

Modélisation des données

Quelques concepts généraux de modélisation de données

Tarik AL ANI, Département Informatique
ESIEE-Paris
E-mail : t.alani@esiee.fr
Url: <http://www.esiee.fr/~alanit>

Tarik AL ANI, A2SI-ESIEE – Paris

Notion de Modélisation de données

Il est nécessaire de construire un modèle à partir de données s'il n'existe pas des relations analytiques ou logique sophistiquées qui régies les comportements des populations étudiées.

Exemple : Pharmacologie, prévoir l'effet d'une nouvelle molécule. Il est difficile d'établir des équations de physico-chimie pour

- décrire les mécanismes d'action d'un médicament au niveau moléculaire,
- relier ces mécanismes à l'état de santé de l'être humain (qui nécessite la compréhension des mécanismes intimes du corps humain)

Pour ces raisons, la mise au point de nouvelles molécules actives ne peut se passer de collecte et de modélisation de données.

La **modélisation de données** : extraire des informations utiles d'un ensemble de données obtenues par des mesures, et de condenser cette information dans un modèle exploitable.

Objectif : basé sur ces données, l'objectif est d'aider à une prise de décision par rapport à un hypothèse concernant la population étudiée. : garder l'hypothèse qui est compatible avec les données.

- **Mesures** : représenter une grandeur (**individu**) d'un phénomène physique ou biomédical. Elles sont faite généralement sur un grand nombre d'individus de même nature (**population**)
- **Données** : plusieurs caractéristiques (**variables**) d'un phénomène physique ou biomédical sont mesurées → permet d'augmenter la quantité d'information disponibles sur ce phénomène .

- **Information** : Tableau de données

Exemple : Tension artérielle

Individu	Sexe M=0, F=1	Taille	Poids	Âge	Tension Artérielle T_{art}
1	1	1,68	49	48	14
2	0	1,79	72	23	13
3	0	1,67	69	65	19
4	1	1,53	95	61	22
5	0	1,82	85	35	15

- **Modélisation descriptive**

Chercher à extraire de tableaux de données souvent grands, parfois gigantesques, des informations compactes et interprétables :

- **Groupement** (ou **typologie**) « **clustering** » : partitionner la population en groupes homogènes et bien différenciés selon certains critères. Il s'agit d'une exploitation des redondances entre individus (les lignes d'un tableau de données).

Le groupement peut être considéré comme l'identification forcée des individus d'un même groupe (créé par la technique de groupement) à un seul individu représentatif du groupe (souvent, son barycentre).

- **Réduction de dimensionnalité** « **feature extraction** » de données concernant les populations étudiées tout en conservant les informations les plus pertinentes.

- **L'Estimation de Densité de Probabilité** : l'action de déterminer la densité des individus autour d'un individu donné. Le groupement peut être perçu comme une version simplifiée de l'Estimation de Densité de Probabilité.

- **Modélisation prédictive** :

Chercher à mettre en évidence des liens causales (**équations rapprochées** ou **règles logiques** ou un **modèle numérique** (déterministe ou stochastique) ou un **réseau de neurones**) entre les différentes variables (ou **primitives** : colonnes dans le tableau de données).

Par exemple, en statistique traditionnel : analyse univariée (moyenne, variance, histogramme, ...) ou analyse bivariée (par exemple : corrélation).

Notion de Modélisation - **Modélisation prédictive** :

Lorsque ces relations sont établies → 2 utilités:

1. Mettent en évidence une relation entre des grandeurs.

Si la nature causal de cette relation est confirmée :

modifier les causes → orienter les effets dans des directions favorables

2. Cette relation peut être traduite par une équation approchée $y = f(x_1, x_2, \dots, x_n)$, un **modèle** qui permet de prédire la valeur de y pour de nouveaux individus pour lesquels y n'aurait pas été mesuré (et qui ne figureraient donc pas dans le tableau de données initial : **généralisation**).

Exemple : prédire la tension artérielle d'un nouveau patient à partir de ses propre variables mesurées : sexe, taille, poids et âge.

Tarik AL ANI, A2SI-ESIEE – Paris

Notion de Modélisation - **Modélisation prédictive** :

Le même tableau de données peut générer plusieurs modèles selon les **objectifs** fixés ou le **problème** posé par l'**analyste** et les **choix techniques** mises en œuvre pour la modélisation.

