COURS DE DATA MINING
3 : MODELISATION
PRESENTATION GENERALE

EPF – 4/ 5ème année - Option Ingénierie d'Affaires et de Projets - Finance

Bertrand LIAUDET

PHASE 4 : MODELISATION

PROCESSUS du DATA MINING		
Acteurs	**Étapes**	**Phases**
Maître d'œuvre	Objectifs	1 : Compréhension du métier
	Données	2 : Compréhension des données
		3 : Préparation des données
	Traitements	**4 : Modélisation**
		5 : Évaluation de la modélisation
Maître d'ouvrage	Déploiement des résultats de l'étude	

Classement des techniques du data mining

Les techniques du data mining

Le data mining met en œuvre un **ensemble de techniques** issues des statistiques, de l'analyse de données et de l'informatique pour explorer les données.

Rappels de vocabulaire : concept, donnée, variable, type, modèle

On travaille sur des **tableaux de données**.

- Le **nom du tableau**, c'est « ce dont on parle », c'est-à-dire le « **concept** » dont on parle. C'est une **abstraction**. Par exemple, un tableau de clients, de malades, etc.

  Rappelons qu'un concept (ou **notion**, ou **idée**) est une représentation mentale générale et abstraite d'un objet. Le concept est le résultat de l'opération de l'esprit qui fait qu'on place tel objet dans telle catégorie et non dans telle autre.

- Chaque **colonne** du tableau a un nom qui est un **attribut** du concept. On parle aussi de « **propriété** » ou de « **champ** ». Le nom de la colonne est une abstraction (un concept). Pour un objet concret, la colonne a **une valeur** particulière qui est la valeur particulière de l'attribut pour l'objet concret.

  En data mining (et en statistique), **les attributs des objets sont appelés : « variables ».**

- Chaque ligne du tableau est un élément du tableau, c'est-à-dire un objet concret correspondant au concept abstrait dont on parle.

  En data mining, **un objet concret** est appelé : **« individu ».**

  En data mining, **la valeur d'un attribut** pour un individu est appelé : **« donnée ».**

  En data mining, **l'ensemble des individus** est appelé : **« population ».** Un tableau de données est une population.

- Un **sous-ensemble de valeurs** pour un ou plusieurs attributs donnés peut être appelé **: « type », « classe », « catégorie » , « segment » ou encore « modalité »**

  Par exemple, « grand » et « petit » sont deux types (ou classe, ou catégorie, ou segment) de l'attribut « taille ».

- On parle de **« variable catégorielle »** par opposition aux **« variables numériques ».** Par exemple, si la variable (attribut) « taille » peut prendre deux valeurs possibles : « grand » et « petit », c'est une variable catégorielle. Si les valeurs de la variable « taille » sont données en cm, c'est une variable numérique.

- **Quand on fait de la prévision**, on travaille sur une variable particulière appelée **: « variable cible »** et sur un ensemble d'autres variables utiles pour la prédiction appelées : **« prédicteurs ».**

  Le principe général de la prédiction sera : **si le ou les prédicteurs valent tant, alors la variable cible vaut tant.**

- Les statisticiens et les data miners construisent des **modèles**. Un modèle est un résumé global des relations entre variables permettant de comprendre des phénomènes (description, jugement) et d'émettre des prévisions (prédiction, raisonnement).

Dans l'absolu, tous les modèles sont faux. Un modèle n'est pas une loi scientifique. Cependant, certains sont utiles.

Première distinction : techniques descriptives et techniques prédictives

On distingue d'abord entre **deux grandes catégories de techniques** : les techniques descriptives et les techniques prédictives.

Les techniques descriptives (archétype : la classification)

- **Décrire**.
- Résumer, synthétiser, réduire, classer.
- Mettre en évidence des informations présentes mais cachées par le volume des données.
- **Pas de variable cible** à prédire.
- On les appelle aussi : **technique non supervisées.**
- Elles produisent des modèles de classement : typologie, méta-typologie.

