les différentes classes soient suffisamment dispersées sur le plan. De même, l'éparpillement intra-classe doit être faible pour éviter une mauvaise classification (comme l'objet de classe B qui se retrouve à la gauche de l'hyperplan alors qu'il devrait être à la droite de celui-ci).

Avantages et inconvénients de la méthode LDA

Tout d'abord, parmis les avantages, il y a :

- la robustesse par rapport à l'hypothèse de normalité;
- la maximisation de l'éparpillement inter-classes et la minimisation de l'éparpillement intra-classe.

Ensuite, on retrouve les différents inconvénients tels que :

- le coût en temps de calcul;
- le coût en espace mémoire;
- la précision des résultats en cas d'une quantité de données élevé;
- l'incapacité de la méthode face à une distribution multimodale.

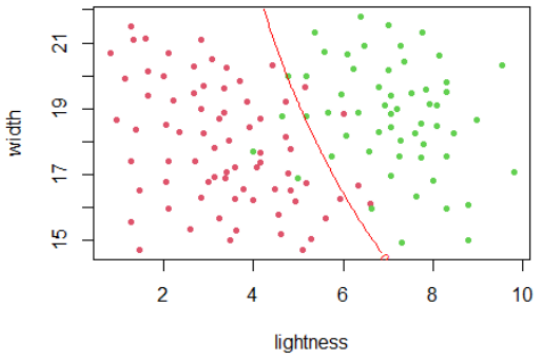
Analyse Discriminante Quadratique (QDA)

Concept

Cette méthode s'utilise dans le cas où l'hypothèse d'hétéroscédasticité est vérifiée, c'est-à-dire que toutes les classes ont des matrices de covariances distinctes. Tout comme l'analyse discriminante linéaire, cette méthode se base sur l'hypothèse qui admet que chaque variable est indépendante et normalement distribuée.

L'analyse discriminante quadratique consiste à tracer une courbe séparant deux classes comme pour l'analyse discriminante linéaire. Pour tracer cette courbe, il faut en premier lieu estimer la matrice de covariance pour chaque classe distincte. Ainsi, il sera possible de trouver la classe qui maximise la fonction discriminante quadratique. Celle-ci est assez similaire à la fonction discriminante linéaire à la différence près que les matrices de covariance sont différentes.

Une courbe sera alors traçable sur le plan correspondant à la limite de décision :



On observe les deux classes: saumons et bars (bars en rouge et saumons en vert). Sur le plan, on peut voir une courbe en rouge estimée grâce à la méthode QDA. Ainsi, pour toute nouvelle individus dont on ne connaît pas la classe, on estimera que ce sera un bar s'il se trouve à la gauche de la courbe ou un saumon s'il se trouve à la droite de la courbe.

Avantages et inconvénients de la méthode QDA

Les avantages de cette méthode sont globalement assez similaires avec ceux de la méthode LDA. On peut quand même remarquer ceci :

- La méthode QDA permet une plus grande flexibilité par rapport aux matrices de covariance des données puisque la méthode suit l'hypothèse d'hétéroscédasticité.
- Néanmoins cela implique un nombre de paramètres bien plus important car chaque matrice de covariance est différente.

Régression linéaire

Concept

La régression linéaire est une méthode statistique consistant à représenter linéairement une variable à expliquer en fonction de variables explicatives. Cette méthode s'exprimera par une courbe suivant la forme suivante:

$$Y = a_0 + a_1*X_1 + a_2*X_2 + \ldots + a_n*X_n + E$$

avec Y étant la variable à expliquer, $\{X_1, \ldots, X_n\}$ étant les variables explicatives et E correspondant aux résidus.

De cette manière, il est possible de réaliser des prédictions sur la variable Y lorsque l'on a les valeurs des variables X. Cette modélisation pourra alors être testée. Pour cela, il suffit de regarder la valeur du R^2. Cette valeur appartient à l'intervalle {0;1}. Plus celle-ci est proche de 1, plus le modèle est de bonne qualité.

Conditions d'application

Afin de faire appel à cette méthode, il faut respecter certaines conditions d'application. Tout d'abord, chacune des variables utilisées dans la régression doit être quantitative. Sans cela, le modèle n'est pas réalisable et la variable à expliquer ne pourra pas être représentée par une équation numérique. Il faut également que les résidus suivent une distribution normale et avoir un nombre suffisant d'observations. Enfin, il faut faire attention à ce qu'aucune des variables explicatives ne puissent s'écrire comme la combinaison linéaire d'autres variables explicatives. On appelle ça la non-colinéarité des variables explicatives.

Avantages et inconvénients

Les avantages de cette méthode sont les suivants:

- Cette méthode est relativement simple à mettre en place et à interpréter.
- Elle est adaptée aux variables quantitatives.
- Elle peut être sujet au sur-ajustement mais est compatible avec des techniques comme la validation croisée qui peuvent résoudre ce genre de problèmes.