Questions:

Poids = f_1 (Sexe) ?

Poids = f_2 (Taille) ?

Taille = f_3 (Sexe) ?

$T_{art} = f_4$ (Sexe, Âge, Poids) ?

$T_{art} = f_5$ (Taille, Âge, Poids) ?

Exploitation du modèle

La **Modélisation Prédictive** fournit

- la **prédiction** de grandeurs non mesurées sur de nouveaux individus ne figurant pas dans la base de données utilisée pour construire le modèle (appelé **base d'apprentissage** en Intelligence artificiel .
- l'**interprétation du modèle**, phase délicate qui demande les efforts conjugués de l'**analyste** et d'un **spécialiste** de la population étudiée.

Exploitation du modèle (suite)

Interprétation

Exemple,

$$T_{\text{art}} = f_4(\text{Sexe}, \hat{\text{Age}}, \text{Poids})$$

pourrait-elle être :

$$T_{\text{art}} = 0,03 * \text{Sexe} + 0,18 * \hat{\text{Age}} + 0,12 * \text{Poids}$$

- Pourquoi les paramètres ont-ils ces valeurs particulières ?
- Quelle information sur la population se trouve cachée dans ces valeurs ?

Interprétation (suite)

Tous les types de modèles ne sont pas susceptibles de recevoir le même niveau d'interprétation. Par exemple,

- *Régression Linéaire Simple* ou les *Arbres de Décision* sont très clairement interprétables,
- alors que les *Réseaux de Neurones* ne le sont pas du tout.

En général, un compromis entre "*interprétabilité*" et "*qualité*" d'un modèle doit être trouvé. Le choix d'un type de modèle pour étudier une population doit prendre en compte ce compromis.

Interprétation (suite) :

Interaction *Analyste/Specialiste*

Individus d'un même sexe et de même âge : Chaque kilo supplémentaire se traduit par une augmentation de T_{art} de 0,12

Analyste



Est-ce que l'excès de poids est, ou non, la cause de cette augmentation de tension?

Spécialiste

Deux types of **prédiction** :

- Lorsque la **variable** (ou **primitive**) à expliquer y est **numérique**, la modélisation prédictive s'appelle "**régression**".

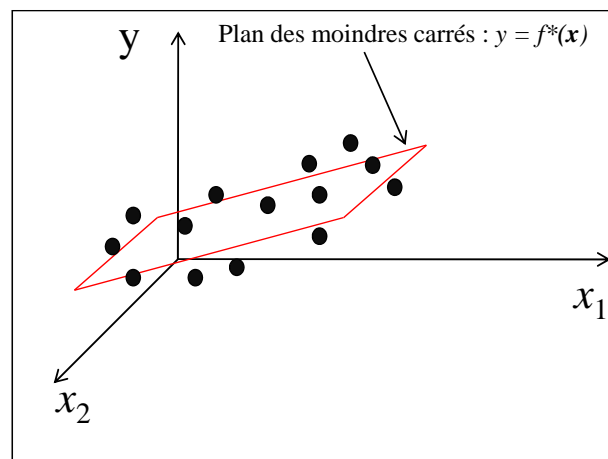
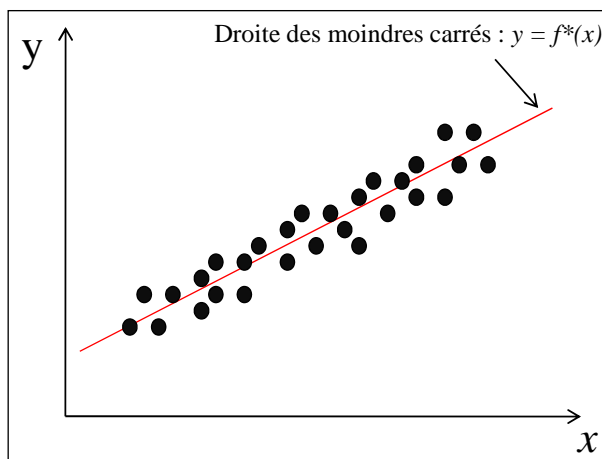
- Lorsque la variable à expliquer y est **nominale** (ou **qualitative**), la modélisation prédictive s'appelle "**classification**".