Les techniques prédictives (archétype : le scoring)

- **Prédire**.
- Extrapoler de nouvelles informations à partir des informations présentes.
- Les techniques prédictives présentent **une variable cible** à prédire.
- L'objectif est de prévoir la variable cible mais aussi de classer à partir de la variable cible.
- On les appelle aussi : **techniques supervisées.**
- Elles sont plus délicates à mettre en œuvre que les techniques descriptives.
- Elles demandent plus d'historique que les techniques descriptives.
- Elles produisent des modèles de prédiction.

Deuxième distinction : variable numérique et variable catégorielle

Cette distinction est essentielle en statistique et en data mining.

Les **variables numériques** permettent de faire des résumés, des synthèses : moyenne, minimum, maximum, écart type, etc.

Les **variables catégorielles** permettent de faire des regroupement par catégories, c'est-à-dire des classements.

Les 6 grands types de techniques du data mining

Le data mining permet d'accomplir les six types d'analyse suivants :

1 : Description - 2 : Classification - 3 : Association

4 : Estimation - 5 : Segmentation - 6 : Prévision.

Ces types d'analyse se répartissent dans les techniques descriptives et prédictives :

Techniques descriptives		Techniques prédictives		
Corrélation simple	Corrélation complexe	Présent		Futur
		Variable cible numérique	Variable cible catégorielle	
1 : Description	2 : Classification 3 : Association	4 : Estimation	5 : Segmentation	6 : Prévision

Problèmes de vocabulaire et de traduction

Traduction

Anglais	Français
Clustering	segmentation ou classification
Classification	classification ou classement
Decision trees	arbres de décision ou segmentation

Le vocabulaire souligné est celui qu'on utilise dans ce cours.

Distinction entre classification et classement

Dans un **classement**, on sait à l'avance à quelle classe l'individu appartient car on connaît à l'avance les classes. Le classement est un tri pour les variables numérique, un « group by » SQL pour les variables catégorielles.

Dans une **classification**, on ne sait pas à l'avance à quelle classe un individu appartient car on ne connaît pas à l'avance les classes. La classification se fait en fonction de la population entière.

Exemple :

On peut classer les personnes par choix de l'option internationale et de l'option messagerie. Ca définit a priori 4 classes. C'est un **classement**.

On peut prendre tous les attributs des clients et chercher des classes de clients en fonction de tous ces attributs : ça donnera un nouvel attribut avec ses valeurs possibles.

Classement	Classification
Ne crée pas nécessairement de nouvel attribut	Crée nécessairement un nouvel attribut
Les classes sont définies à partir d'un attribut unique ou d'un petit nombre d'attributs.	Les classes sont définies à partir d'un grand nombre d'attributs
Une classe est connue à partir d'un individu	Les classes sont connues à partir de la population
Les classes et leur nombre sont connus *a priori*.	Les classes et leur nombre sont connus *a posteriori*.
La classe d'appartenance d'un individu est définie par l'individu lui-même.	La classe d'appartenance d'un individu est défini par ses relations avec la population.

Classement	**Classification**
Plutôt prédictif. Les données des attributs de classement sont utilisés pour prédire une variable cible. Exemple : superposition du « churn » en fonction du choix de l'option internationale.	Plutôt descriptif. Le classification crée un attribut de classification qui est la variable cible de la classification elle-même.

Les techniques concrètes

Le data mining utlise des techniques concrètes qui peuvent être limitées à un type de technique spécifique ou être être partagées par plusieurs types de techniques.

- Exemple de méthodes descriptives : la classification hiérarchique, la classification des K moyennes, les réseaux de Kohonen, les règles d'association.
- Exemples de méthodes prédictives : les méthodes de régression, les arbres de décision, les réseaux de neurones, les K plus proches voisins.

Les six grands types de technique du data mining

1 : la description (technique descriptive)

Principe :

La description consiste à mettre au jour

- Pour une variable donnée : la répartition de ses valeurs (tri, histogramme, moyenne, minimum, maximum, etc.).
- Pour deux ou trois variables données : des liens entre les répartitions des valeurs des variables. Ces liens s'appellent des **« tendances ».**

Intérêt :

- Favoriser la connaissance et la compréhension des données.