Les inconvénients sont les suivants:

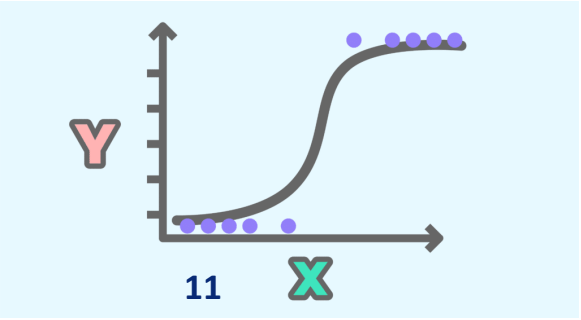
- Les valeurs aberrantes (outliers) ont un grand impact sur la modélisation et peuvent fausser celle-ci.
- Cette méthode nécessite le respect de nombreuses hypothèses ce qui réduit grandement les cas où elle peut être utilisée.

Régression logistique

Concept

La régression logistique est une méthode de machine learning qui vise à faire une des prédictions sur une variable à expliquer en fonction de variable explicative. Dans le cas de la régression logistique, la variable à expliquer sera forcément une variable qualitative binaire. Dans le cas d'une régression logistique multinomiale, la variable à expliquer sera une nouvelle fois qualitative mais ne sera pas binaire. Dans la suite de la présentation de la méthode, nous allons nous intéresser à la régression logistique pour une variable à expliquer binaire.

La régression logistique peut être représentée graphiquement de la manière suivante:



Ici, Y correspond à la variable à expliquer, X correspond à la variable explicative, les points bleus correspondent aux individus du jeu de données et la courbe correspond au modèle obtenu grâce à la régression logistique.

Exemple d'utilisation

Un domaine utilisant beaucoup cette méthode de machine learning est le domaine médical. En effet, il est courant de retrouver des problématiques dont le but est de prédire une variable qualitative binaire. Par exemple: "Peut-on prédire la présence d'une maladie cardiaque chez un individu à partir de sa pression sanguine, de son taux de cholestérol… ?".

Condition d'utilisation

La condition principale d'utilisation est d'avoir une variable qualitative binaire à prédire (sauf dans le cas de la régression logistique multinomiale). Toutes les variables utilisées doivent être représentées dans le modèle. C'est-à-dire qu'il faut faire une sélection des variables qui seront vraiment utiles pour la modélisation. Il faut également éviter toute corrélation élevée entre deux variables explicatives. Si jamais deux variables explicatives partagent une corrélation trop forte, il faut alors en choisir une et enlever l'autre du modèle. En général, ce choix se fera en fonction des corrélations que partagent ces deux variables avec les autres variables explicatives. Enfin, il faut faire attention à la taille de l'échantillon étudié.

.

Avantages et inconvénients

Les avantages de cette méthode sont les suivants:

- Tout d'abord, la régression logistique reste une méthode de machine learning très facile et rapide à mettre en place, très efficace et facilement interprétable.
- Ce modèle est également efficace pour faire de la classification.
- Cette méthode peut-être sujet au surajustement mais est compatible avec des techniques comme la validation croisée qui peuvent résoudre ce genre de problèmes.

Les inconvénients de cette méthode sont les suivants:

- Cette méthode construit une frontière linéaire autour de variables qualitatives.
- Ce modèle nécessite une non-corrélation entre les différentes variables explicatives ce qui peut être problématique dans certains cas d'étude.

Support Vector Machine

Introduction

Les supports vector machine font partie des algorithmes qui sont dit algorithmes d'apprentissage supervisée. Les svm permettent de résoudre des problèmes de classification, de régression et de détection d'anomalie. Ils sont notamment utilisés pour la reconnaissance d'images.

Objectif des SVM

L'objectif de cet algorithme est de trouver la droite (2-dimensions) ou le plan (3-dimensions ou plus) qui permet de classifier des données. Cette frontière de séparation est appelée hyperplan.

Cas séparable

Pour séparer deux classes qui peuvent l'être linéairement, un hyperplan est tracé tel qu'il maximise la distance entre les deux groupes. Cette maximisation dépend de la marge maximale. Cette marge est la distance entre l'hyperplan et les points les plus proches. Ces points sont les **vecteurs support**. Ils influencent la position et l'orientation de cette frontière. Ces points sont primordiaux pour la conception des **SVM**. De plus, l'hyperplan doit être à équidistance des deux classes.

L'hyperplan maximisant la marge est donné par :

$$arg \max_{w,w_0} \min_{k} \{||x - x_k|| : x \in \mathbb{R}_N, w^T x + w_0 = 0\}$$

Cependant, la recherche d'un hyperplan utilisant la marge maximale permet seulement de résoudre les cas de séparation linéaire.