Tarik AL ANI, A2SI-ESIEE – Paris

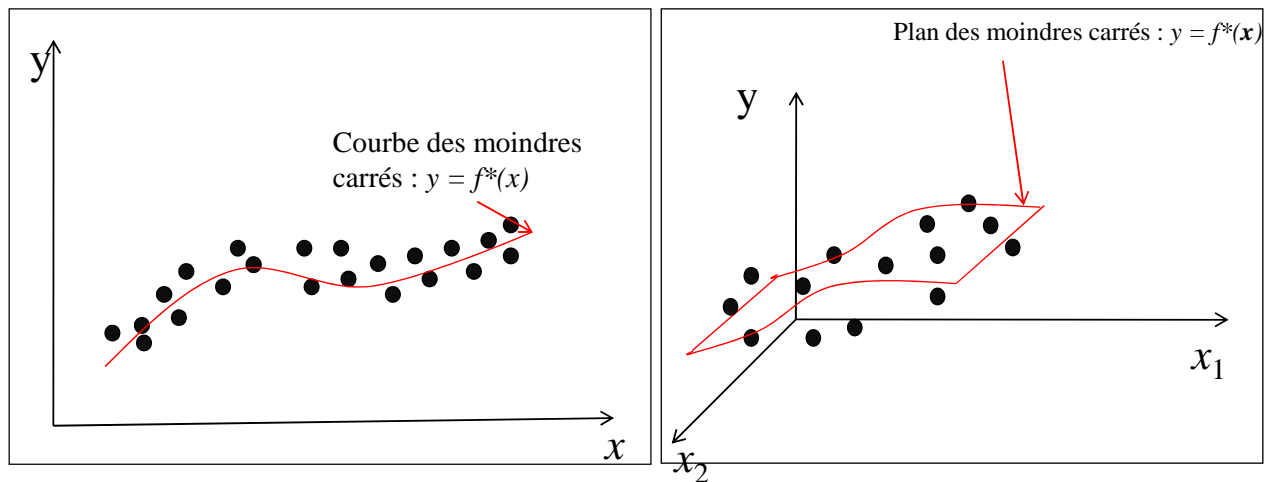
- **Régression linéaire simple** : $y = f^*(x) = ax + b + \varepsilon(x)$

- **Régression linéaire multiple** :

$$y = f^*(\mathbf{x}) = a_1x_{i1} + a_2x_{i2} + \dots + a_px_{ip} + b_i + \varepsilon(x)_i, i=1, 2, \dots, n$$



Régression – fonction générale



Généralement, la "**variable à prédire**" (y) ne dépend pas uniquement d'une valeur (x). La fonction (exacte) de régression $f(x)$ donne une réponse unique pour toute valeur de x , sa prédiction est donc presque certainement entachée d'erreur.

D'autres variables "**variables explicatives**", ou "**prédicteurs**" « **features** », pourraient être prises en compte dans le but de réduire les erreurs sur la prédiction des valeurs de y .

Ainsi, d'une façon générale, la **fonction de régression** (le **modèle de régression**) est la "meilleure" fonction :

$$y = f(\mathbf{x}) = f(x_1, x_2, \dots, x_p)$$

permettant de prédire la valeur de la variable à prédire y , connaissant les valeurs des **prédicteurs** $\{x_1, x_2, \dots, x_p\}$.

Il faut donc définir $f(\mathbf{x})$ de façon à **minimiser** ces erreurs.

Le terme "régression" repose sur la définition suivante :

La fonction de régression $f(\cdot)$ est celle qui, pour tout jeu de valeurs des **prédicteurs** $\{x_1, x_2, \dots, x_p\}$ produit la **valeur moyenne** de y pour ce jeu de valeurs :

$$f(x_1, x_2, \dots, x_p) = E[y | x_1, x_2, \dots, x_p]$$

Difficultés de la Régression

La **fonction de régression** $f(\cdot)$ peut avoir

- une forme analytique quelconque inconnue
- pas de forme analytique.