Méthode :

- Méthodes graphiques pour la clarté : **analyse exploratoire des données**.

Exemples :

- Répartition des votes par âge (lien entre les variables « vote » et « âge »).

2 : la classification (technique descriptive)

Principe :

La **classification** (ou ***clustering*** ou **segmentation**) consiste à créer des classes (c'est-à-dire des sous-ensembles) de données similaires entre elles et différentes des données d'une autre classe (autrement dit, l'intersection des classes entre elles doit toujours être vide).

Autrement dit, il s'agit pour n variables de créer des sous-ensembles disjoints de données. On dit aussi « **segmenter** » l'ensemble entier des données.

La classification définit les grands types de regroupement et de distinction : on parle de **métatypologie** (type de type).

Elle permet une vision générale de l'ensemble (de la clientèle, par exemple).

Intérêt :

- Favoriser, grâce à la métatypologie, la compréhension et la prédiction.
- Fixer des segments qui serviront d'ensemble de départ pour des analyses approfondies.
- Réduire les dimensions, c'est-à-dire le nombre d'attributs, quand il y en a trop au départ.

Méthodes :

- Classification hiérarchique
- Classification des K moyennes
- Réseaux de Kohonen.
- Règles d'association.

Exemples :

- Métatypologie d'une clientèle en fonction de l'âge, les revenus, le caractère urbain ou rural, la taille des villes, etc.

- Pour un audit comptable, classer un comportement financier en catégorie normale et suspecte.

3 : l'association (technique descriptive)

Principe :

L'association consiste à **trouver quelles valeurs des variables vont ensemble**. Par exemple, telle valeur d'une variable va avec telle valeur d'une autre variable.

Les règles d'association sont de la forme : si antécédent, alors conséquence.

L'association ne fixe pas de variable cible. Toute les variables peuvent à la fois être prédicteurs et variable cible.

On appelle aussi ce type d'analyse une « analyse d'affinité ».

Intérêt :

Mieux connaître les comportements.

Méthodes :

- Algorithme *a priori.*
- Algorithme du GRI (induction de règles généralisée).

Exemples :

- Analyse du panier de la ménagère (si j'achète des fraises, alors j'achète des cerises).
- Étudier quelle configuration contractuelle d'un abonné d'une compagnie de téléphone portable conduit plus facilement à un changement d'opérateur.

4 : l'estimation[1] (technique prédictive)

Principe :

L'estimation consiste à définir le lien entre un ensemble de prédicteurs et une variable cible. Ce lien est défini à partir de données « complètes », c'est-à-dire dont les valeurs sont connues tant pour les prédicteurs que pour la variable cible. Ensuite, on peut déduire une variable cible inconnue de la connaissance des prédicteurs.

À la différence de la segmentation (technique prédictive suivante) qui travaille sur une variable cible catégorielle, l'estimation travaille sur une variable cible numérique.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Analyse statistique classique : régression linéaire simple, corrélation, régression multiple, intervalle de confiance, estimation de points.
- Réseaux de neurones

Exemples :

- Estimer la pression sanguine à partir de l'âge, le sexe, le poids et le niveau de sodium dans le sang.
- Estimer les résultats dans les études supérieures en fonction de critères sociaux.

[1] Reprise du 1er cours.

5 : la segmentation (technique prédictive)

Principe :

La segmentation est une estimation qui travaille sur une variable cible catégorielle.

On parle de segmentation car chaque valeur possible pour la variable cible va définir un segment (ou type, ou classe, ou catégorie) de données.

La segmentation peut être vue comme une classification supervisée.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Graphiques et nuages de points.
- Méthode des k plus proches voisins.
- Arbres de décision.
- Réseau de neurones.

Exemples :

- Segmentation par tranche de revenus : élevé, moyen et faible (3 segments). On cherche les caractéristiques qui conduisent à ces segments.
- Déterminer si un mode de remboursement présente un bon ou un mauvais niveau de risque crédit (deux segments).

6 : la prévision (technique prédictive)

Principe :

La prévision est similaire à l'estimation et à la segmentation mise à part que pour la prévision, les résultats portent sur le futur.

