Classification

INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE

La diversité des problèmes qui peuvent être traités par les algorithmes de classification est importante et couvre de nombreux domaines. Il est difficile d'aborder de manière exhaustive toutes les méthodes dans un seul ouvrage.

[C.C. Aggarwal]

4. Aperçu de la classification

Vue d'ensemble

Dans la **classification**, un échantillon de données (l'ensemble de **formation**) est utilisé pour déterminer les règles et les modèles qui divisent les données en groupes prédéterminés, ou **classes** (apprentissage supervisé).

Les données de formation sont généralement constituées d'un sous-ensemble de données **étiquetées** (cibles) sélectionnées de **manière aléatoire.**

L'estimation de la valeur (régression) est semblable à la classification (la variable cible est **numérique**).

Vue d'ensemble

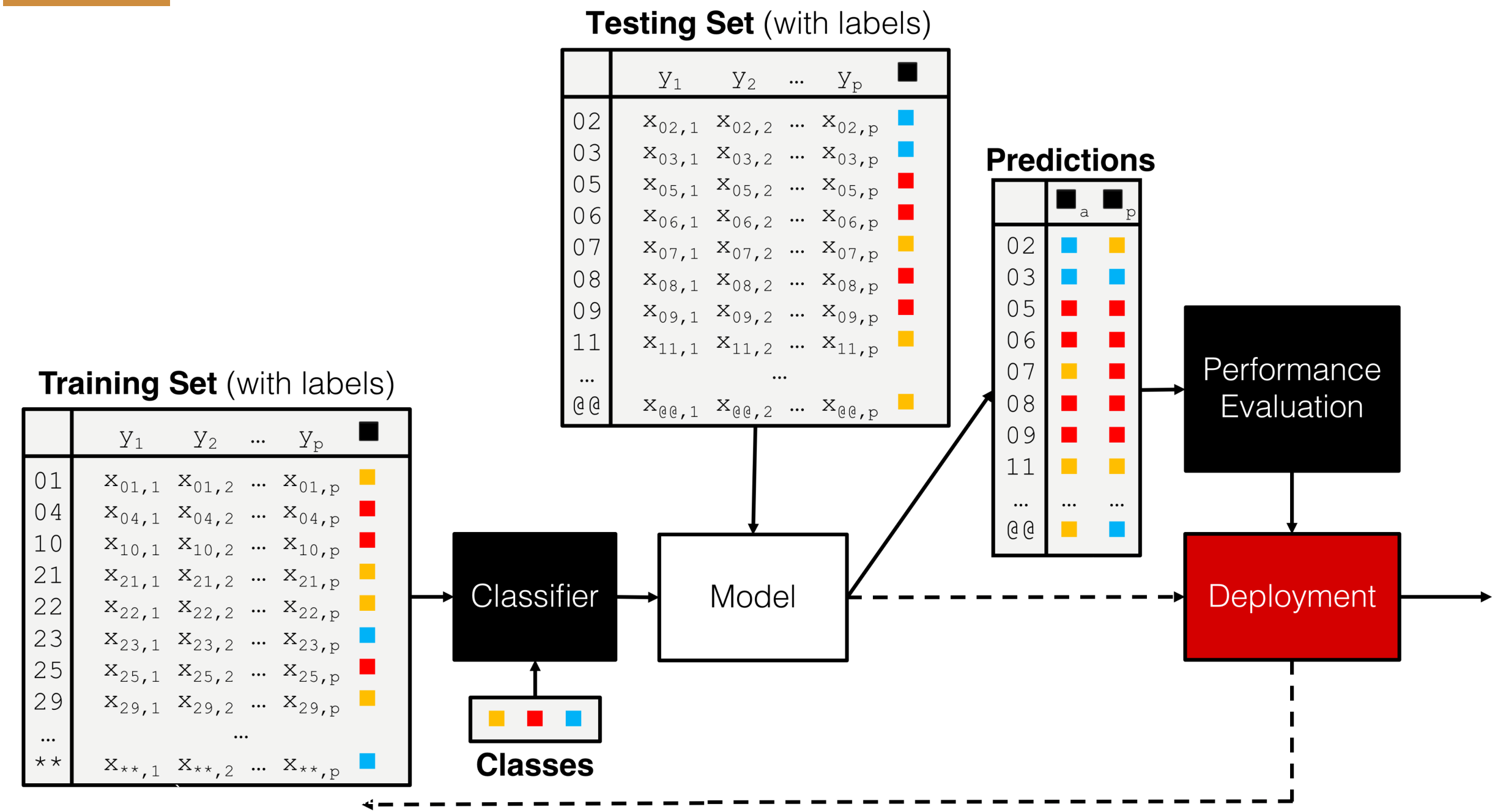
Ensuite, le modèle est utilisé pour attribuer une classe aux observations pour lesquelles l'étiquette est cachée, mais connue (l'ensemble de **test**).

La performance d'un modèle de classification est évaluée sur l'ensemble de test, **jamais** sur l'ensemble de formation. En l'**absence de** données de test, la classification peut être **descriptive**, mais elle n'est pas prédictive.

Problèmes techniques :

- sélection des caractéristiques à inclure dans le modèle
- sélection de l'algorithme
- etc.

Session 2


Testing Set (with labels)
Predictions
Training Set (with labels)
Performance Evaluation
Classifier
Model
Deployment
Classes

Applications

Médecine et sciences de la santé

- prédire quel patient risque de subir un deuxième infarctus mortel dans les 30 jours en fonction de facteurs de santé (tension artérielle, âge, problèmes de sinus, etc.)

Politiques sociales

- prédire la probabilité qu'une personne âgée ait besoin d'un logement d'assistance sur la base d'informations démographiques/de réponses à des enquêtes

Marketing et affaires

- prédire quels clients sont susceptibles de changer d'opérateur de téléphonie mobile sur la base de données démographiques et de l'utilisation du téléphone

Applications

Prédire qu'un objet appartient à une classe particulière.

Organiser et regrouper les instances en catégories.

Prédire le taux d'inflation pour les deux années à venir sur la base d'un certain nombre d'indicateurs économiques.

Améliorer la détection des objets pertinents

- **évitement :** "cet objet est un véhicule qui arrive"
- **poursuivre :** "il est peu probable que cet emprunteur ne rembourse pas son prêt hypothécaire"
- **degré :** "ce chien a 90% de chances de vivre jusqu'à l'âge de 7 ans".

Exemples

Scénario :

Une compagnie d'assurance automobile dispose d'un service d'enquête sur les fraudes qui étudie jusqu'à 30 % de toutes les demandes d'indemnisation.

Questions : peut-on prédire

- si une créance est susceptible d'être frauduleuse ?
- si un client est susceptible de commettre une fraude dans un avenir proche ?
- si une demande de police est susceptible de donner lieu à une réclamation frauduleuse ?
- le montant de la réduction d'une créance si elle est frauduleuse ?

Exemples

Scénario :

Les clients qui passent un grand nombre d'appels au service clientèle d'un opérateur de téléphonie mobile ont été identifiés comme présentant un risque de désabonnement. L'entreprise souhaite réduire ce taux de désabonnement.

Questions : peut-on prédire

- la valeur monétaire à vie apportée par un client ?
- quels sont les clients les plus susceptibles de se désabonner dans un avenir proche ?
- quelle est l'offre de fidélisation à laquelle un client donné répondra le mieux ?

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Objectif

L'*Internal Revenue Service* (IRS) des États-Unis a estimé qu'il y avait un écart important entre les **recettes dues** et les **recettes perçues** pour 2001 et pour 2006.

En utilisant les données du DoR (*Minnesota Department of Revenue*), les auteurs ont cherché à accroître l'**efficacité** du processus de sélection des audits et à **réduire l'écart** entre les recettes dues et les recettes perçues.

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Méthodologie

1. **sélection et séparation des données :** les experts ont sélectionné des cas à auditer et les ont divisés en ensembles de formation, de test et de validation
2. **modélisation de la classification** à l'aide de MultiBoosting, Naïve Bayes, arbres de décision C4.5, perceptrons multicouches, etc.
3. **évaluation de tous les modèles** sur l'ensemble de tests – les modèles ont donné de mauvais résultats jusqu'à ce que l'on reconnaisse l'effet de la taille de l'entreprise auditée, ce qui a donné lieu à deux tâches distinctes (grandes/petites entreprises).
4. **sélection/validation de modèles** pour comparer la précision estimée entre les prédictions de différents modèles de classification et les audits réels sur le terrain (MultiBoosting avec Naïve Bayes a été sélectionné comme modèle final, suggérant des améliorations pour accroître l'efficacité de l'audit).