Cas non séparable

Lorsque les données ne peuvent pas être séparées linéairement, une transformation doit être appliquée aux données.
L'objectif est d'augmenter la dimension des données afin de pouvoir les séparer linéairement et effectuer des prédictions.
Nous pouvons alors introduire alors l'**astuce du noyau** ("*kernel trick*").

Beaucoup de ces transformations de données amènent à un coût de calcul très élevé. Cependant, l'astuce du noyau permet de réduire ce coût. Cette méthode permet d'obtenir les similarités entre chaque donnée originale. Ce nouvel ensemble est alors représenté par une matrice à noyau de dimension nxn. La similarité i,j est de m-dimension(s) avec m étant le nombre de variables utilisées.
L'augmentation de dimension citée plus tôt et qui permet le calcul des similarités entre chaque observation revient tout simplement à créer une nouvelle variable égale à x^3. Par exemple, si l'on est dans une situation avec une dimension = 2 avec longueur des pétales et longueur des sépales, une $3^{ème}$ dimension est créée et est donc égale à (longueur des pétales)^3. Cette augmentation de dimensions permet alors de séparer linéairement les données d'entraînement pour ensuite effectuer des prédictions.

Avantages et inconvénients

+ Les svm sont très faciles d'implémentation, notamment avec python;
+ très grande précision de prédiction;
+ fonctionnent très bien sur de petits jeux de données;
+ efficaces car utilisent un sous-ensemble de points d'entraînement.

- temps d'entraînement très long sur des datasets très volumineux;
- pas très efficace avec la présence de bruit et de valeurs aberrantes.

K plus proches voisins

Concept

L'algorithme des k plus proches voisins est une méthode d'apprentissage supervisée. Il peut être utilisé pour des problèmes de régression comme pour des problèmes de classification.

L'idée derrière cette algorithme se traduit par l'analogie "Dis moi qui sont tes voisins et je te dirais qui tu es." Ainsi, l'algorithme fonctionne de la manière suivante. Tout d'abord, on renseigne le nombre de voisins que l'on veut étudier (k) et la formule de la distance que l'on va utiliser. Ensuite, on peut passer à la réalisation de l'algorithme.

La première étape est de mesurer chacune des distances entre l'individu étudié et les autres individus du jeu de données. La seconde étape sera de retenir les k valeurs des distances les plus petites. La troisième étape dépend du type de problème que l'on cherche à résoudre. S'il s'agit d'une régression, il faudra calculer la moyenne des distances retenues. S'il s'agit d'une classification, il faudra calculer le mode des distances retenues. Ainsi, on obtient des prédictions sur l'individu étudié.

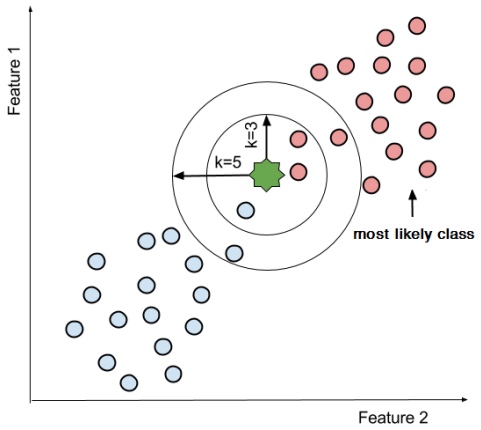
Pour le choix des distances à utiliser pour cet algorithme, celui-ci est assez libre. En effet, il existe une trentaine de distances utilisables dans ce cas. Il faut donc faire son choix en fonction de la nature du jeu de données. Par exemple, il sera plus logique d'utiliser la distance euclidienne dans le cas de variables quantitatives.

Choisir le nombre K de voisins

Le choix de k est indispensable pour le bon fonctionnement de l'algorithme. Seulement, il se pose la problématique de savoir pour quel nombre k l'algorithme sera optimal. Choisir un nombre k faible permet rarement d'avoir un modèle de bonne qualité. Néanmoins, choisir un nombre k trop élevé pourrait amener le modèle à être en sur-apprentissage. Il faut donc réussir à trouver un entre-deux satisfaisant.

Exemple d'utilisation

Cet algorithme est utilisé dans différents domaines. Il apparaît dans les recherches autour de détection d'images, de vidéos ou de textes et aussi dans le domaine économique (accorder un crédit, …).



Avantages et inconvénients

Les avantages de cet algorithme sont les suivants:

- L'algorithme est facile à mettre en place et n'est pas sujet à de nombreuses contraintes.
- L'algorithme peut régler des problèmes de classification comme de régression.
- L'algorithme ne nécessite aucune hypothèse sur les données.

Les inconvénients de cet algorithme sont les suivants:

- L'algorithme requiert énormément de place puisqu'il a besoin de garder en mémoire chaque individu du jeu de données. Plus les données sont grandes, plus l'algorithme consomme de la mémoire.
- Choisir la bonne distance et le bon nombre de k peut être compliqué.