Intérêt :

- Permettre l'estimation de valeurs inconnues.

Méthodes :

- Celles de l'estimation ou de la segmentation.

Exemples :

- Prévoir le prix d'action à trois mois dans le futur.
- Prévoir le temps qu'il va faire.
- Prévoir le gagnant du championnat de football, par rapport à une comparaison des résultats des équipes.

Historique des techniques de statistique et de data mining

➔	**1875**	**Régression linéaire de Francis Galton.**
➔	**1896**	**Formule du coefficient de corrélation de Karl Pearson[2].**
	1900	Distribution du X^2 de Karl Pearson.
	1936	Analyse discriminante de Fischer et Mahalanobis
➔	**1941**	**Analyse factorielle des correspondances de Guttman**
➔	**1943**	**Réseaux de neurones de Mac Culloch et Pitts**
	1944	Régression logistique de Joseph Berkson
	1958	Perceptron de Rosenblatt
	1962	Analyse des correspondances de J.-P. Benzécri
	1962	Régression logistique de J. Cornfield
➔	**1964**	**Arbre de décision AID de J.-P. Sonquist et J.-A. Morgan**
	1965	Méthode des centres mobiles de E. W. Forgy
➔	**1967**	**Méthode des k means (k moyennes) de Mac Queen**
	1971	Méthode des nuées dynamiques de Diday
	1972	Modèle linéaire généralisé de Nelder et Wedderburn
	1975	Algorithme génétique de Holland
	1977	Méthode de classement DISQUAL de Gilbert Saporta
	1980	Arbre de décision CHAID de KASS
	1983	Régression PLS de Herman et Svante Wold
	1984	Arbre CART de Breichman, Friedman, Olshen, Stone
	1986	Perceptron multicouches de Rumelhart et Mac Clelland
	1989	Réseaux de T. Kohonen (cartes auto-adaptatives)
➔	**1990**	**Apparition du concept de Data Mining**
	1993	Arbre C4.5 de J. Ross Quinlan
	1996	Bagging (Breiman) et boosting (Freund-Shapire)
	1998	Support vector machine de Vladimir Vapnik
	2001	**Régression logistique PLS de Tenenhaus**

[2] Karl Pearson, (1857-1936), mathématicien et philosophe britannique qui a mis au point les principales techniques statistiques modernes et les a appliquées aux questions de l'hérédité.

Fonctionnement général des méthodes de classification

Principe de la classification

Une classe est un ensemble d'éléments qui sont semblables entre eux et qui sont dissemblables à ceux d'autres classes.

Classifier consistera à maximiser les similarités des éléments qui sont dans la même classe et à minimiser les similarités de ces éléments avec ceux des autres classes. Inversement, on peut dire que classifier consiste à minimiser la variation intra-classe et à maximiser la variation inter-classe.

Classification et techniques supervisées

Quand on part d'un volume de données très important, on a intérêt à faire une classification préalable pour réduire l'espace de recherche des algorithmes supervisés.

Comment mesurer la similarité ? Notion de distance entre les enregistrements

C'est le premier problème inhérent à la classification.

La distance euclidienne entre deux enregistrements « x » et «y »est la suivante :

$$d(x,y)= \sqrt{\sum_i (x_i-y_i)^2}$$

$x = x_1, x_2 \ldots x_n$ représentent les valeurs des variables de « x ». De même pour « y ».

Il existe d'autres calculs de distance.

Pour que les distances soient comparables d'une variable à une autre, on va utiliser la technique des normalisations : normalisation « min-max » ou normalisation par le « test Z »

Normalisation « min - max » : x' = (x-min(x)) / amplitude(x)

Normalisation « test Z » : x'= (x - moy(x)) / écart type(x)

Comment mesurer les variables catégorielles ?

C'est le second problème inhérent à la classification.

Quand on a une variable booléenne, ça ne pose pas de difficulté. Faux vaut 0 et vrai vaut 1.

Pour des variables énumérées, on considérera que Si $x_i = y_i$ alors $x_i-y_i = 0$ sinon $x_i-y_i = 1$ (c'est une sorte de généralisation du cas précédent).