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Données

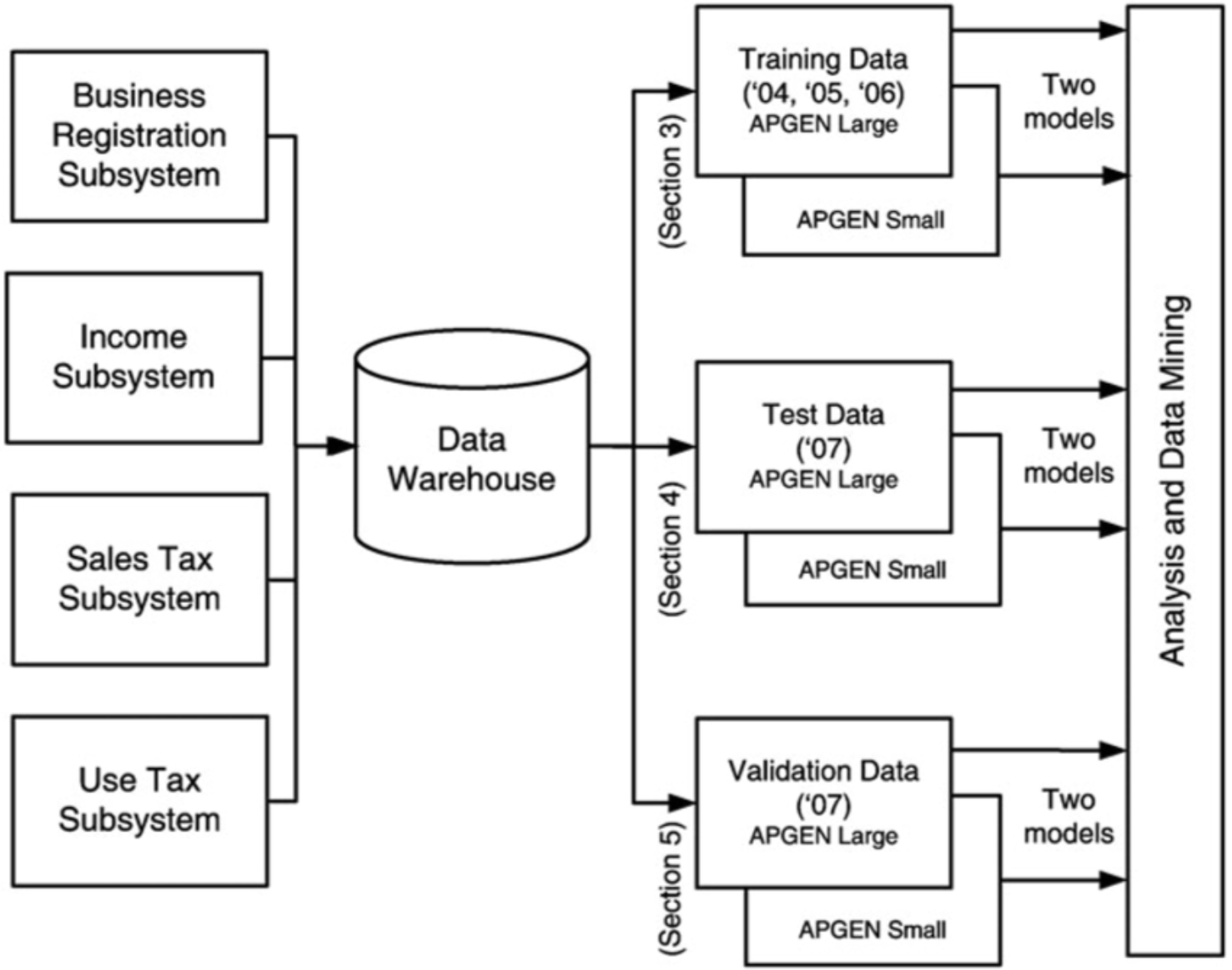
Une sélection de cas d'audit fiscal de 2004 à 2007, recueillis par les experts en audit (divisés en formation, test, validation) :

- l'ensemble de **formation** est constitué des audits de APGEN et de leurs résultats pour les années 2004 à 2006
- les **données de test** étaient constituées d'audits de l'APGEN en 2007 et ont été utilisées pour tester ou évaluer les modèles (grandes et petites entreprises) construits sur l'ensemble de données de formation
- la **validation** a été évaluée en effectuant des vérifications sur le terrain des prédictions faites par les modèles construits sur les données des déclarations de taxe d'utilisation de 2007 traitées en 2008

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015



Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

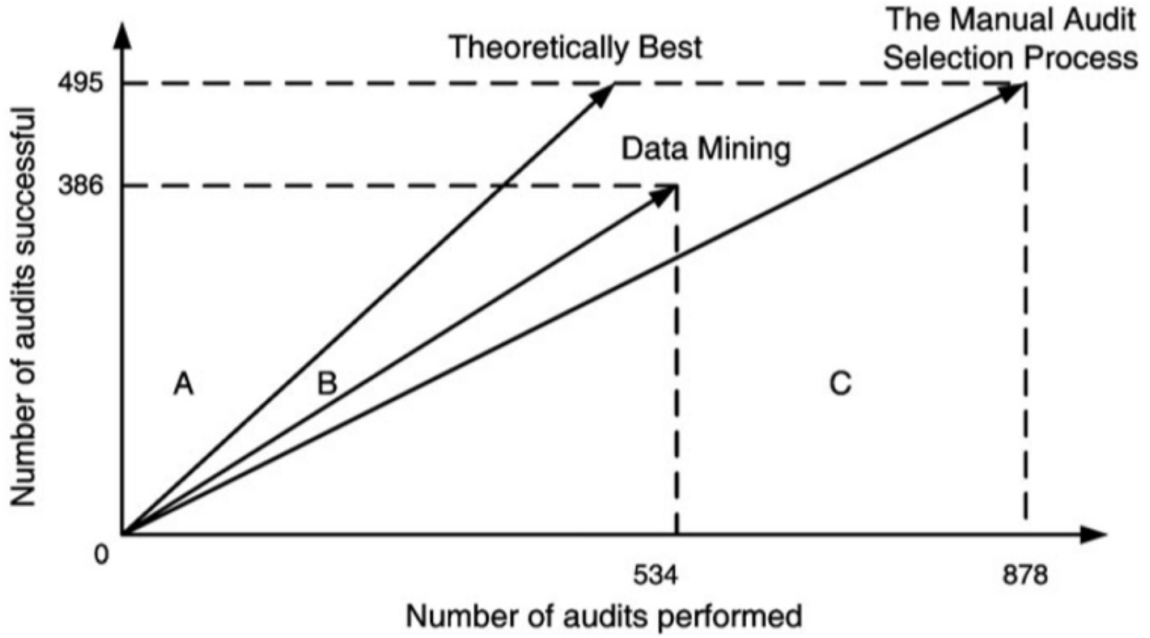
Points forts et limites des algorithmes

- La classification Naïve Bayes suppose l'indépendance des caractéristiques, ce qui est rarement le cas dans les situations réelles. Cette approche est également connue pour potentiellement introduire des biais dans les schémas de classification. Malgré cela, les modèles de classification construits à l'aide de cette approche ont fait leurs preuves.
- MultiBoosting est une **technique ensembliste** qui utilise le comité (c'est-à-dire des groupes de modèles de classification) et la "sagesse du groupe" pour faire des prédictions ; elle se distingue par le fait qu'elle forme un comité de sous-comités, ce qui a tendance à réduire à la fois le biais et la variance des prédictions.

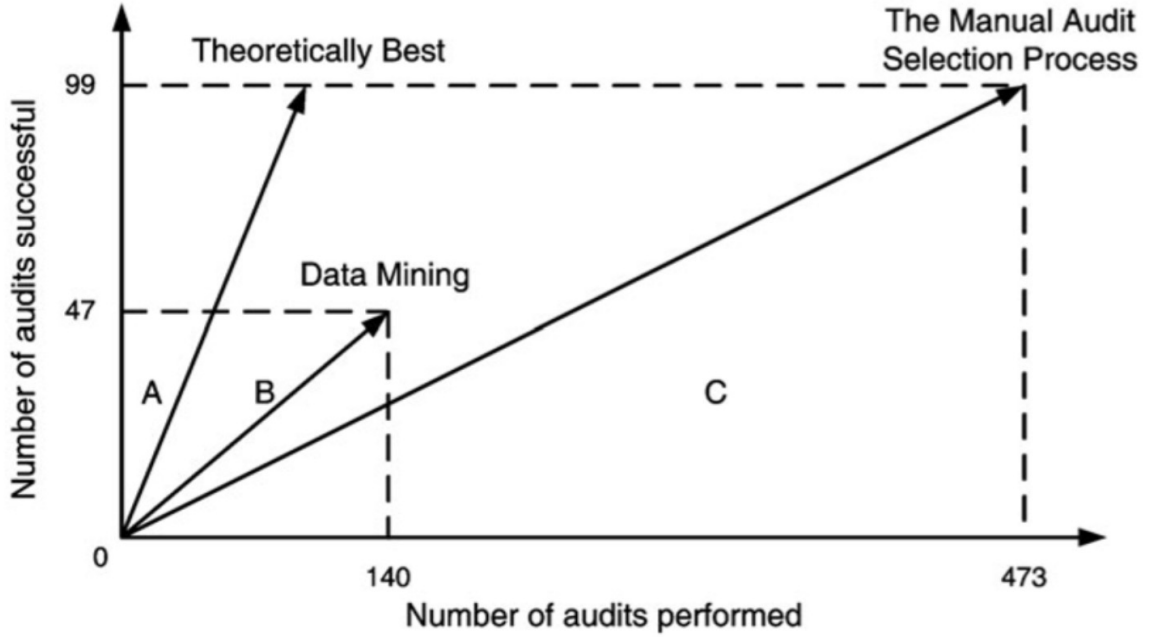
Étude de cas

Audits fiscaux au Minnesota

APGEN Large



APGEN Small



Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

APGEN Large

	Predicted as good	Predicted as bad
Actually good	386 (Use tax collected) R = $5,577,431 (83.6 %) C = $177,560 (44 %)	109 (Use tax lost) R = $925,293 (13.9 %) C = $50,140 (12.4 %)
Actually bad	148 (costs wasted) R = $72,744 (1.1 %) C = $68,080 (16.9 %)	235 (costs saved) R = $98,105 (1.4 %) C = $108,100 (26.7 %)

APGEN Small

	Predicted as good	Predicted as bad
Actually good	47 (Use tax collected) R = $263,706 (42.5 %) C = $21,620 (9.9 %)	52 (Use tax lost) R = $264,101 (42.5 %) C = $23,920 (11 %)
Actually bad	93 (costs wasted) R = $24,441 (3.9 %) C = $42,780 (19.7 %)	281 (costs saved) R = $68,818 (11.1 %) C = $129,260 (59.4 %)

Étude de cas

Audits fiscaux au Minnesota

Hsu *et al.*
Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue
Real World Data Mining Applications, 2015

Résumé et résultats du projet

- De nombreux modèles ont été produits avant que l'équipe ne fasse une sélection finale.
- Les performances passées d'une famille de modèles dans le cadre d'un projet antérieur peuvent guider la sélection, mais n'oubliez pas le *théorème du "No Free Lunch"*
- Le processus de sélection des caractéristiques peut très bien nécessiter un certain nombre de visites à des experts du domaine.
- Les équipes chargées de l'analyse des données devraient avoir une bonne compréhension des données/contextes.
- Les connaissances spécifiques au domaine doivent être intégrées dans le modèle afin de battre les classificateurs aléatoires, en moyenne.
- Même de légères améliorations par rapport à l'approche actuelle peuvent trouver une place utile dans une organisation!