1) **Modèles de régression paramétriques:**

Une **fonction de régression** $f(\cdot)$ peut être approximée par une fonction $f^*(\cdot)$ connue. L'analyste devra donc faire le choix de la forme de $f^*(\cdot)$ destinée à approximer $f(\cdot)$ en fonction de ses connaissances a priori des données.

Régression Linéaire (Simple ou Multiple)

Hypothèse linéaire: $f(\cdot)$ (ainsi que $f^*(\cdot)$) est une fonction linéaire des **prédicteurs**.

Hypothèse non linéaire : $f(\cdot)$ (ainsi que $f^*(\cdot)$) est une fonction **non linéaires** : la radioactivité, exponentielle décroissante.

Difficultés de la Régression (suite)

2) La **fonction de régression** $f(\cdot)$ est inconnue

Modèles de régression non paramétriques:

$f(\cdot)$ est inconnue, dans ce cas on fera appel à des modèles de régression plus généraux, dits "**non paramétriques**", comme les fonctions **splines** ou les **Réseaux de Neurones**.

Difficultés de la Régression (suite)

3) La régression suppose donc que les données ont été générées par une densité de probabilité :

$$y = f(\mathbf{x}) + \varepsilon_x$$

où :

$f(\mathbf{x})$: fonction déterministe, et

ε_x : variable aléatoire de moyenne nulle, mais de distribution inconnue, et dépendant possiblement de \mathbf{x} , c-à-d que la variance est la même dans tout l'espace des prédicteurs (**hétéroscédasticité**).

La régression doit également faire face, comme toute modélisation, à la très importante question du **choix des variables** à incorporer dans le modèle.

Ajouter des variables augmente :

- **la quantité d'information** disponible pour prédire les données (dans le cas d'une modélisation prédictive),
- le **nombre de paramètres du modèle**, et donc sa souplesse, ce qui lui permet de mieux rendre compte des données disponibles.

Prix à payer : plus grande instabilité du modèle, et donc une moins grande crédibilité de ses verdicts sur des données nouvelles.

Calcul des paramètres du modèle dans le cas de *Modèles de régression paramétriques linéaires* :

Le choix de la forme fonctionnelle $f^*(.)$ du modèle de régression étant fait, il faut alors calculer les valeurs de paramètres de cette fonction de façon à ce qu'elle approche au mieux la vraie fonction de régression $f(.)$.

Le technique principale utilisée dans ce cas est la méthode des **Moindres Carrés**, qui minimise la somme des carrés des erreurs de prédiction du modèle sur les données disponibles. La méthode des Moindres Carrés est celle adoptée par les **Régressions Linéaires Simple** et **Multiple**.

Validation du modèle

Le modèle $f^*(.)$ une estimation de la vraie fonction de régression $f(.)$. Ce modèle sera construit à partir de données aléatoires en nombre fini (**données de construction du modèle**, en IA appelées **données d'apprentissage**).

Comme pour tout modèle, il sera donc essentiel d'estimer ses performances réelles (c'est à dire sur des données n'ayant pas contribué à son élaboration (**données de généralisation**, en IA appelées **données de test**)).