Fonctionnement général des méthodes supervisées

Rappels : variable cible et variables prédictives

Variable cible

La variable cible est la variable dont on cherche à connaître la valeur.

On parle aussi de :variable à expliquer, réponse, variable dépendante, variable endogène.

C'est la variable « en sortie ».

Variables explicatives

Les variables explicatives sont les variables utilisées pour fabriquer le modèle.

On parle aussi de variables prédictives ou de prédicteurs.

Ce sont les variables « en entrée ».

Définition générale d'un modèle prédictif

Un modèle prédictif est un ensemble de règles de découpage et d'association des variables explicatives. En appliquant ces règles à n'importe quel nouvel individu de la population, on pourra déterminer la valeur de l'individu pour la variable cible.

Les techniques prédictives sont nombreuses et leur domaine d'application tout autant. Elles servent aussi bien à calculer l'efficacité d'un traitement médical, à prévoir le temps en météorologie, qu'à prévoir le rendement d'une culture en agriculture.

Ces techniques ont un cadre théorique précis qu'il faut connaître pour les appliquer correctement.

Description intuitive d'un modèle prédictif

Le but est de connaître une information qu'on ne connaît pas.

Par exemple, on veut savoir si un client va rembourser le prêt qu'on lui fait.

Pour calculer cette information, on va s'intéresser aux clients qui ont déjà eu des prêts. Et on va chercher une corrélation générale entre les données économiques, sociales, géographiques et comportementales (le comportement des comptes) et le fait que ces clients aient ou n'aient pas remboursé leurs prêts. Cette corrélation, c'est le modèle prédictif. Une fois trouvée, on peut l'appliquer au client qui demande un prêt : c'est ce qu'on appelle une mesure de score de risque.

Distinction entre les méthodes supervisées : classement et prédiction

Le classement : variable cible catégorielle

Encore appelé « discrimination », le classement est une technique prédictive dont la variable cible est une variable catégorielle, le plus souvent booléenne.

Le classement permet de placer chaque individu dans une classe correspondant à une catégorie de la variable cible.

A noter que le classement est aussi le nom donné à une technique de modélisation descriptive, par opposition à la classification. Il s'agit bien du même « classement » dans le sens où on connaît a priori les catégories de classement. Quand il s'oppose à la classification, le classement est descriptif, sans variable cible. Quand il s'oppose à la prédiction, le classement est prédictif, avec variable cible.

L'exemple type sera le classement prédictif par arbre de décision.

La prédiction : variable cible continue

Encore appelé « régression », la prédiction est une technique prédictive dont la variable cible est une variable continue.

L'exemple type sera la prédiction par régression linéaire.

Exemple : le scoring

La banque est le principal utilisateur de mesure de score. Ces mesures utilisent les données économiques, sociales et géographiques du client, mais aussi les données sur le fonctionnement de ses comptes.

Principaux types de scores utilisés dans la banque (tous binaires) :

- **Score d'appétence** ou de propension à consommer. Pour savoir quel produit proposer à quel client.
- **Score de risque**, de comportement à risque. Pour accepter ou pas une demande de prêt, de découvert, de carte bancaire, etc.
- **Score d'octroi**. C'est la même chose qu'un score de risque, mais pour un nouveau client, donc sans historique du fonctionnement des comptes.
- **Score de recouvrement**. Evalue le montant susceptible d'être récupéré sur un compte en cas de contentieux.
- **Sore d'attrition**. Evalue la probabilité de quitter la banque.

Deux grands types de technique : inductive et transductive

Les techniques transductives

Elles ne présentent qu'une seule phase.

Elles ne produisent pas de modèle.

C'est pendant la classification des individus connus que se fait la prédiction des données inconnues. Toute prédiction demande donc un accès à la population complète (ou à un échantillon) et demande une grande puissance de calcul et peut donc être assez longue.

Les techniques inductives

1 : Elles présentent trois phases (parfois quatre) :

- une phase d'apprentissage qui permet d'élaborer un modèle. C'est la phase inductive.

- une phase de test pour vérifier le modèle obtenu (et éventuellement une phase de validation en plus).
- une phase de prédiction ou de classement qui consiste à appliquer le modèle à de nouvelles données. C'est la phase déductive.

Les phases d'apprentissage, de test et de validation sont effectuées sur des échantillons distincts de la population.

2 : Elles produisent un modèle.

Les techniques inductives sont plus répandues car le modèle produit permet un contrôle du modèle (courbe de ROC et indice de Gini) et une application facilitée : une prédiction se fait à partir du modèle, sans retour à la population ou à un échantillon d'origine. C'est rapide et demande peu de puissance de calcul.

Ce sont uniquement ces techniques qu'on va aborder dans ce cours.

Echantillons d'apprentissage et de test

Les techniques inductives travaillent sur deux échantillons de la population :

- L'échantillon d'apprentissage
- L'échantillon de test

L'échantillon d'apprentissage est celui avec lequel le modèle est construit.

L'échantillon de test est celui avec lequel le modèle est testé.

Ces échantillons doivent être représentatifs pour que garantir la qualité du modèle.

Concrètement, on prend une partie de la population de départ (les x premiers, 1 sur n, tel pourcentage aléatoire), puis on vérifie que les principales caractéristiques statistiques (tendance centrale, dispersion, corrélations) sont maintenues.

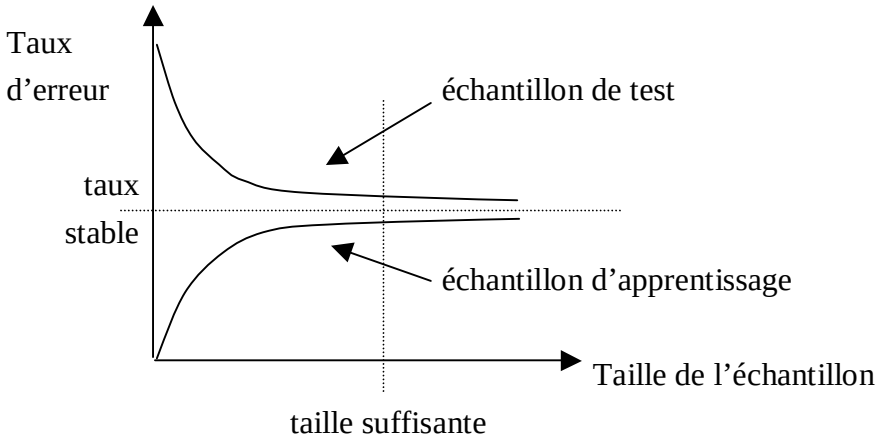
Si des exemples manquent systématiquement dans l'ensemble d'apprentissage concernant une catégorie particulière de données, la modélisation donnera de mauvais résultats.

Qualités attendues d'un modèle supervisé

- Le taux d'erreur doit être le plus bas possible (courbe de ROC et indice de Gini).
- Il doit être aussi peu sensible que possible aux fluctuations aléatoires de l'échantillon d'apprentissage.
- Il doit se maintenir le plus possible avec l'évolution dans le temps de la population. Cette caractéristique est fonction des domaines d'application : un score peut durer deux ans dans la banque et six mois en téléphonie mobile.
- Les règles doivent être aussi simples et aussi peu nombreuses que possibles.
- Elles doivent autant que possible être accessibles et compréhensibles.

Taille de l'échantillon d'apprentissage

Le schéma ci-dessous montre l'évolution du taux d'erreur dans les échantillons d'apprentissage et de test en fonction de la taille de ces échantillons (les deux échantillons aillant la même taille).



Le principe est que le taux d'erreur dans l'échantillon d'apprentissage croit avec le nombre d'éléments de l'échantillon jusqu'à stabilisation. En effet, si on a deux points, on peut faire une droite et le taux d'erreur est nul ; idem avec 3 points et une courbe ; avec 4 points et plus, on va commencer à avoir un taux d'erreur croissant.

