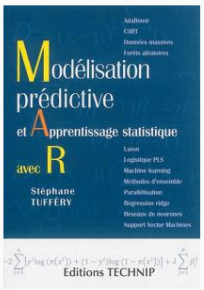
Les mots des auteurs

- «Lumière» par Annie Ernaux
- «Camarade» par Alain Bonnand
- «Elégance» par Raphaël-Didier de L'Hommel
- «Créée» par Florence Gentner
- «Étymologie» par Malo Aguettant
- (voir les autres)

Modélisation prédictive et apprentissage statistique avec R



- Référence 9782710811589
- Prix de vente éditeur 45.00 €
- Edité par Technip, Paris, France
- Ecrit par Stéphane Tufféry

Présentation de l'éditeur

Issu de formations devant des publics variés, cet ouvrage présente les principales méthodes de modélisation de statistique et de machine learning, à travers le fil conducteur d'une étude de cas. Chaque méthode fait l'objet d'un rappel de cours et est accompagnée de références bibliographiques, puis est mise en oeuvre avec des explications détaillées sur les calculs effectués, les interprétations des résultats et jusqu'aux astuces de programmation permettant d'optimiser les temps de calcul. À ce sujet, une annexe est consacrée au traitement des données massives.

L'ouvrage commence par les méthodes de classement classiques les plus éprouvées, mais aborde rapidement les méthodes plus récentes et avancées : régression ridge, lasso, elastic net, boosting, forêts aléatoires, Extra-Trees, réseaux de neurones, séparateurs à vaste marge. Chaque fois, le lien est fait entre la théorie et les résultats obtenus pour montrer qu'ils illustrent bien les principes sous-jacents à ces méthodes. Mais l'aspect pratique est aussi privilégié, avec l'objectif de permettre au lecteur une mise en oeuvre rapide et efficace dans son travail concret. L'exploration et la préparation préliminaire des données sont d'ailleurs décrites, ainsi que le processus de sélection des variables. Une synthèse finale est faite de toutes les méthodes présentées.

La mise en oeuvre s'appuie sur le logiciel libre R et sur un jeu public de données. Ce dernier peut être téléchargé sur internet et présente l'intérêt d'être riche, complet et de permettre des comparaisons grâce aux nombreuses publications dans lesquelles il a servi. Le logiciel statistique utilisé est R, actuellement celui qui se développe le plus : devenu la lingua franca de la statistique et l'outil le plus répandu dans le monde académique, il prend également de plus en plus de place dans le monde de l'entreprise, à tel point que tous les logiciels commerciaux proposent désormais une interface avec R. Outre qu'il est disponible pour tous, dans de multiples environnements, il est aussi le plus riche statistiquement et c'est le seul logiciel permettant de mettre en oeuvre toutes les méthodes présentées dans cet ouvrage. Enfin, son langage de programmation particulièrement élégant et adapté au calcul mathématique permet de se concentrer dans le codage sur les aspects statistiques. R permet d'arriver directement à l'essentiel et de mieux comprendre les méthodes exposées dans l'ouvrage.

Stéphane TUFFÉRY est docteur en mathématiques. En charge de la statistique et du data mining dans un grand groupe bancaire français, il enseigne le data mining à l'université Rennes 1 à l'ISUP (Institut de Statistique de l'Université de Paris). Il a également publié "Étude de cas en statistique décisionnelle" dans la même collection.

Extrait du livre

Extrait de l'avant-propos

Cet ouvrage explique et met en oeuvre les principales méthodes de modélisation d'une situation dans laquelle on recherche la prédiction d'une variable binaire à l'aide de variables qualitatives ou quantitatives. Cette situation est très fréquente en data mining et en statistique, et elle est à la base entre autres des algorithmes de scoring. Nous le ferons à l'aide du logiciel libre R sur la base d'une étude de cas. Nous avons choisi ce logiciel, parce qu'il est aujourd'hui le plus répandu chez les statisticiens et les data scientists, et aussi en raison de la richesse et de la variété sans égales de ses fonctionnalités, couvertes par des modules appelés «packages» (ou «paquets»), au nombre de plus de six mille au moment où nous écrivons ces lignes. De plus, les ressources consacrées à R sur Internet et dans la littérature statistique sont innombrables et permettent de trouver la réponse à presque n'importe quelle question. Pour aider le lecteur, nous avons indiqué quelques références dans la bibliographie en fin d'ouvrage.
Les méthodes présentées ici couvrent le champ de la statistique, mais aussi de ce qui dans la terminologie anglo-saxonne s'appelle «statistical learning» et «machine learning». Chaque chapitre de l'ouvrage est consacré à une méthode, et comporte des rappels théoriques, suivis de larges développements mettant en oeuvre R sur un jeu de données que nous allons décrire. Les paramètres et les sorties des fonctions de R sont commentés, et les résultats obtenus pour chaque méthode sont confrontés d'une part avec la théorie et d'autre part avec les résultats obtenus avec les autres méthodes. Le lecteur sera ainsi à même de choisir les méthodes les plus adaptées à ses problématiques, et de les programmer efficacement. Les textes de référence sont indiqués s'il veut approfondir la théorie.
Les bases de la statistique sont supposées connues du lecteur, ainsi que les concepts d'estimateur, d'échantillonnage bootstrap, de courbe ROC et d'aire sous la courbe ROC, qui sont juste rappelés en temps utile. Si nécessaire, le lecteur pourra se référer aux indications bibliographiques fournies ou à notre ouvrage Dafa Mining et statistique décisionnelle, paru aux Éditions Technip.