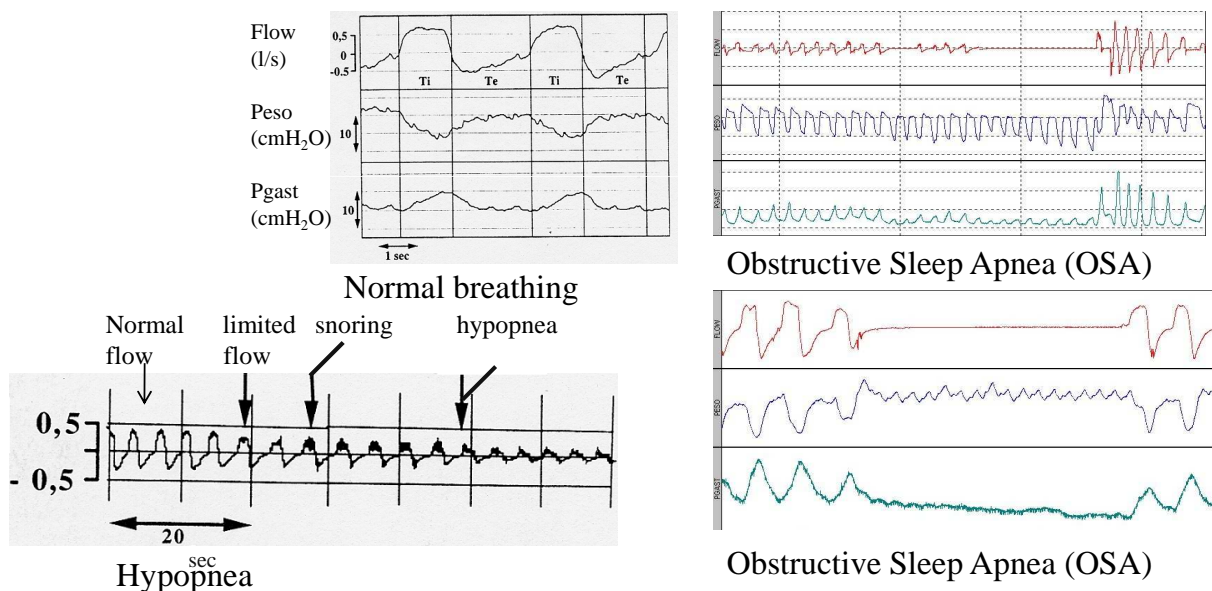
La Régression Linéaire est un des rares cas où, moyennant certaines hypothèses raisonnables sur le mécanisme ayant généré les données, il est possible d'estimer par le calcul les performances du modèle.

Dans le cas général, il conviendra :

- de construire plusieurs modèles qui diffèrent en général par le choix de $f^*(.)$,
- de soumettre ces modèles à des épreuves de **validation**,
- et de retenir, parmi les modèles candidats, celui ayant les meilleures performances estimées.

Classification

Les observations sont souvent regroupées naturellement en groupes, ou **classes**. Par exemple : Une famille de symptômes est associée à une pathologie, et une autre famille de symptômes à une autre pathologie.

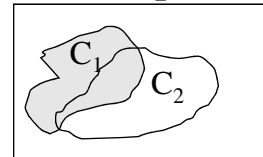


Le **classificateur** (modèle prédictif) est un jeu d'équations ou de règles logiques ou numériques construit à partir des données déjà étiquetées disponibles dans la base de données (**ensemble d'apprentissage**).

Il a deux rôles :

1. Accomplir la prédiction. A cette fin, le **classificateur** considère que le jeu d'attributs retenus (**prédicteurs**) contient implicitement l'information nécessaire pour reconstruire la variable à prédire (**étiquette** de classe).
2. Mettre en évidence certaines structures des données, et plus particulièrement les prédicteurs qui sont les plus importants pour une bonne qualité de prédiction.

La Classification rencontre au moins deux difficultés qui lui sont propres :

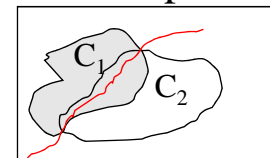


1. En général, les classes se chevauchent.

La Classification est donc une activité **probabiliste** par nature. Les nombres $P(C_i/x)$ sont appelées les **probabilités a posteriori** des différentes classes pour l'observation x .

2. Les classes peuvent avoir des formes arbitrairement complexes dans l'**espace des prédicteurs (feature space)**.

La forme de la frontière entre deux régions de décision est déterminée par :



- Les données disponibles lors de la construction du **classificateur**, et
- le choix de l'algorithme utilisé pour construire le **classificateur** à partir de l'ensemble d'apprentissage.

Décision Bayésienne

Minimiser le nombre d'erreurs d'affectation sur de nouvelles observations (donc non étiquetées).

La *théorie de la Décision Bayésienne* : la stratégie qui consiste à affecter une nouvelle observation à la classe ayant la plus grande *probabilité a posteriori* $p(C_i/x)$ est **optimale**, c'est à dire génère un plus petit nombre d'erreurs que toute autre stratégie.

Une erreur d'affectation est toujours coûteuse, mais toutes les erreurs n'ont pas le même coût : certaines peuvent être relativement inoffensives, alors que d'autres peuvent avoir des conséquences dramatiques.