Inversement, le taux d'erreur de l'échantillon de test décroît avec le nombre d'éléments de l'échantillon, jusqu'à stabilisation un peu au-dessus du taux d'erreur de la population d'apprentissage. En effet, si on a deux points dans l'échantillon d'apprentissage, les deux points de l'échantillon de test seront (probablement) très éloignés de la droite trouvée dans l'échantillon. La progression du modèle par augmentation de la taille de l'échantillon d'apprentissage verra donc une diminution du taux d'erreur dans l'échantillon de test.

Il y a donc une taille critique de l'échantillon d'apprentissage. Celle-ci dépend de la complexité du problème traité.

Il est recommandé de disposer de 300 à 500 individus dans chaque classe à prédire.

Sur-apprentissage

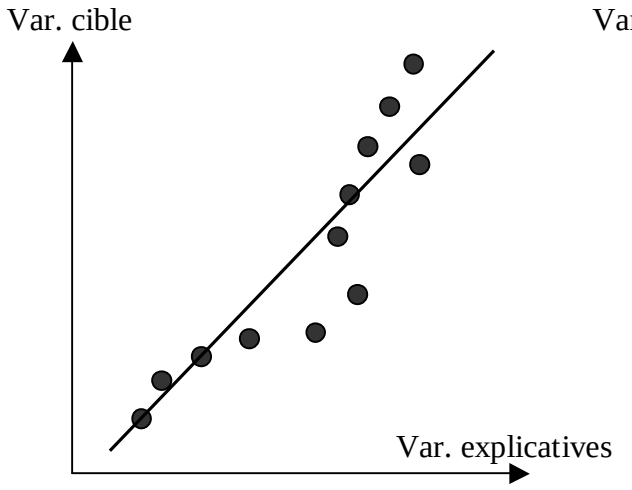
On parle aussi de sur-ajustement (overfitting ou overtraining).

Un modèle trop simple est tel le taux d'erreur sur les données d'apprentissage est élevé. De ce fait, le taux d'erreur sur les données de test et d'application sera aussi élevé.

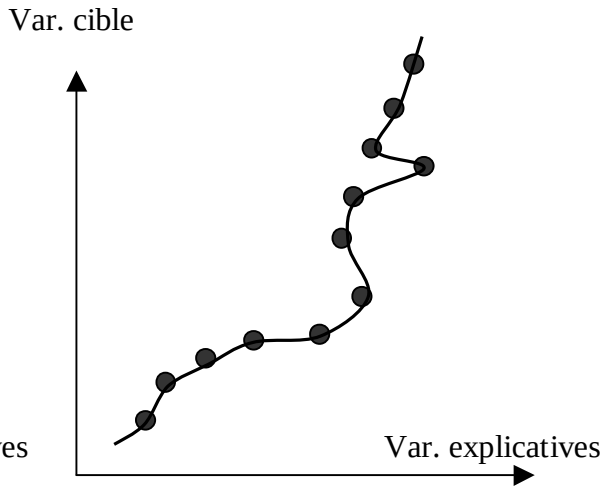
Un modèle trop complexe est tel que le taux d'erreur sur les données d'apprentissage est très faible. De ce fait aussi, le taux d'erreur sur les données de test et d'application sera très élevé. En effet, un modèle trop complexe, étant particulièrement bien adapté aux données d'apprentissage, se trouve être inadapté pour les données de test et d'application. En quelque sorte, il ne prend pas son compte de taux d'erreur global des populations d'apprentissage, de test et d'application, surchargeant du même coup le taux des populations de test et d'application.

On parle de sur-apprentissage quand une liaison entre la variable cible et les variables explicatives apparaît dans les données d'apprentissage alors qu'elle n'existe pas dans la population entière.

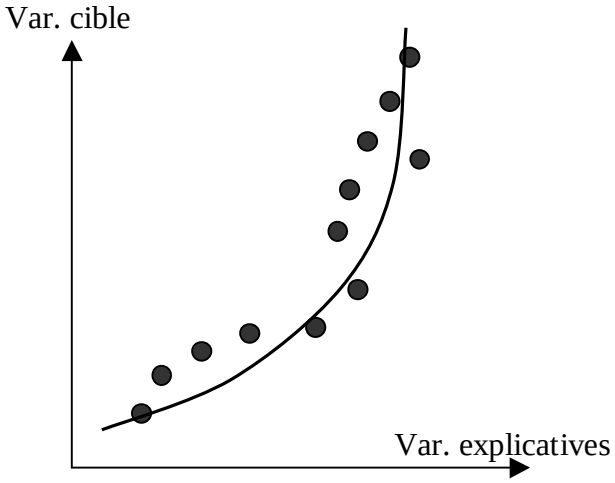
Le sur-apprentissage peut survenir lorsque l'une des variables cibles est mathématiquement corrélée à la variable cible.



Modèle trop simple

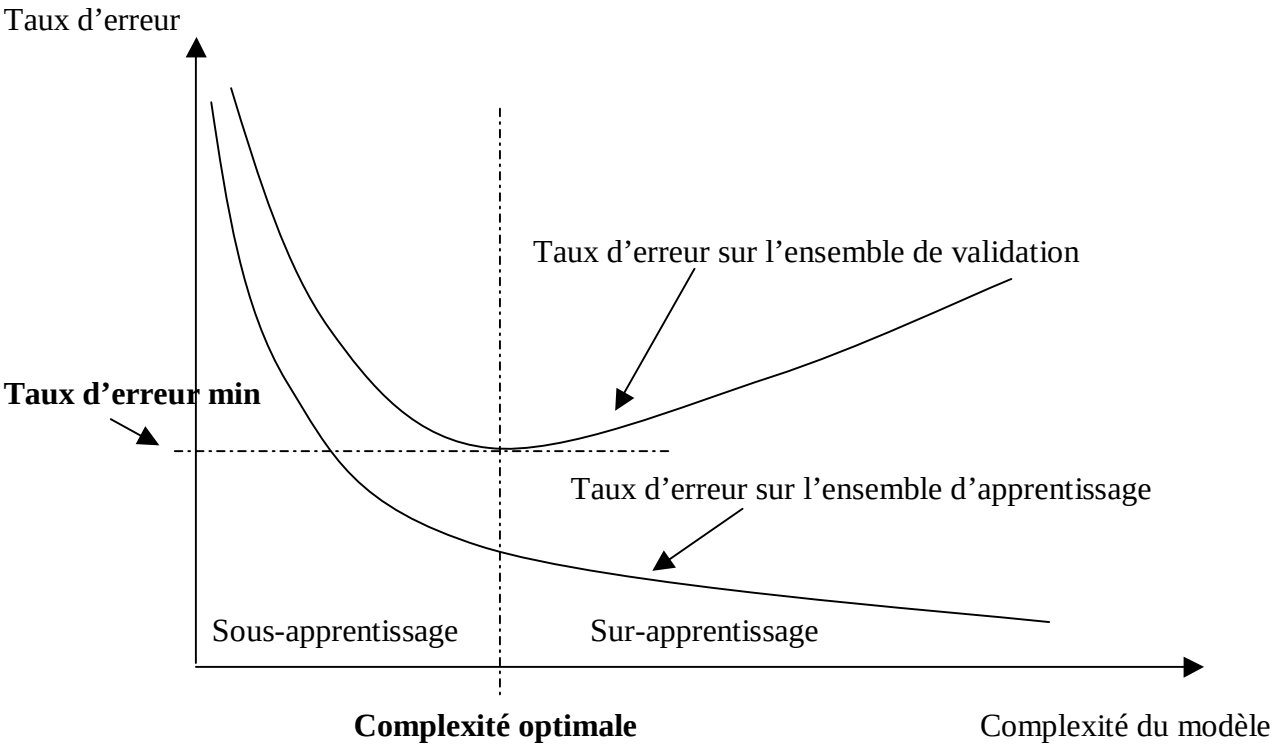


Modèle trop complexe



Bon modèle

Il s'agit de trouver un compromis entre la fiabilité du modèle sur l'ensemble d'apprentissage et la généralisation du modèle :



Le but est de trouver le juste milieu entre sous et sur –apprentissage.