Les étapes de notre étude de cas sont les suivantes.
Nous commençons par importer, échantillonner et décrire les données. Nous mesurons la liaison entre la variable à expliquer et chacune des variables explicatives. Une phase de mise en forme des données aboutit à des regroupements de modalités pour les variables qualitatives, et une discrétisation des variables explicatives quantitatives. Une deuxième méthode de discrétisation supervisée est proposée, avec une procédure basée sur la maximisation de l'aire sous la courbe ROC de la prédiction à l'aide de la variable discrétisée.
Nous mettons en oeuvre le modèle classique de régression logistique logit, avec recours à la sélection pas à pas, puis à la sélection globale (algorithme de Furnival et Wilson et méthode directe). Nous en déduisons une grille de score et nous établissons les règles de décision basées sur cette grille de score. Nous testons aussi les modèles probit et log-log.
Nous voyons ensuite comment les méthodes de pénalisation s'appliquent à la régression logistique : régression ridge, lasso, group lasso, et l'elastic net qui est un compromis entre ridge et lasso. Nous évoquons la non-consistance de l'estimateur lasso et les solutions que sont le relaxed lasso et l'adaptive lasso.
Nous testons la régression logistique PLS et faisons quelques rappels sur les intervalles de confiance qu'il est possible d'obtenir par bootstrap dans ce genre de situation.
Nous rappelons ensuite les principes de l'arbre CART, que nous mettons en oeuvre, en insistant particulièrement sur son mécanisme d'élagage. Après cela, nous présentons une méthode de Friedman et Fisher dérivée des arbres de décision mais moins connue : le «bump hunting» (ou algorithme PRIM).
Suivent les méthodes d'agrégation, dites aussi «méthodes d'ensemble» : le bagging, les forêts aléatoires, les Extra-Trees et le boosting (Discrète AdaBoost et Real AdaBoost). Une variante des forêts aléatoires est présentée, consistant à appliquer le mécanisme de double randomisation (individus et variables) des forêts aléatoires au modèle logistique et non plus à des arbres. Nous appliquons aussi l'algorithme du boosting au modèle logistique.
Ce panorama des méthodes d'apprentissage supervisé se poursuit avec les Support Vector Machines (séparateurs à vaste marge), et plusieurs noyaux possibles : linéaire, radial, sigmoïde et polynomial. Nous montrons le bénéfice que nous pouvons attendre de l'utilisation d'un SVM sur les coordonnées factorielles issues d'une analyse des correspondances multiples.
Le panorama s'achève avec les réseaux de neurones, et le boosting et les forêts aléatoires de réseaux de neurones, plus spécifiquement de perceptrons à une couche cachée.
Le dernier chapitre présente une synthèse des méthodes précédentes, avec la comparaison de leur pouvoir prédictif et autres propriétés.

Les mots préférés de Stéphane Tufféry

- «Tintinnabuler»

  J'éprouve quelque difficulté à choisir un mot entre les si nombreux magnifiques mots de la langue française, mais je choisirais «tintinnabuler». J'aime sa rareté, la légèreté de sa sonorité et la douceur de ce qu'il évoque. Il peut suggérer des situations très différentes, le tintinnabulement d'un lustre de cristal, d'une clochette, d'un grelot au cou d'un animal, et même d'une source cachée dans la nature.

Recherchez

trouver

Participez !

Pour participer à la vie du site, c'est simple:

- Editeurs
- Auteurs

Contactez-nous !

N'hésitez pas à nous contacter pour toute suggestion ou question concernant ce site.

Nos partenaires

librairie dialogues.fr

leslibraires.fr