Commentaires généraux

La classification est liée à l'**estimation des probabilités**

- les approches basées sur des modèles de régression pourraient s'avérer fructueuses

Les événements rares (souvent plus intéressants ou importants) :

- les données historiques du réacteur nucléaire de Fukushima avant le meltdown n'auraient pas pu être utilisées pour en apprendre davantage sur les meltdowns, par exemple
- prédire qu'il n'y aura pas de meltdown donnera des prédictions correctes dans environ 99.99 % des cas, mais manquera l'objectif de l'exercice.

Aucun classificateur ne donne les meilleurs résultats dans tous les scénarios.

Avec des données massives, on doit également tenir compte de l'efficacité.

Lectures conseillées

Aperçu de la classification

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.4 Classification and Regression
    - Overview
    - Case Study: Minnesota Tax Audits

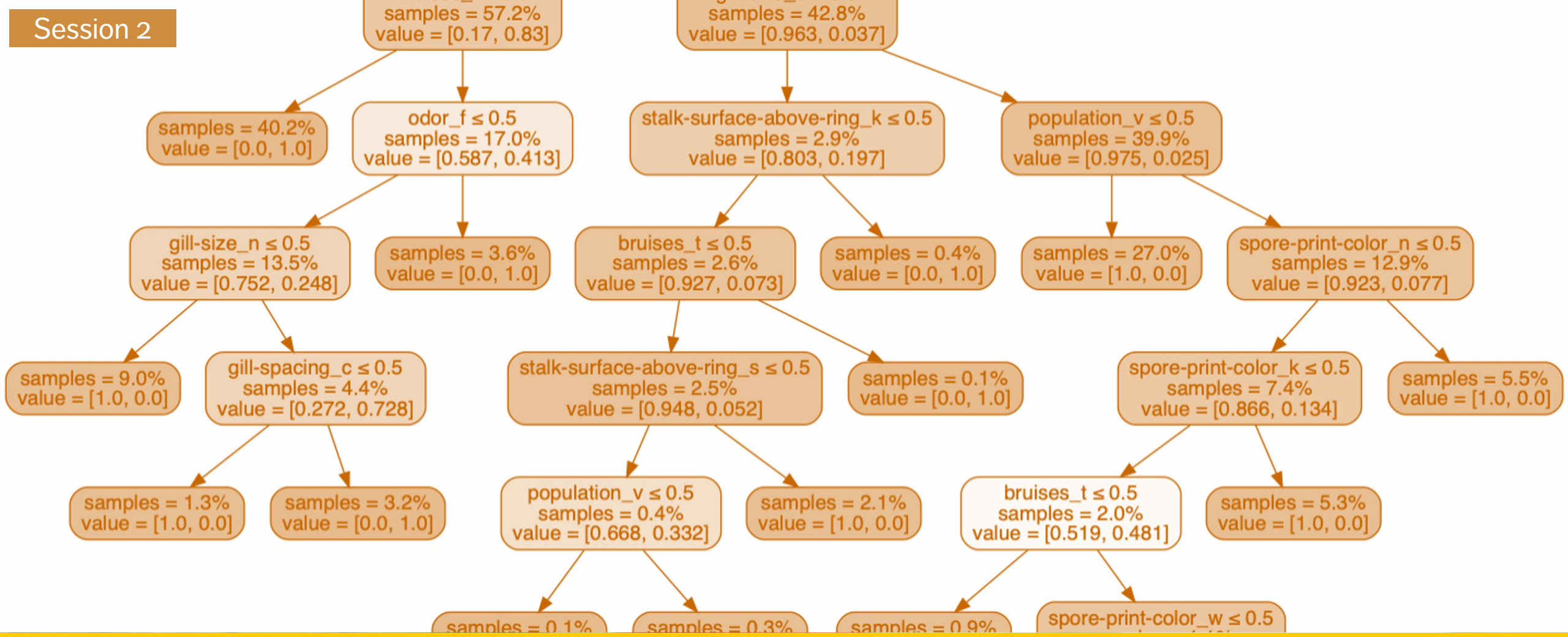
21. Focus on Classification and Supervised Learning
    21.1 Overview
    - Formalism

Exercices

Aperçu de la classification

1. Comment utiliseriez-vous les techniques de modélisation statistique standard pour répondre aux questions présentées dans les deux scénarios des diapositives ?

2. Identifier des scénarios et des questions qui pourraient faire appel à la classification et/ou à l'estimation de la valeur dans vos activités professionnelles quotidiennes.



5. Arbres de décision et autres algorithmes

Algorithmes de classification

Régression logistique

- modèle classique
- affectée par l'inflation de la variance et le processus de sélection des variables

Réseaux neuronaux

- difficile à interpréter
- exige que toutes les variables soient du même type
- plus facile à former depuis la rétropropagation (règle de dérivée en chaîne)

Méthodes Bayésiennes

Arbres de décision

- peut sur-ajuster les données s'il n'est pas élaguée correctement (manuellement ?)

Algorithmes de classification

Classificateurs Naïf de Bayes

- très efficace pour les applications d'exploration de texte (filtre anti-spam)
- hypothèses qui ne sont pas souvent respectées dans la pratique

Machines à vecteurs de support

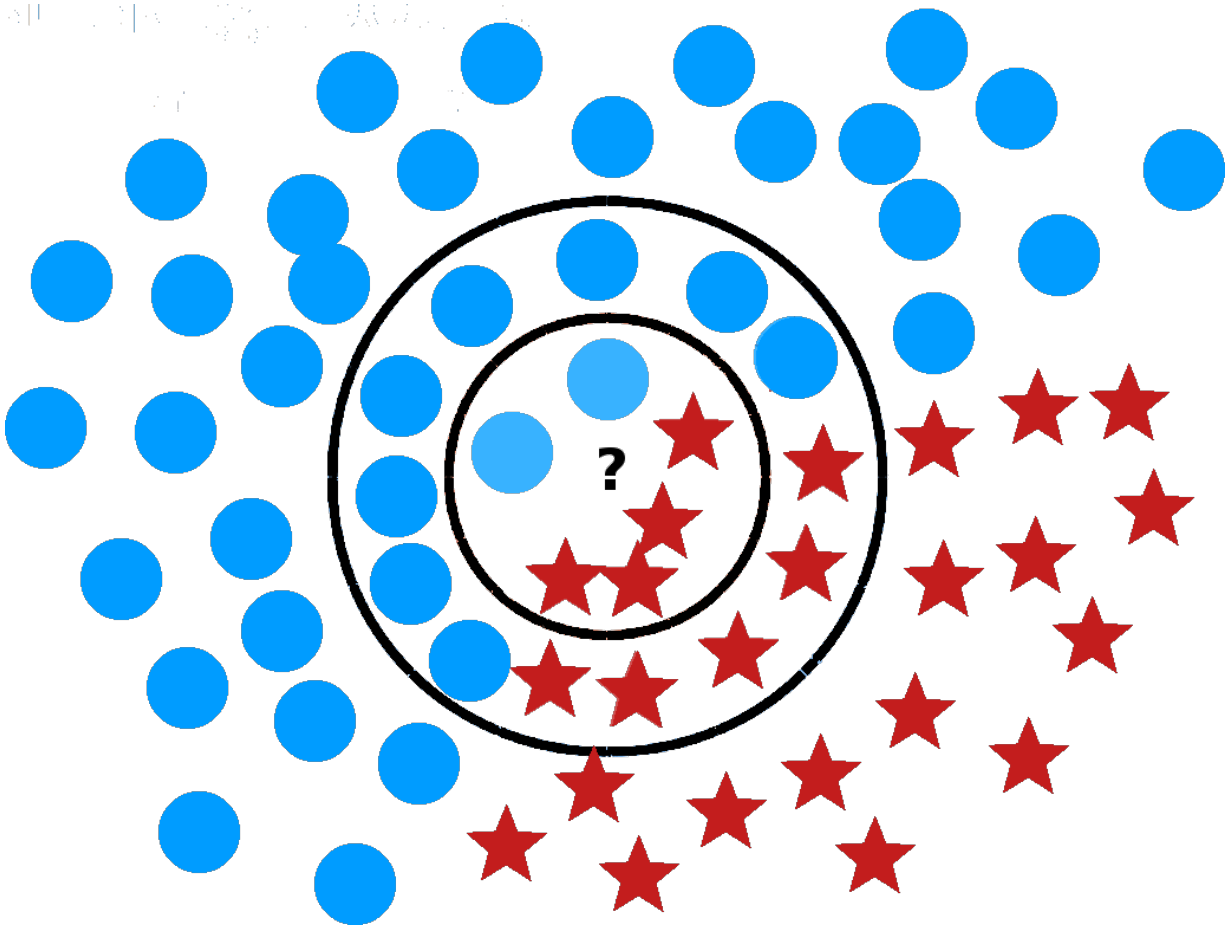
- peut être difficile à interpréter (frontières non linéaires)
- peut contribuer à atténuer les difficultés liées au données massives

Méthodes de stimulation (“boosting”)

Classificateurs des plus proches voisins

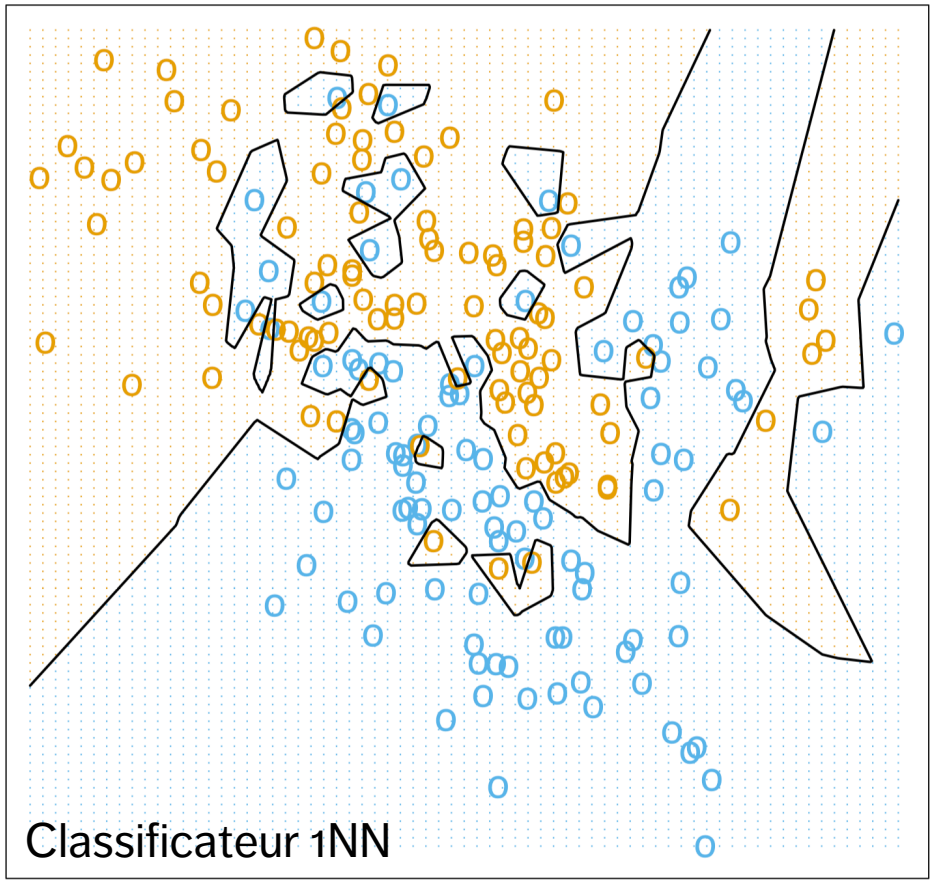
- ne nécessitent que très peu d'hypothèses sur les données
- pas très stable (l'ajout de points peut modifier substantiellement la frontière)

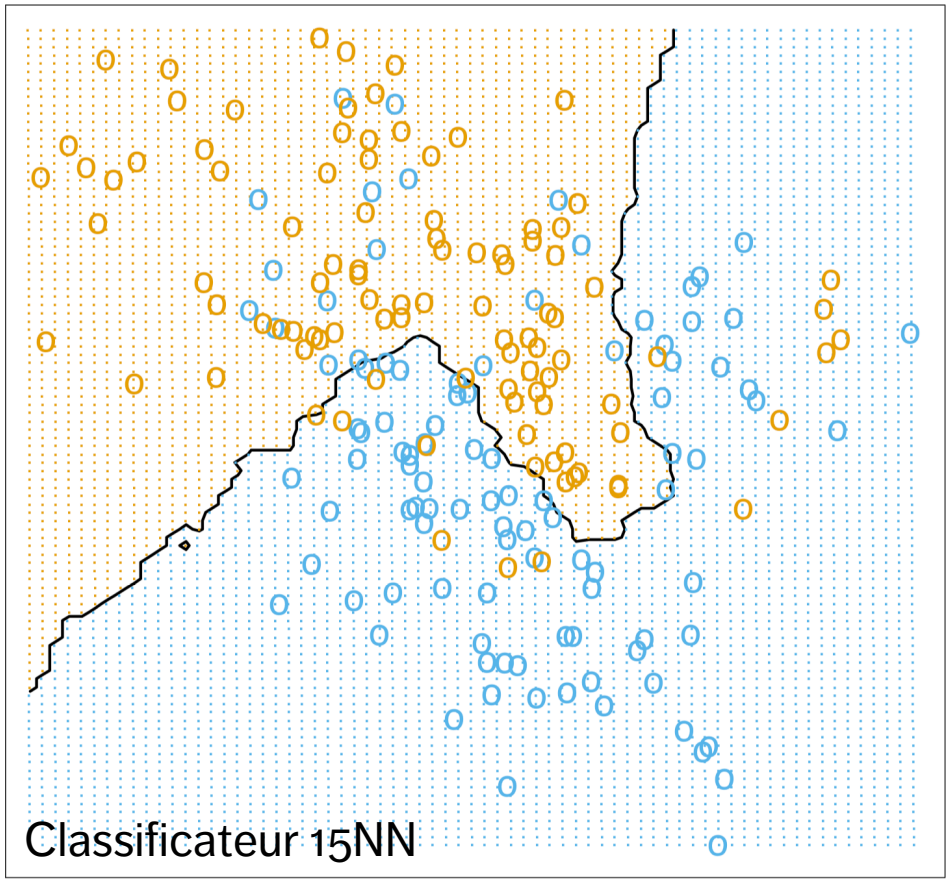
Classificateur des k plus proches voisins



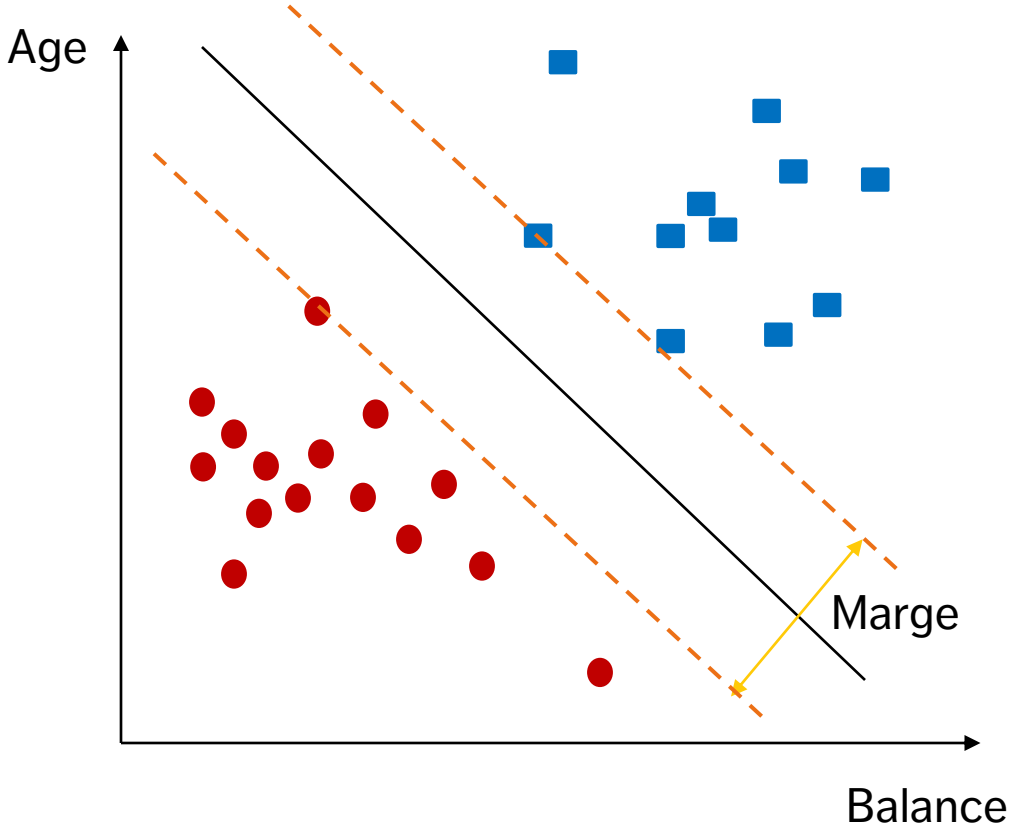
[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

Classificateur des k plus proches voisins



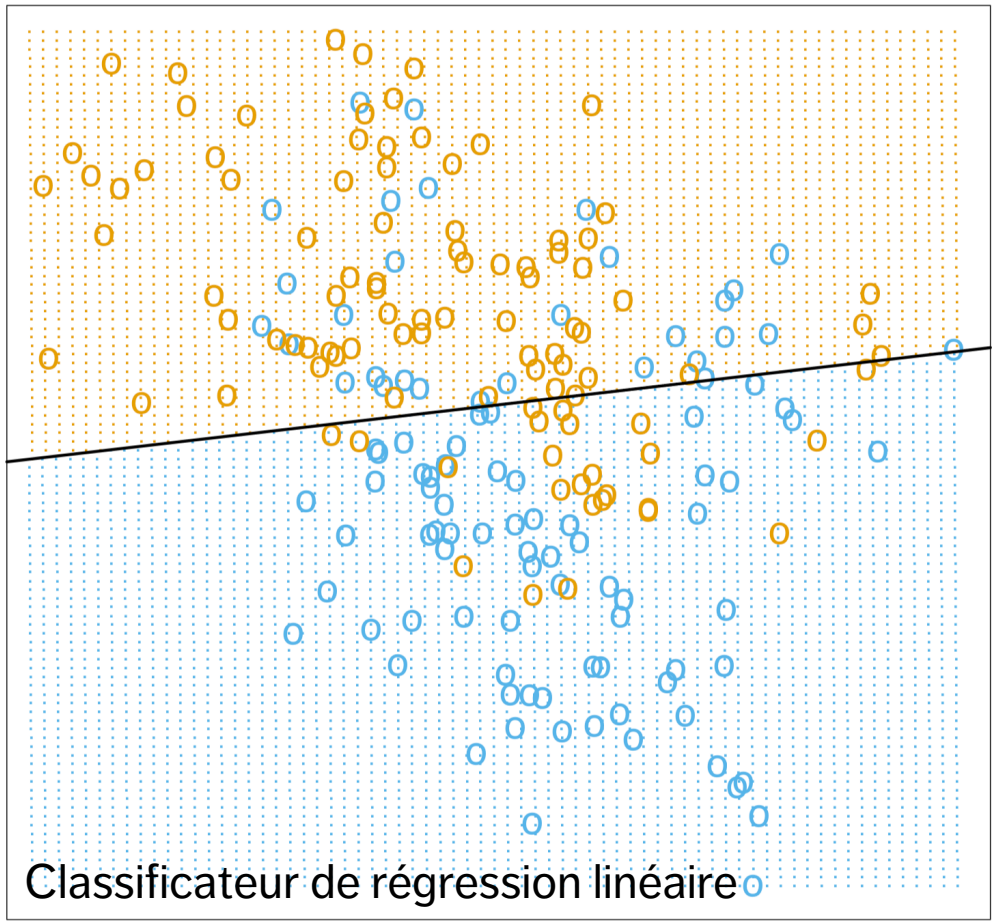


Machines à vecteurs de support

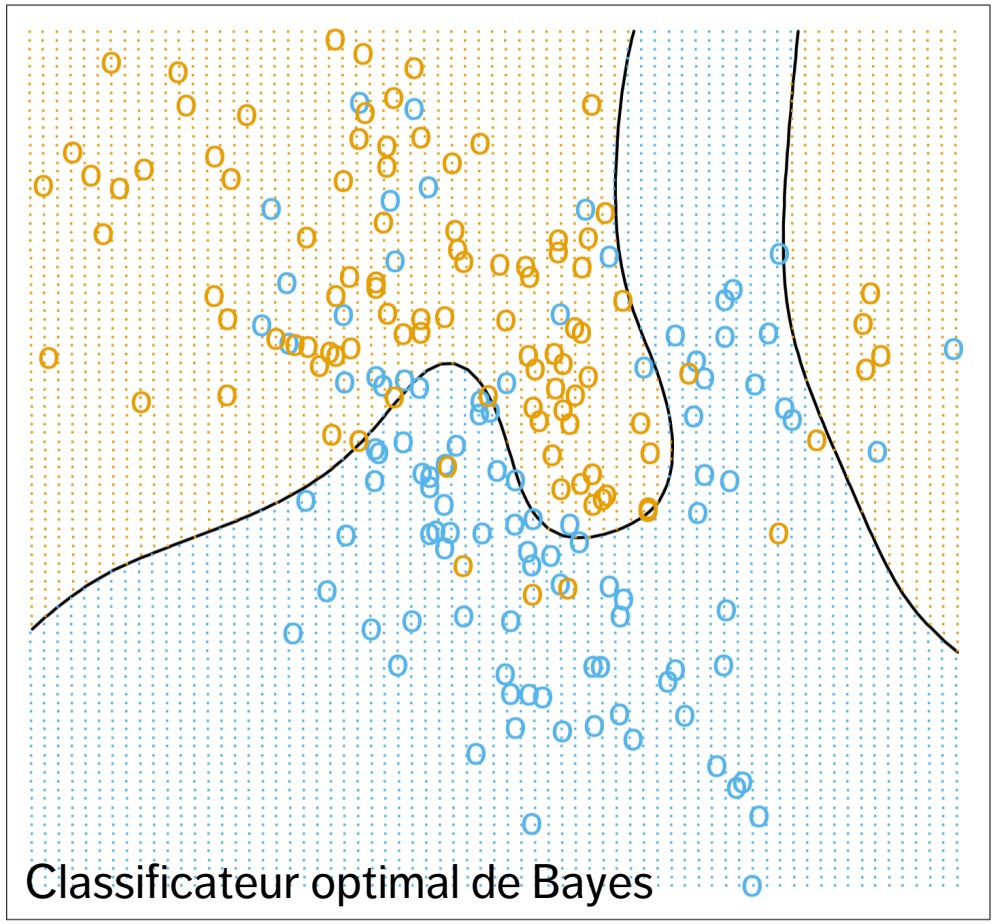


[Tibshirani, Hastie & Friedman, *The Elements of Statistical Learning*]

Autres classificateurs



Classificateur de régression linéaire



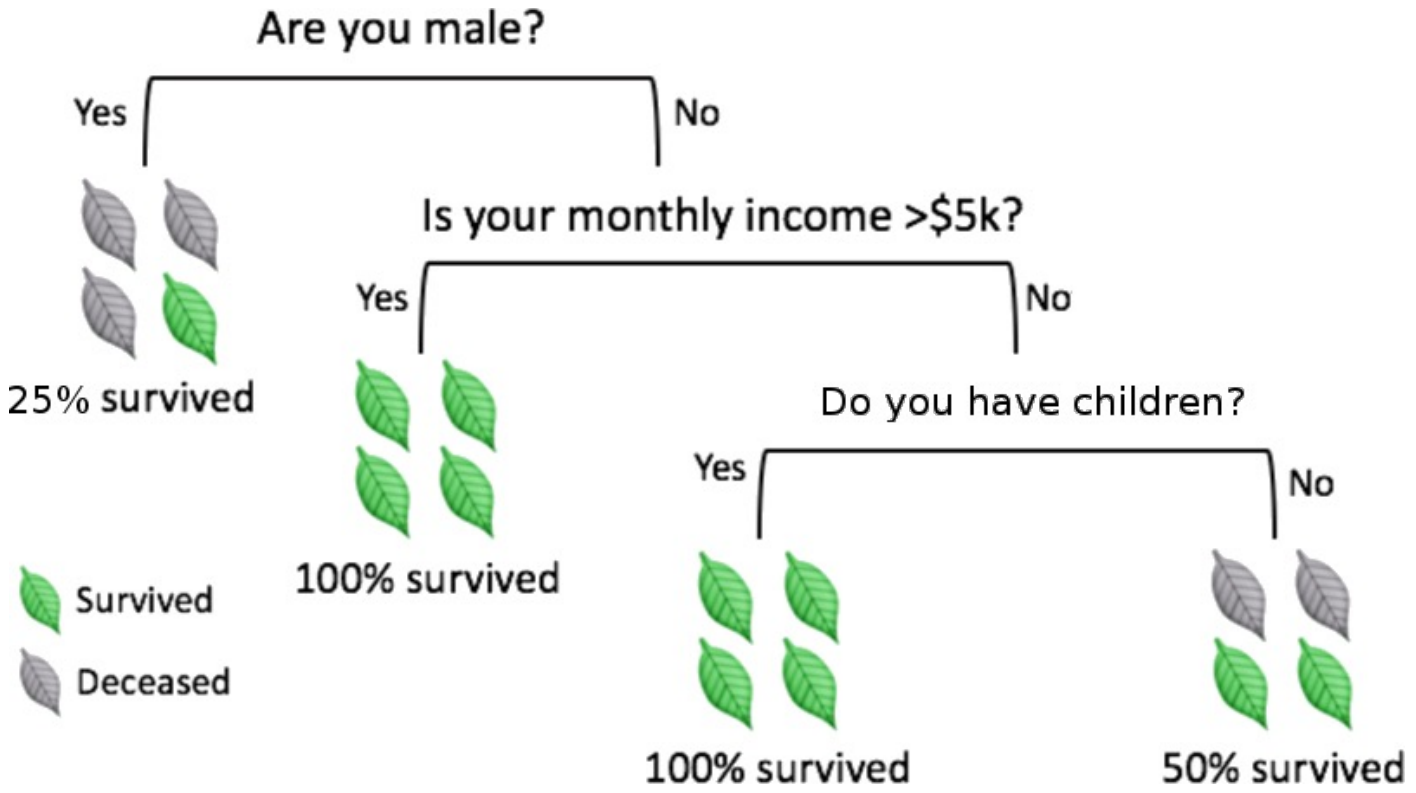
Classificateur optimal de Bayes

[adapté de A. Ng, K. Soo, *Numsense!*, Surviving a Disaster].

Arbres de décision

Les arbres de décision sont peut-être la plus **intuitive** de ces méthodes.

La classification est obtenue en suivant un chemin dans l'arbre, depuis sa **racine** jusqu'à ses **feuilles,** en passant par ses **branches**.



Arbres de décision

Pour faire une **prédiction** pour une nouvelle instance, on suit le chemin vers le bas de l'arbre et on lit directement la prédiction une fois une feuille atteinte.

La création de l'arbre et sa traversée peuvent **prendre du temps** s'il y a trop de variables.

La précision des prédictions peut être problématique dans les arbres dont la croissance **n'est pas contrôlée**. En pratique, le critère de **pureté** au niveau des feuilles est lié à de mauvais taux de prédiction pour nouvelles instances.

- d'autres critères sont souvent utilisés pour élaguer les arbres, ce qui peut conduire à des feuilles **impures** (c'est-à-dire avec une entropie non triviale).

Algorithme ID3

Objectif : développer un arbre de décision à l'aide d'un ensemble de formation.

Aperçu

1. Diviser l'ensemble de données de formation (**parent)** en sous-ensembles (**enfants**), en utilisant les différents niveaux d'un attribut particulier.
2. Calculer le **gain d'information** pour chaque sous-ensemble
3. Choisir la répartition la **plus avantageuse** au niveau du gain
4. Répéter l'opération pour chaque sous-ensemble jusqu'à ce qu'un critère soit rempli pour les **feuilles** (il est possible que chaque élément de feuille soit classé de la même manière).

Gain d'information

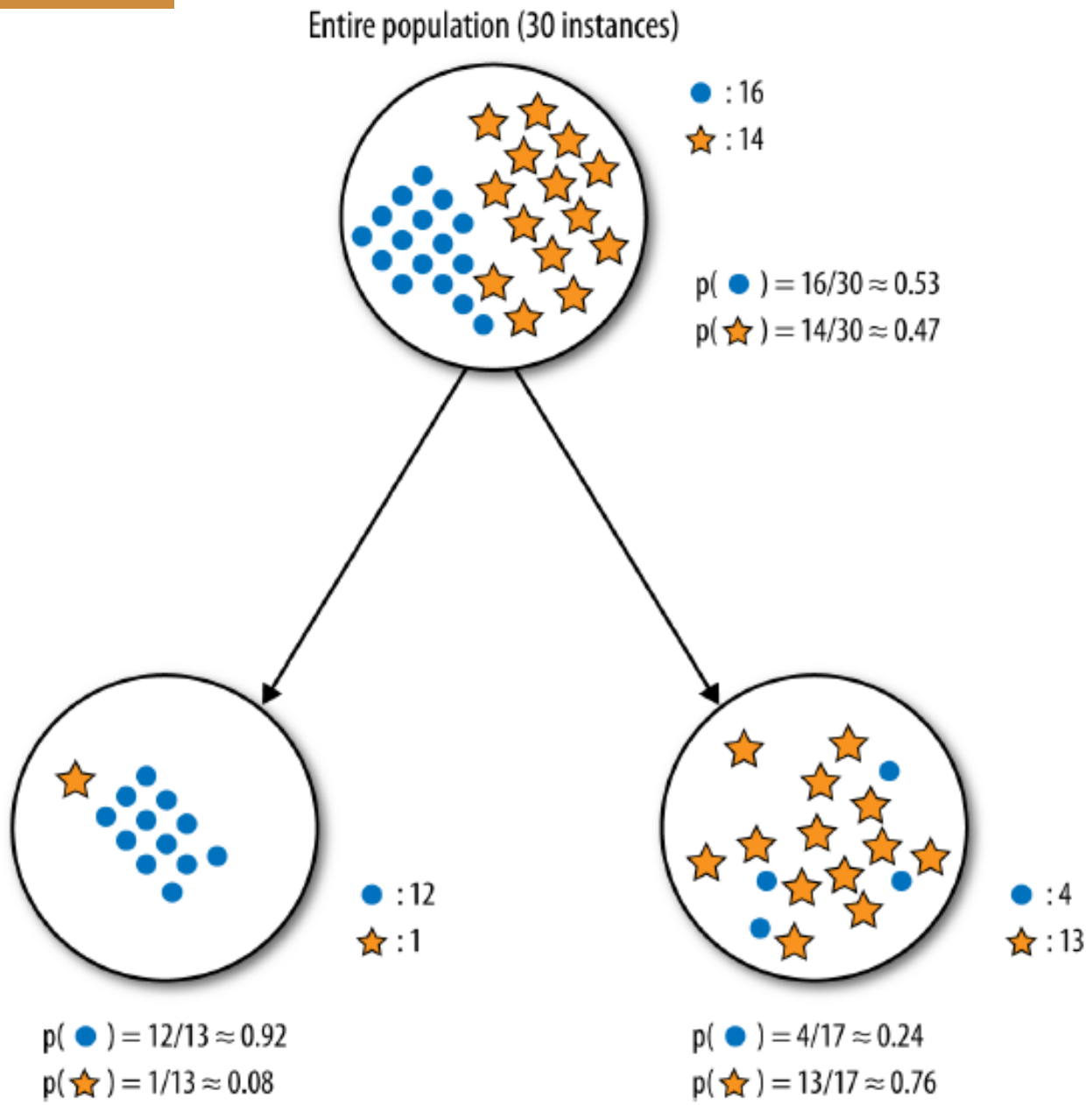
L'entropie est une mesure du désordre dans un ensemble S. Soit p_i le % d'observations de S appartenant à la catégorie i, pour $i = 1, \ldots, n$. L'entropie de S est donnée par

$$E(S) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log n.$$

Si l'**ensemble parent** S contient m observations divisées en k **ensembles enfants** $C_1, \ldots, C_k$ contenant $q_1, \ldots, q_k$ observations, le **gain d'information** résultant de la division est

$$\mathrm{IG}(S; C) = E(S) - \frac{1}{m}[q_1 E(C_1) + \cdots + q_k E(C_k)].$$

[Provost & Fawcett, *Data Science for Business*]



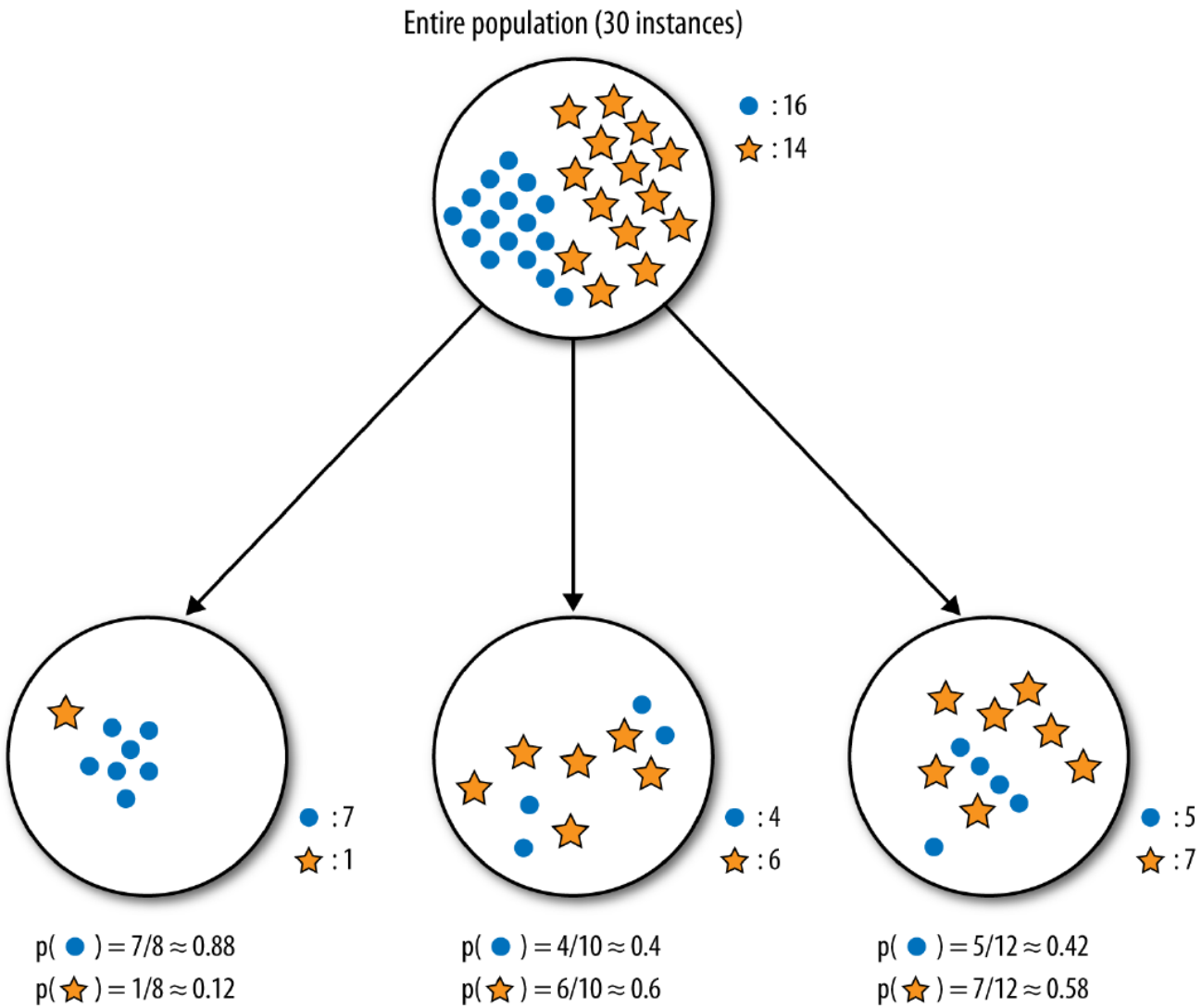
$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log\frac{1}{13} \approx 0.39$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{17}\log\frac{4}{17} - \frac{13}{17}\log\frac{13}{17} \approx 0.79$$

$$\text{IG} = E(S) - \frac{1}{30}[q_L E(L) + q_R E(R)]$$
$$\approx 0.99 - \frac{1}{30}[13(0.39) + 17(0.79)]$$
$$\approx \mathbf{0.37}$$

[Provost & Fawcett, *Data Science for Business*]



$$E(S) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{16}{30}\log\frac{16}{30} - \frac{14}{30}\log\frac{14}{30} \approx 0.99$$

$$E(L) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{7}{8}\log\frac{7}{8} - \frac{1}{8}\log\frac{1}{8} \approx 0.54$$

$$E(C) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10} \approx 0.97$$

$$E(R) = -p_\circ \log p_\circ - p_* \log p_*$$
$$= -\frac{5}{12}\log\frac{5}{12} - \frac{7}{12}\log\frac{7}{12} \approx 0.98$$

$$\mathrm{IG} = E(S) - \tfrac{1}{30}[q_L E(L) + q_C E(C) + q_R E(R)]$$
$$\approx 0.99 - \tfrac{1}{30}[8(0.54) + 10(0.97) + 12(0.98)]$$
$$\approx \mathbf{0.13}$$

Points forts

Modèle **boîte blanche**

- les prédictions peuvent toujours être expliquées en suivant les chemins appropriés

Peut être utilisé avec des ensembles de données **incomplets**

Sélection de caractéristiques **intégrées**

- les caractéristiques moins pertinentes n'ont pas tendance à être utilisées

Ne fait **aucune supposition** au sujet de :

- indépendance, variance constante, distributions sous-jacentes, colinéarité

Limites

Pas aussi précis que les autres algorithmes (en général)

De petites modifications dans l'ensemble de formationpeuvent conduire à des arbres (et prédictions) complètement différents

Particulièrement vulnérable à l'**overfitting** en l'absence d'**élagage**

- les procédures d'élagage sont généralement compliquées

L'apprentissage optimal des arbres de décision est **NP-complet**

Biais vers les caractéristiques catégorielles avec un **grand nombre** de niveaux

Notes

Métriques de séparation/division :

- gain d'information, impureté de Gini, réduction de la variance, etc.

Variantes courantes :

- ID3, C4.0, C4.5, CHAID, MARS, arbres d'inférence conditionnelle, CART

Les arbres de décision peuvent également être combinés à l'aide d'algorithmes d'amplification (**AdaBoost**) ou de **forêts aléatoires**, ce qui permet d'obtenir une sorte de procédure de **vote** (apprentissage d'ensemble).

Lectures conseillées

Arbres de décision et autres algorithmes

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    - 19.4 Classification and Regression
      - Classification Algorithms
      - Decision Tree
      - Toy Example: Kyphosis Dataset
    - 19.7 R Examples
      - Classification: Kyphosis Dataset

21. Focus on Classification and Supervised Learning
    - 21.2 Simple Classifiers
    - 21.3 Rare Occurrences
    - 21.4 Other Supervised Approaches
    - 21.5 Ensemble Learning

Exercices

Arbres de décision et autres algorithmes

1. Passez en revue l'exemple de classification des cyphoses trouvé dans DUDADS (voir les lectures conseillées). Répétez le processus avec l'ensemble de données du Titanic (vous pouvez d'abord visualiser l'ensemble de données) afin de construire un arbre de décision qui vous aidera à déterminer si un passager a survécu au naufrage ou non.

Exercices

Arbres de décision et autres algorithmes

2. UniversalBank cherche à convertir ses clients **passifs** (qui n'ont que des dépôts à la banque) en clients **actifs** (qui ont contracté un prêt auprès de la banque). Lors d'une campagne précédente, *UniversalBank* a réussi à convertir 9.6% de ses 5000 clients passifs en clients actifs. Le service de marketing aimerait comprendre quelle combinaison de facteurs rend un client plus susceptible d'accepter un prêt personnel, afin de mieux concevoir la prochaine campagne de conversion.

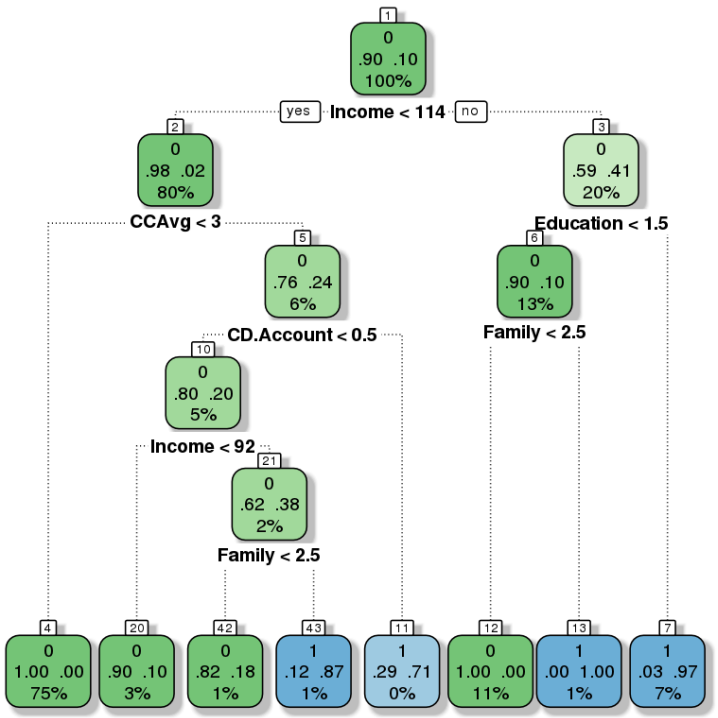
   L'ensemble de données contient des données sur les 5000 clients : âge, années d'expérience professionnelle, revenu annuel (en milliers de dollars), taille de la famille, valeur de l'hypothèque auprès de la banque, si le client a un certificat de dépôt auprès de la banque, une carte de crédit, etc.

Exercices

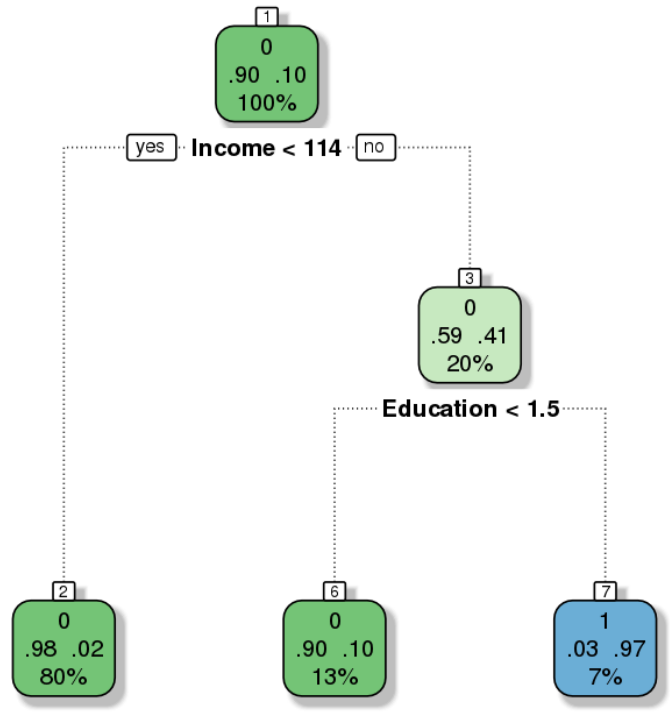
Arbres de décision et autres algorithmes

2. (suite) Nous construisons deux arbres de décision sur un sous-ensemble de formation de 3000 observations afin deprédire si un client est susceptible d'accepter un prêt personnel (1) ou non (0).

Arbre



Arbre



Exercices

Arbres de décision et autres algorithmes

a. Combien de variables sont utilisées dans la construction de l'arbre A? De l'arbre B?

b. La règle de décision suivante est-elle valide ou non pour l'arbre A :
`SI (Income` ≥ 114`) ET (Education` ≥ 1.5`)`
`ALORS (Personal Loan` $= 1$`)?`

c. La règle de décision suivante est-elle valide ou non pour l'arbre B:
`SI (Income` < 92`) ET (CCAvg` ≥ 3`)`
`ET (CD.Account` < 0.5`)`
`ALORS (Personal Loan` $= 0$`)?`

d. Quel pronostic ferait-on avec l'arbre A pour un client ayant :
- revenu annuel de 94,000 $USD (`Income = 94`),
- 2 enfants (`Family = 4`),
- pas de certificat de dépôt auprès de la banque (`CD.Account = 0`),
- un taux d'intérêt de carte de crédit de 3.2 % (`CCAvg = 3.2`), et
- un diplôme d'ingénieur (`Education = 3`).

e. Et avec l'arbre B?

Actual \ Predicted	A	B	C	D	Total
A	50	10	30	20	110
B	15	20	30	15	80
C	20	10	30	40	100
D	15	15	30	50	110
Total	100	55	120	125	800

6. Évaluation de la performance

Sélection du modèle

En conséquence du **théorème de la gratuité (“No Free Lunch”)**, aucun classificateur ne peut être le plus performant pour tous les problèmes.

La sélection du modèle doit prendre en compte :

- la **nature** des données disponibles
- les **fréquences relatives des sous-groupes de la classification**
- les **objectifs déclarés de la classification**
- la facilité avec laquelle le modèle se prête à l'**interprétation** et à l'**analyse statistique**
- le degré nécessaire de **préparation des données**

Sélection du modèle

La sélection du modèle doit prendre en compte (suite) :

- s'il peut incorporer différents types de données et d'observations manquantes
- s'il est efficace avec des données massives, et
- s'il est **robuste** face à des écarts minimes entre les données et les hypothèses théoriques.

Le succès passé n'est pas une garantie de succès futur – il est de la responsabilité des analystes d'essayer une **variété de modèles**.

Mais comment sélectionner le "**meilleur**" modèle ?

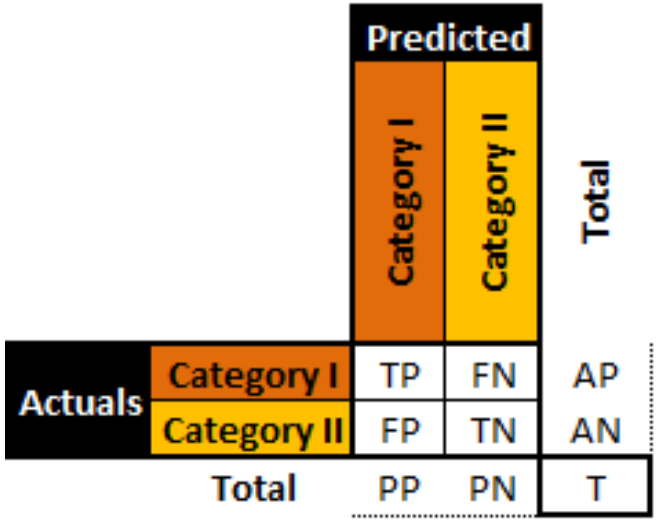
Erreurs de classification

Lorsqu'on essaie de déterminer le type de musique qu'un nouveau client préférerait, il n'y a pas de coût réel à se tromper ; en revanche, si le classificateur tente de déterminer la presence/absence de cancer dans le tissu pulmonaire, les erreurs sont **plus lourdes de conséquences**.

Plusieurs métriques peuvent être utilisées pour évaluer les performances d'un classificateur, en fonction du contexte.

Les **classificateurs binaires** sont plus simples et ont été étudiés depuis plus longtemps que les classificateurs multiniveaux ; par conséquent, un plus grand nombre de mesures d'évaluation est disponible pour ces classificateurs.

Classificateurs binaires



TP, TN, FP, FN: **: Vrais positifs**, **Vrais négatifs**, **Faux positifs** et **Faux négatifs,** respectivement.

Les classificateurs parfaits auraient $FP, FN = 0$, mais cela se produit rarement en pratique (à éviter, en fait).

Mesures :

- sensibilité
- spécificité
- précision
- rappel
- valeur prédictive négative
- taux de faux positifs
- taux de fausse découverte
- taux de faux négatifs
- précision

Autres mesures :

F_1-score, ROC AUC, information, marquage, coefficient de corrélation de Matthews (MCC), etc.

Session 2

		Predicted A	Predicted B	Total	Total
Actuals	A	54	10	64	79.0%
	B	6	11	17	21.0%
Total		60	21	81	
		74.1%	25.9%		

Classification Rates	
Sensitivity:	0.84
Specificity:	0.65
Precision:	0.90
Negative Predictive Value:	0.52
False Positive Rate:	0.35
False Discovery Rate:	0.10
False Negative Rate:	0.16

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.49
Markedness:	0.42
M.C.C.:	0.46
Pearson's chi2:	0.01
Hist. Stat:	0.10

		Predicted A	Predicted B	Total	Total
Actuals	A	54	0	54	66.7%
	B	16	11	27	33.3%
Total		70	11	81	
		86.4%	13.6%		

Classification Rates	
Sensitivity:	1.00
Specificity:	0.41
Precision:	0.77
Negative Predictive Value:	1.00
False Positive Rate:	0.59
False Discovery Rate:	0.23
False Negative Rate:	0.00

Performance Metrics	
Accuracy:	0.80
F1-Score:	0.87
Informedness (ROC):	0.41
Markedness:	0.77
M.C.C.:	0.56
Pearson's chi2:	0.33
Hist. Stat:	0.40

Les deux classificateurs ont une précision de 80 % ; le second classificateur fait quelques prédictions erronées pour A, mais jamais pour B ; le premier classificateur fait des erreurs pour les deux classes. Le second classificateur prédit à tort A par B à 16 reprises, alors que le premier ne le fait que 6 fois. Le choix du meilleur classificateur dépend du **coût de la classification erronée**.

Classificateurs à plusieurs niveaux

Il est préférable de choisir des mesures qui se généralisent plus facilement (**multiniveaux**).

Précision : proportion de prédictions correctes parmi toutes les observations

- la valeur est comprise entre 0% et 100%
- plus la précision est élevée, meilleure est la correspondance
- un modèle prédictif très précis peut s'avérer inutile en raison du **paradoxe de la précision**

Coefficient de corrélation de Matthews (MCC) : utile même lorsque les classes sont de tailles très différentes.

- coefficient de corrélation entre les classifications réelles et prédites
- varie de -1 à 1
- si $\mathrm{MCC} = 1$, les réponses prévues et réelles sont identiques
- si $\mathrm{MCC} = 0$, le classificateur ne fait pas mieux qu'une prédiction aléatoire ("pile ou face").

Classificateurs à plusieurs niveaux

MCC: 69.7%
Accuracy: 78.3%
Pearson: 0.13161
Hist: 30.0%

			Predicted: Maltreatment: Unfounded	Predicted: Maltreatment: Suspected	Predicted: Maltreatment: Substantiated	Predicted: Risk: No	Predicted: Risk: Yes	Predicted: Risk: Unknown	Total	
Actuals	Maltreatment	Unfounded	4,577	-	-	198	6	-	*4,781*	*29.2%*
		Suspected	-	965	-	29	2	-	*995*	*6.1%*
		Substantiated	-	-	6,187	116	35	2	*6,339*	*38.7%*
	Risk	No	894	-	763	949	19	9	*2,632*	*16.1%*
		Yes	123	-	520	122	111	5	*880*	*5.4%*
		Unknown	212	-	303	184	21	24	*745*	*4.6%*
	Total		*5,805*	*965*	*7,772*	*1,597*	*194*	*40*	***16,372***	
			35.5%	*5.9%*	*47.5%*	*9.8%*	*1.2%*	*0.2%*		

Évaluation de la performance

Pour les cibles numériques avec des prédictions , les métriques comprennent

- **erreurs quadratiques moyennes** et **erreurs absolues moyennes**

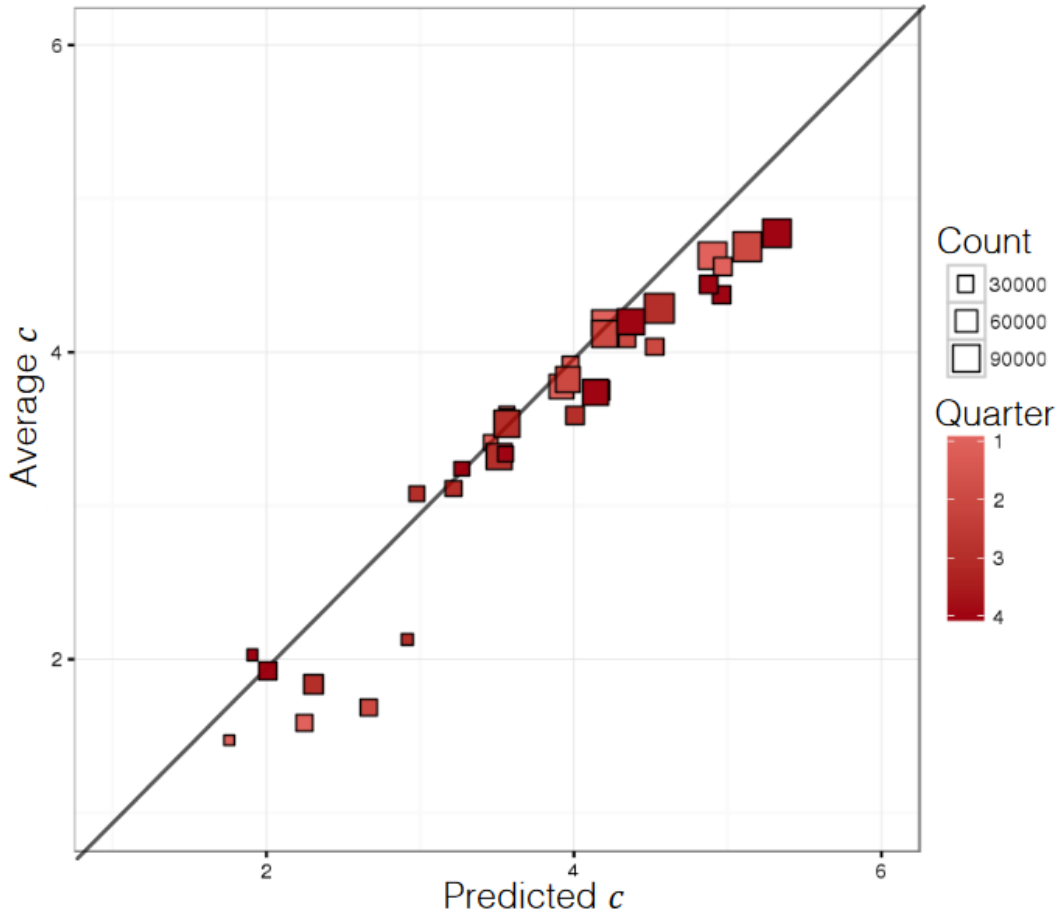
$$\text{MSE} = \text{moyenne}\{(\hat{y}_i - y_i)^2\},\ \text{MAE} = \text{moyenne}\{|\hat{y}_i - y_i|\}$$

- **erreur quadratique moyenne normalisée** et **erreur absolue moyenne normalisée**

$$\text{NMSE} = \frac{\text{moyenne}\{(\hat{y}_i - y_i)^2\}}{\text{moyenne}\{(\bar{y} - y_i)^2\}}, \text{NMAE} = \frac{\text{moyenne}\{|\hat{y}_i - y_i|\}}{\text{moyenne}\{|\bar{y} - y_i|\}}$$

- **moyenne du pourcentage d'erreur** $\text{MAPE} = \text{mean}\left\{\frac{|\hat{y}_i - y_i|}{y_i}\right\}$
- **corrélation** $\rho_{\hat{y},y}$
- etc.

Évaluation de la performance



Lectures conseillées

Évaluation de la performance

Data Understanding, Data Analysis, Data Science
Volume 3: Spotlight on Machine Learning

19. Introduction to Machine Learning
    19.4 Classification and Regression
    - Performance Evaluation

20. Regression and Value Estimation
    20.1 Statistical Learning
    - Model Evaluation

21. Focus on Classification and Supervised Learning
    21.1 Overview
    - Model Evaluation

Exercices

Évaluation de la performance

Nous poursuivons avec l'exemple de *UniversalBank*. Les matrices de confusion pour les prédictions des arbres A et B sur les observations de l'ensemble de test (2000 observations) sont présentées ci-dessous.

1. À l'aide des matrices appropriées, calculez les paramètres d'évaluation des performances pour chacun des arbres (sur l'ensemble de test).

2. Si les clients qui n'accepteraient pas un prêt personnel sont irrités lorsqu'on leur propose un prêt personnel, quel arbre *le* groupe de marketing doit-il utiliser pour maintenir de bonnes relations avec les clients ?

Arbre A

		Predicted		Total	
		A	B		
Actuals	A	1792	19	1811	90.55%
	B	18	171	189	9.45%
Total		1810	190	2000	
		90.50%	9.50%		

Arbre B

		Predicted		Total	
		A	B		
Actuals	A	1801	10	1811	90.55%
	B	64	125	189	9.45%
Total		1865	135	2000	
		93.25%	6.75%		