La *Décision Bayésienne* (suite)

Exemple : diagnostic médical [1]

classificateur chargé de trier des examens de radios pulmonaires douteuses en "**Pas de cancer**" et "**Cancer**". Nous pouvons considérer deux types d'erreurs de classification :

- * L'erreur qui consiste à considérer comme alarmante une radio parfaitement normale.
- * et l'erreur qui consiste à considérer comme saine la radio d'un poumon cancéreux.

Le premier type d'erreur est gênant, le second est dramatique. Il est clair que l'on chercherait à biaiser les calculs du *classificateur* de façon à ce qu'il réduise encore plus le nombre d'erreur du deuxième type, quitte à ce qu'il commette plus d'erreurs du premier type.

La **Décision Bayésienne** (suite)

L'objectif de l'analyste est alors non pas de minimiser le nombre d'erreurs d'affectation, mais **de minimiser le coût moyen** de ces erreurs. La **théorie de la Décision Bayésienne** résoud également cette question.

Estimer les **probabilités a posteriori**

Il y a deux grandes approches de l'estimation des **probabilités a posteriori** :

- **Méthodes indirectes**
- **Méthodes directes**

La **Décision Bayésienne - Estimer les probabilités a posteriori** (suite)

Méthodes indirectes

Le **théorème de Bayes**

En théorie, le cadre bayésien résoud complètement le problème de la Classification, si

- les **probabilités a priori** $P(C_i)$ puissent être calculées ou estimées.
- Les **densités de probabilité** à l'intérieur de chacune des classes (**densités de probabilité conditionnelles (Vraisemblance)** « **likelihood** ») $P(x/C_i)$ peuvent être estimées.
- La **densités de probabilité inconditionnelles (densité de probabilité)** $P(x)$.

La **Décision Bayésienne - Estimer les probabilités a posteriori**
Méthodes directes - le **théorème de Bayes** (suite)

La formule générale de Bayes :

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{\sum_j P(x|C_j)P(C_j)}$$

Cas x est continue

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$$

La **Décision Bayésienne - Estimer les probabilités a posteriori**
(suite)

Méthodes directes

Plusieurs techniques évitent la phase d'estimation des densités de probabilités conditionnelles, et estiment **directement** les **probabilités a posteriori** en effectuant une sorte de régression sur les indicatrices de classe. C'est par exemple le cas des **Réseaux de Neurones** [2].

La Décision Bayésienne - *Estimer les probabilités a posteriori* (suite)

Méthodes *stochastiques* basées sur *l'apprentissage*

Plusieurs techniques utilisent une estimation des densités de probabilités conditionnelles, et construisent un modèle stochastique en utilisant des données d'apprentissage issues de processus à modéliser et estiment **directement** les probabilités *a posteriori* en maximisant la vraisemblance des données de test par rapport au modèle. Elles effectuent ensuite une sorte de régression sur les indicatrices de classe. C'est par exemple le cas des *Modèles de Markov Cachés (Hidden Markov Models (HMMs))* [2-3].

Modélisation de données et statistiques

Un enregistrement d'une activité cardiaque sur 24h ou un enregistrement d'une activité cérébrale sur 8h sont des exemples de processus stochastiques en un sens que ces enregistrements, s'ils sont renouvelés plusieurs fois sur les mêmes périodes et dans les mêmes conditions, les grandeurs obtenues à partir d'eux ne donnent jamais les mêmes valeurs ou les mêmes paramètres recherchés.

Question : Un modèle construit à partir d'un nombre fini d'enregistrements (un échantillon) issus d'un processus stochastique (ou une population), est-il un modèle crédible ?

La *Statistique* avec ses deux branches principales, *l'Estimation* et *l'Inférence* (ou *Théorie de Tests*) permet, moyennant certaines hypothèses (sous forme de répartition d'une certaine grandeur au sein de la population), de juger de la crédibilité du modèle (construit sur un échantillon) en tant que représentants des propriétés de la population.

Références

[1] Livre interactif : <http://www.aiaccess.net/>

[2] Cours Intelligence artificielle : <http://www.esiee.fr/~alanit>

[3] Tutorial HMMs : L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition", Proc. IEEE, Vol. 77, No. 2, February 1989, pp. 267-296.
<http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf>