

Reconnaissance des formes et méthodes neuronales (RCP208)

Introduction

Michel Crucianu
(prenom.nom@cnam.fr)
<http://cedric.cnam.fr/vertigo/Cours/ml/>

Département Informatique
Conservatoire National des Arts & Métiers, Paris, France

28 septembre 2017

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

Le sujet

« *Essentially, all models are wrong, but some are useful.* » (Box, G. E. P., and Draper, N. R. Empirical Model

Building and Response Surfaces, John Wiley & Sons, New York, NY, 1987, p. 424.)

- Reconnaissance des formes (*pattern recognition*) : (sens strict) identifier à quelle catégorie appartient une « forme » décrite par des données brutes
 - « Forme » = observation (ou partie d'observation, ou ensemble d'observations)
 - Exemples : valeurs de variables décrivant un état clinique, partie correspondant à un visage dans une image, ensemble des valeurs prises par le cours d'une action sur une journée
 - Fouille de données (*data mining*) : recherche de régularités ou de relations inconnues *a priori* dans de grands volumes de données
- ⇒ Présenter les éléments de base des méthodes d'analyse et de modélisation des données
- Dans cette UE on ne s'intéresse pas aux spécificités des données massives

Contenu de l'enseignement : problématique et pré-requis

Problématique abordée

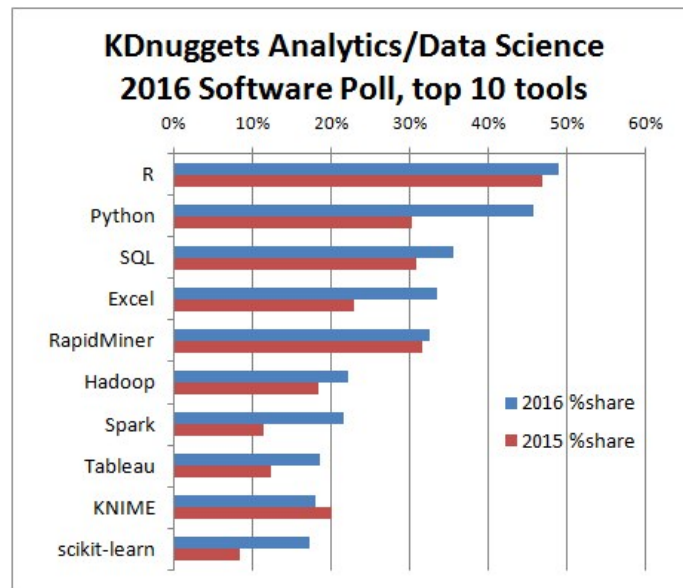
- Examiner et comprendre les (caractéristiques des) données → analyses factorielles, visualisation
 - Comprendre les problèmes posés par les données → solutions à ces problèmes
 - Description des données → classification automatique, estimation de densités, quantification vectorielle
 - Construction de modèles décisionnels (ou prédictifs) → perceptrons multi-couches (RCP208), autres méthodes (dont forêts d'arbres de décision, SVM, réseaux de neurones profonds - RCP209)
- ⇒ Capacité à mettre œuvre des méthodes d'analyse des données, de reconnaissance des formes et de fouille de données

Prérequis

- En mathématiques : connaissances mathématiques de base
- En informatique : connaissances de base en programmation

Contenu de l'enseignement : travaux pratiques

- Outil libre et ouvert développé par une large communauté → <http://scikit-learn.org> : langage Python, visualisation intégrée



- Python permet d'utiliser aussi Apache Spark (pour données massives, voir RCP216)

Contenu détaillé de l'UE RCP208

- 1 Introduction : exemples, nature des problèmes de modélisation, types de variables et leur représentation, nature des difficultés posées par les données. TP : introduction à Python et scikit-learn, installation assistée. (2 séances)
- 2 Méthodes factorielles (ACP, ACM, AFD) pour l'analyse des données et la réduction de dimension. TP : ACP, AFD. (2 séances)
- 3 Classification automatique : k-moyennes (avec initialisation k-means++), *spectral clustering*. TP : k-means, *spectral clustering*. (2 séances)
- 4 Estimation de densités : histogrammes (non paramétrique), modèles de mélanges, espérance-maximisation. TP : estimation de modèles de mélanges. (2 séances)
- 5 Imputation des données manquantes : spectre des méthodes. TP : imputation par SVD, imputation par centres des groupes. (1 séance)
- 6 Quantification vectorielle et visualisation : cartes auto-organisatrices (de Kohonen). TP : cartes de Kohonen. (3 séances)
- 7 Perceptrons multi-couches, rétro-propagation, fonctions d'erreur, régularisation. TP : perceptron multi-couches en discrimination ou régression. (3 séances)

Contenu détaillé de l'UE RCP209

- 1 Apprentissage supervisé : classement, régression, prédiction structurée ; généralisation, biais et variance, bornes de généralisation. Évaluation et sélection de modèles : validation croisée, *grid search*. TP : *grid search*, validation croisée (*k-fold*, *leave one out*). TP : discrimination linéaire avec perceptrons ou AFD, régression linéaire, validation croisée. (2 séances, Michel Crucianu)
- 2 Arbres de décision et forêts d'arbres de décision. TP : forêts d'arbres de décision. (2 séances, Marin Ferecatu)
- 3 SVM : maximisation de la marge, *kernel trick*, classement, régression, estimation du support d'une distribution, ingénierie des noyaux. TP : classement, régression. (3 séances, Marin Ferecatu)
- 4 *Deep learning* avec ConvNets profonds, réseaux de neurones récurrents. TP : *deep learning* avec Theano. (8 séances, Nicolas Thome)

Références bibliographiques RCP208

- J.-P. Asselin de Beauville, F. Kettaf, Bases théoriques pour l'apprentissage et la décision en reconnaissance des formes. Éditions Cépaduès, 2005.
- A. Belaid, Y. Belaid, Reconnaissance des formes : méthodes et applications. Inter Editions.
- M. Crucianu, J.-P. Asselin de Beauville, R. Boné, Méthodes factorielles pour l'analyse des données : méthodes linéaires et extensions non-linéaires. Hermès, 2004, 288 p.
- G. Dreyfus, J. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria, Apprentissage statistique : Réseaux de neurones - Cartes topologiques - Machines à vecteurs supports. Éditions Eyrolles, 2008.
- G. Saporta, Probabilités, analyse des données et statistique. 622 p., Éditions Technip, Paris, 2006.
- Un site qui rassemble de nombreux liens utiles, bien classés :
http://www.xavierdupre.fr/app/ensae_teaching_cs/helpsphinx3/td_2a.html

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

Organisation

- Enseignants : Michel Crucianu, Meziane Yacoub
- Semestre 1 : hors temps ouvrable (HTO)
 - Cours : jeudi 17h30-19h30
 - Travaux pratiques (TP) : jeudi 19h45-21h45
- Semestre 2 : formation à distance (FOD)
- Supports en accès ouvert mis en ligne progressivement :
<http://cedric.cnam.fr/vertigo/Cours/ml/>
 - Cours : transparents (PDF) et (plus tard) explications en HTML
 - TP : contenu détaillé en HTML

Evaluation

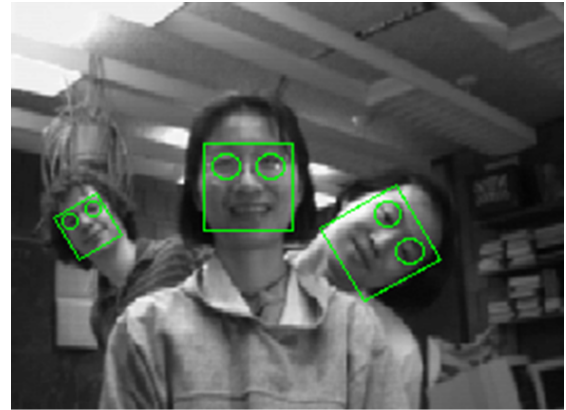
- S1 (HTO) : examen en février avec rattrapage en avril
- S2 (FOD) : examen en juin avec rattrapage en septembre
- Planification des examens : <http://www.cnam-paris.fr/suivre-ma-scolarite/examens/>
ou <http://emploi-du-temps.cnam.fr/emploidutemps2?FD=1>
- Comptes-rendus de TP : réponses aux questions posées dans le support de TP
- Note finale = note d'examen modulée (en + ou en -) par les comptes-rendus des TP

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation**
- 5 Décision et généralisation
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

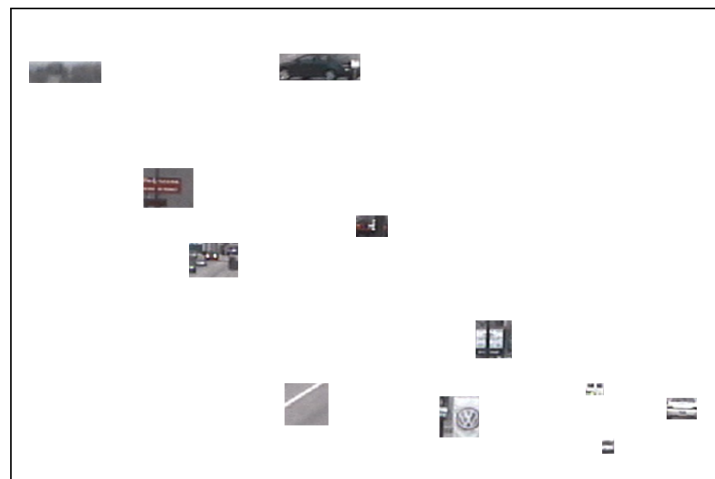
Exemple : détection d'objets

- Objectif : détecter des « objets » d'une ou plusieurs catégories dans des images ou des vidéos



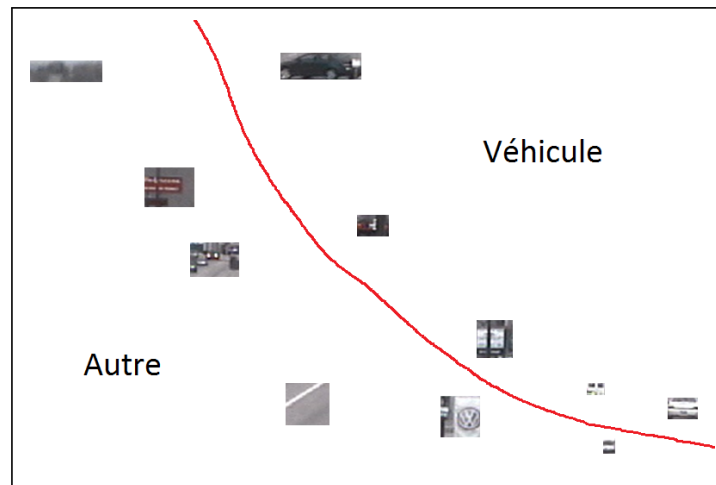
Exemple : détection d'objets (2)

1. Déterminer quelles données devraient permettre de discriminer la classe d'intérêt du « reste du monde »
2. Récolter, examiner, nettoyer les données
3. Représenter les données de façon adéquate



Exemple : détection d'objets (3)

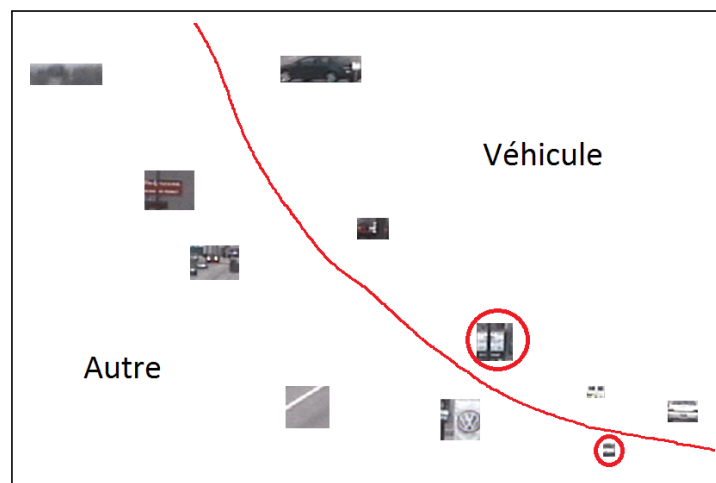
4. Modéliser, à partir des données, la frontière de discrimination entre objets de la classe d'intérêt et autres régions d'images



- Frontière linéaire ou non linéaire, déterminée sur la base de différents critères, directement sur la représentation d'entrée ou en apprenant de nouvelles représentations
- Note : il est également possible de décider sans construire un modèle, voir par ex. la méthode des k plus proches voisins

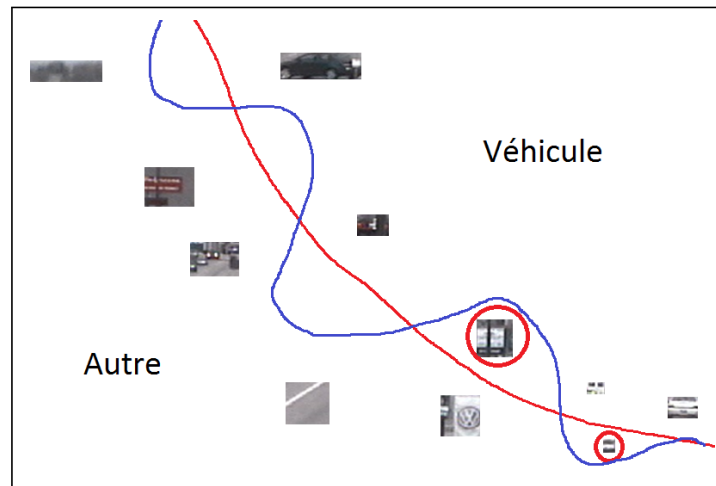
Exemple : détection d'objets (4)

- Un modèle est obtenu sur des données **d'apprentissage**
- Approche : choix d'une famille paramétrique (par ex. frontières linéaires) et optimisation de paramètres afin de minimiser une fonction de coût (erreur d'apprentissage)



Exemple : détection d'objets (5)

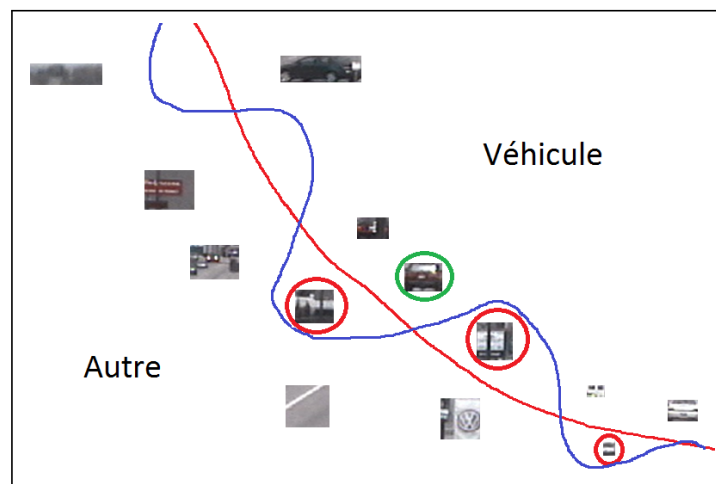
- Avec une autre famille paramétrique peut-on réduire encore l'erreur d'apprentissage ?



- Avec quelles conséquences ?

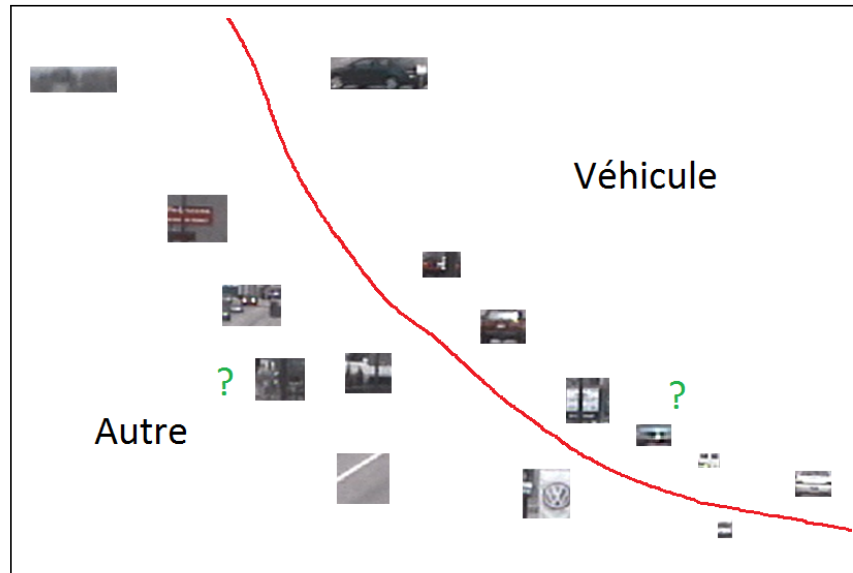
Exemple : détection d'objets (6)

- Les performances d'un modèle ne doivent pas être évaluées sur des données d'apprentissage car cela produirait des estimations trop optimistes !
- L'évaluation se fait sur des données de **test**, non utilisées pour l'apprentissage



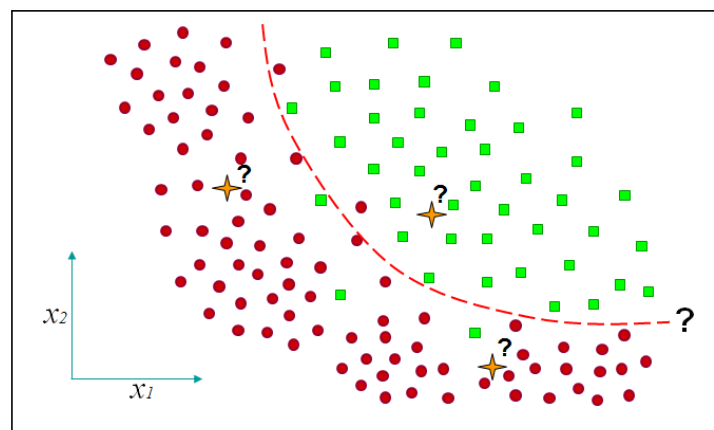
Exemple : détection d'objets (7)

5. Avec le modèle retenu, décider, pour une région d'une nouvelle image, si elle contient ou non un objet d'intérêt



Simplification de l'exemple

- Les observations sont souvent caractérisées par de nombreuses variables de types variés (que nous examinerons plus tard)
- Pour illustrer nos propos, nous considérerons seulement 2 variables, X_1 et X_2
- Une observation i est donc décrite par (x_{i1}, x_{i2})

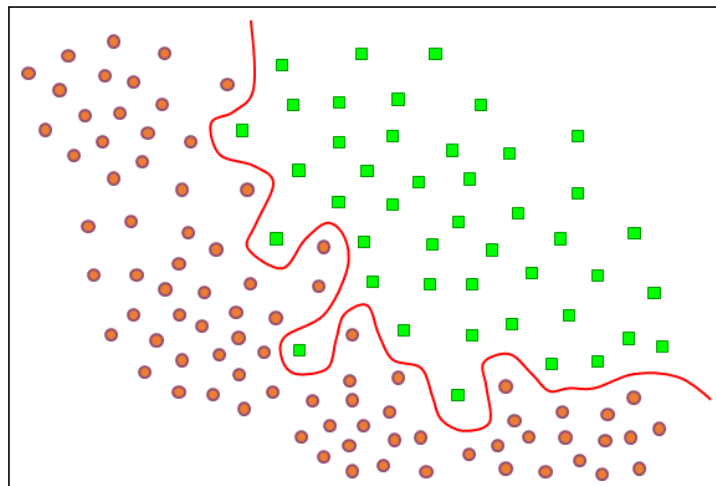


Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation**
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

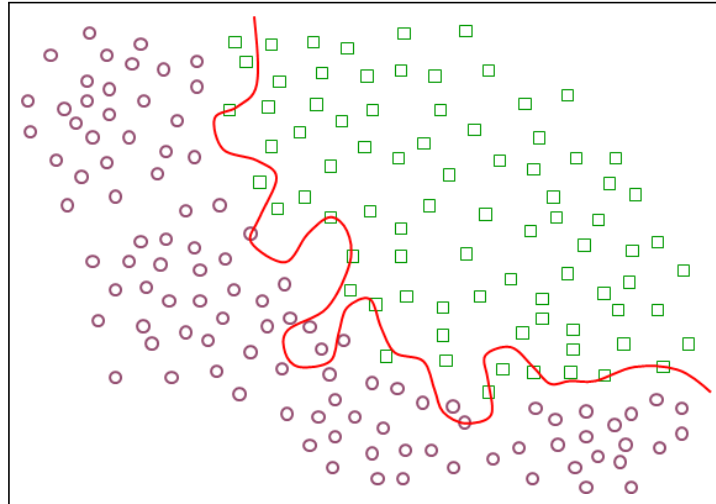
Construction de modèle décisionnel

- Un modèle décisionnel est obtenu à partir de données d'**apprentissage**, pour lesquelles la classe d'appartenance est connue
- L'erreur résiduelle du modèle sur ces données est appelée erreur d'apprentissage (ou *risque empirique*)



Décision avec un modèle

- L'objectif est d'utiliser le modèle pour **décider** à quelle classe affecter de **futures** (nouvelles) données
- L'erreur sur les données futures est l'erreur de **généralisation** (*risque espéré*)



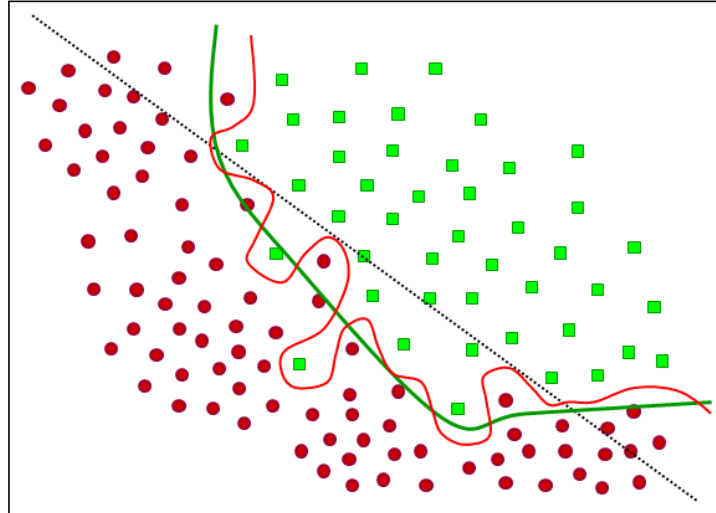
⇒ on souhaite avoir **la meilleure généralisation**, c'est à dire la plus faible erreur de généralisation (et pas nécessairement la plus faible erreur d'apprentissage) !

Quel modèle présente la meilleure généralisation ?

- L'erreur d'apprentissage peut être (facilement) mesurée
- L'erreur de généralisation ne peut pas être mesurée, comment l'**estimer** ?
 - A partir de l'erreur sur des données de **test**, non utilisées pour l'apprentissage
 - Grâce à une éventuelle **borne supérieure** sur l'écart entre erreur d'apprentissage et erreur de généralisation
- Hypothèse importante : la distribution des données d'apprentissage est **représentative** de celle des données futures !
 - Or, la distribution évolue en général dans le temps (n'est pas stationnaire) ⇒ nécessaire d'adapter régulièrement le modèle

Quel modèle présente la meilleure généralisation ?

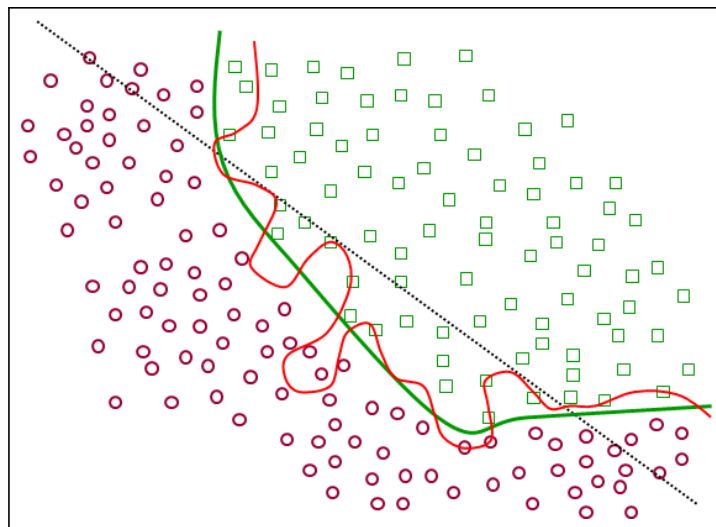
- Obtenir un modèle : choix d'une famille paramétrique, puis optimisation des paramètres pour minimiser l'erreur d'apprentissage
- Plusieurs familles → plusieurs modèles (un par famille), lequel préférer ?



- Sur les données d'apprentissage, certaines familles permettent de trouver un modèle dont l'erreur d'apprentissage est 0

Quel modèle présente la meilleure généralisation ?

- Et l'erreur de généralisation ?



- Explication

- 1 Erreur généralisation \leq erreur apprentissage + borne
- 2 Or, complexité famille paramétrique $\nearrow \Rightarrow$ borne $\nearrow \dots$

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation
- 6 Problèmes posés par les données**
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

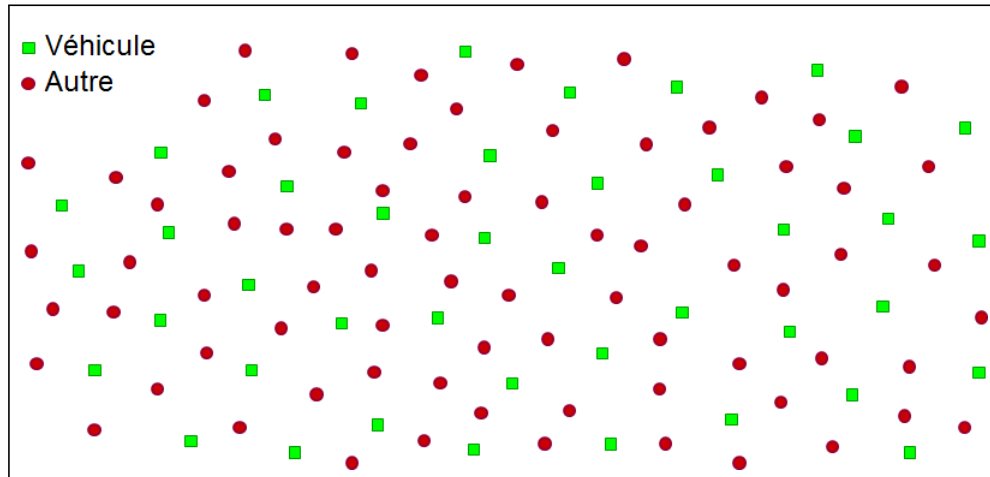
La réalité des données

Quelles caractéristiques des données posent des difficultés particulières à la modélisation ?

- 1 Données inadaptées
- 2 Données non représentatives
- 3 Données aberrantes
- 4 Données manquantes
- 5 Classes déséquilibrées
- 6 Redondance des variables
- 7 Dimension très élevée des données
- 8 ...

Données inadaptées

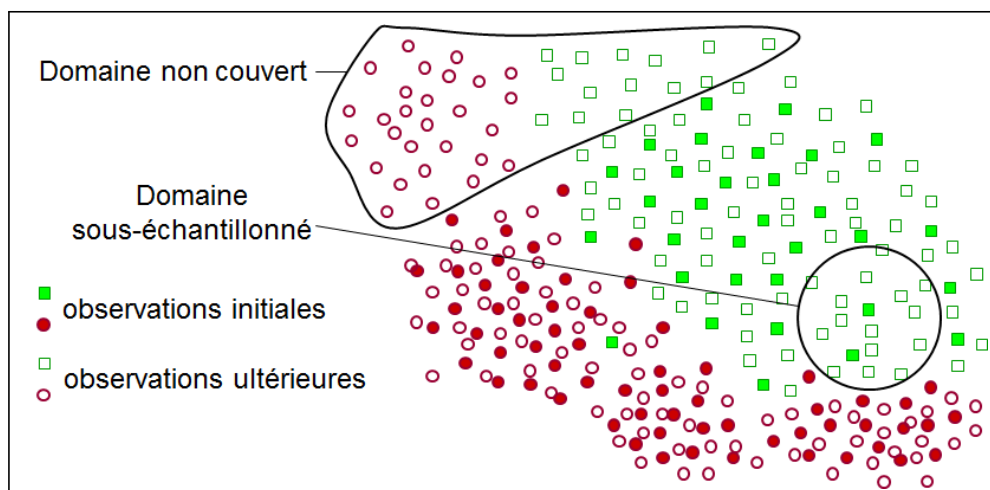
- Dans la situation illustrée, faut-il chercher à tout prix un modèle quelle que soit sa complexité ?
- Un tel modèle généralisera vraisemblablement très mal...



- ⇒ on conclut que l'information utile est absente des données
- ⇒ refaire une collecte de données après une analyse du problème

Données non (ou très partiellement) représentatives

- Collecte de données mal menée
 - 1 Données sous-représentées : par ex., collecte uniquement le dimanche quand la circulation des camions est interdite (→ images de camions sous-échantillonnées)
 - 2 Domaines pratiquement absents : par ex., types de véhicules non présents dans le pays de collecte, ou collecte uniquement en ville mais modèle utilisé partout



Données non (ou très partiellement) représentatives (2)

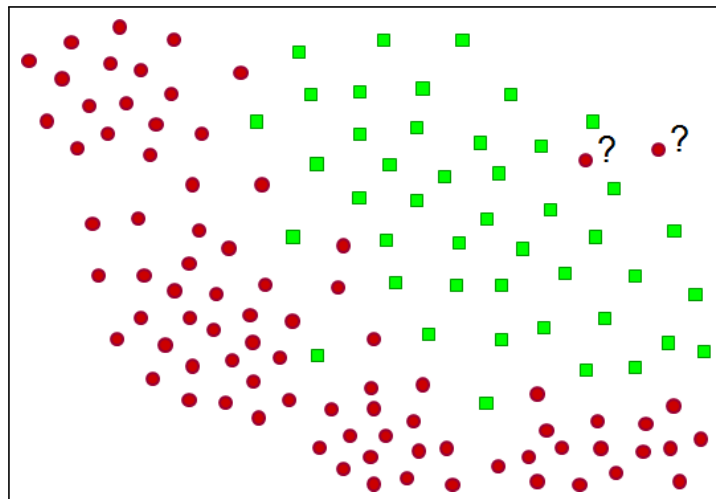
⇒ Comprendre comment la collecte a été réalisée

- Que faire ?

- Être prescripteur pour une **nouvelle collecte** de données afin de compléter les données là où elles sont absentes (ou insuffisantes)
- Restreindre le modèle aux régions avec un bon échantillonnage et pouvoir caractériser les limites du modèle (par ex., rejet de **non représentativité** : refus de décider pour des observations trop éloignées du domaine couvert par les données d'apprentissage)

Données aberrantes

- Données éloignées de toutes les autres ou seulement de celles de leur classe (comme dans l'illustration)



⇒ erreurs d'enregistrement, bruit ou phénomène significatif ?

Données aberrantes (2)

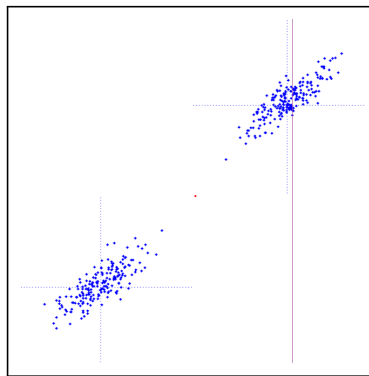
- Ignorer les données aberrantes
 - De façon implicite : méthodes robustes (par ex. modèles basés sur des distributions à décroissance lente)
 - De façon explicite : détection suivie de suppression
- Détecter les données aberrantes
 - Tests statistiques
 - Distance aux k plus proches voisins
 - Modélisation du support de la distribution des données « normales »
- Comprendre les données aberrantes
 - Comment les valeurs aberrantes ont été obtenues
 - ⇒ possibilités de correction des valeurs aberrantes ?
 - ⇒ ces valeurs **ne sont pas** aberrantes ?
 - Erreurs de mesure, par ex. appareils de mesure défectueux ou fonctionnant parfois dans des conditions extrêmes
 - Erreurs d'étiquetage, par ex. un attelage est-il considéré comme un véhicule ?

Données manquantes

- Pour certaines observations, les valeurs de certaines variables manquent
 - Ex. sondages : absence de réponse à certaines questions
 - Ex. véhicules : un capteur de vitesse fonctionne de façon intermittente
- Solution simple : suppression d'observations ou/et variables ⇒ si la proportion de données manquantes est élevée, on peut compromettre la capacité à construire un modèle
 - Supprimer les observations qui présentent des valeurs manquantes pour certaines variables
 - Supprimer les variables qui présentent une proportion importante de valeurs manquantes

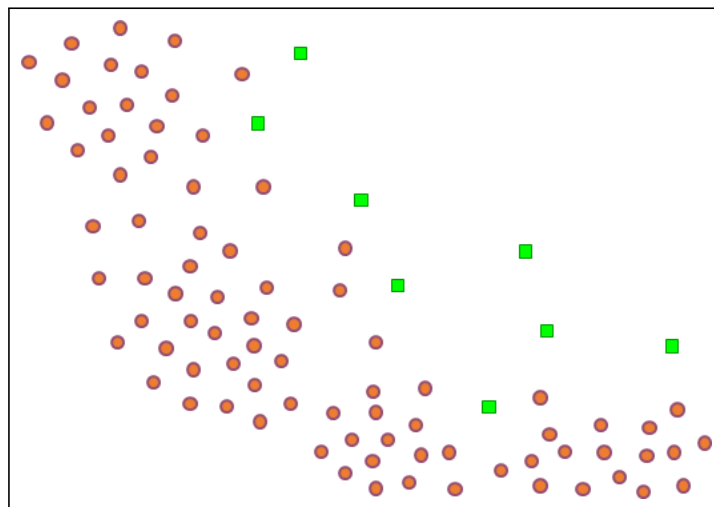
Données manquantes (2)

- Imputation des valeurs manquantes : estimer les valeurs manquantes, employer l'estimation comme une donnée mesurée
 - Estimer par la moyenne (ou la médiane) de la variable sur toutes les observations
 - Estimer par la moyenne de la variable pour **les observations de la même classe**
 - Estimer par le centre du groupe (*cluster*) issu d'une classification automatique
 - Estimation (lors d'itérations successives) du modèle **et** des données manquantes (par ex. algorithmes de type espérance-maximisation)
- Exemple : quelle valeur estimer pour la coordonnée verticale (**manquante**) de la donnée représentée par une barre verticale ?



Classes déséquilibrées

- Le nombre d'observations est beaucoup plus faible pour une classe que pour l'autre



- Classe minoritaire = 5% des observations \Rightarrow un modèle qui affecte toute observation à la classe majoritaire ne fait que 5% d'erreurs ...

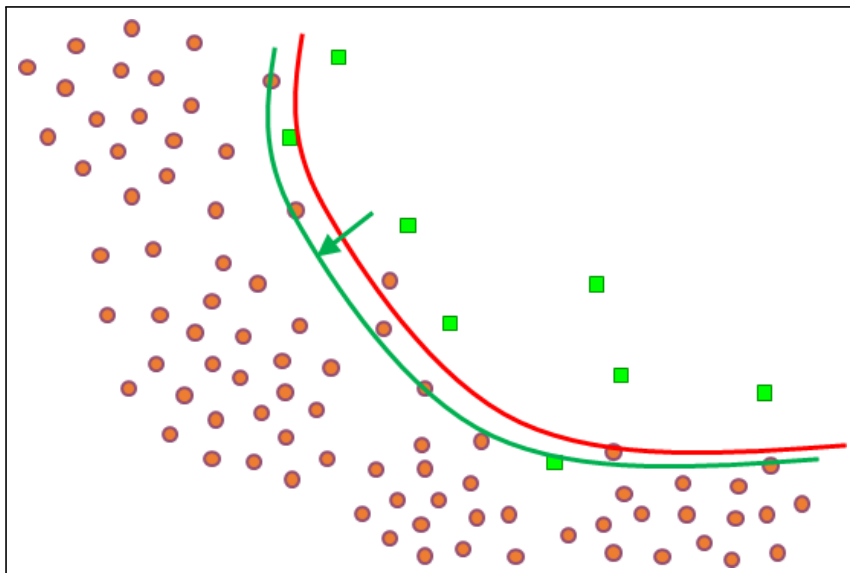
Classes déséquilibrées (2)

Quelques solutions

1. Certaines méthodes de modélisation sont **moins sensibles** que d'autres au déséquilibre entre les classes
 - Par ex. SVM maximisent une marge \Rightarrow seules comptent les données les plus proches de la frontière
2. Changer la façon de **mesurer la performance** du modèle, utiliser par ex.
 - κ (*Kappa*) pour mesurer « l'accord » entre les prédictions du modèle obtenu et le bon classement
 - Le **rappel** mesuré pour la classe minoritaire = proportion d'éléments de la classe détectés comme tels

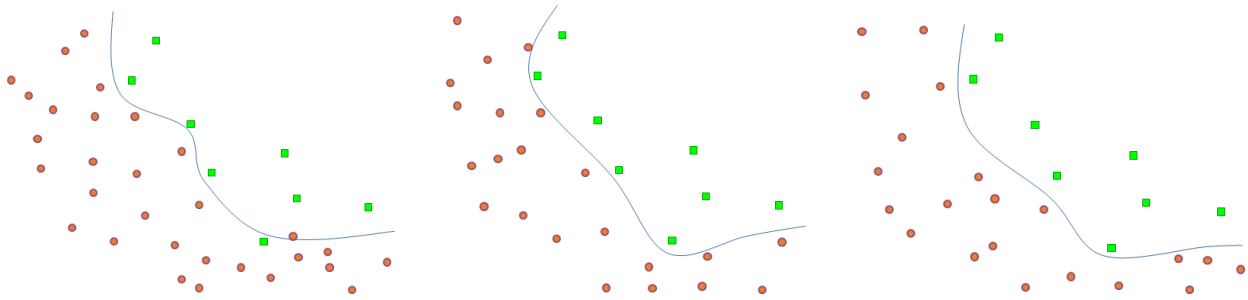
Classes déséquilibrées (3)

3. Utiliser des **pénalités** supérieures pour les erreurs faites sur la classe minoritaire



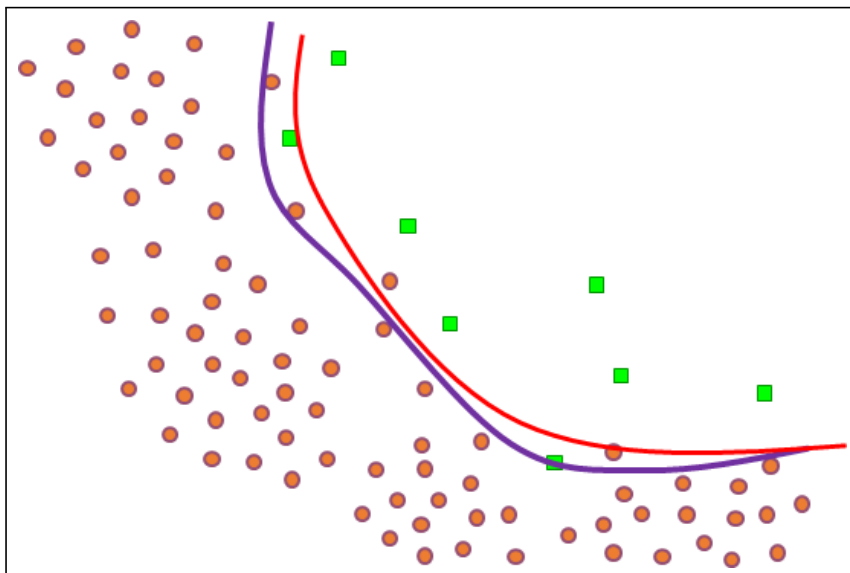
Classes déséquilibrées (4)

4. Employer un **ensemble** de classifieurs, chacun construit sur un **échantillon** de la classe majoritaire et **toutes** les données de la classe minoritaire



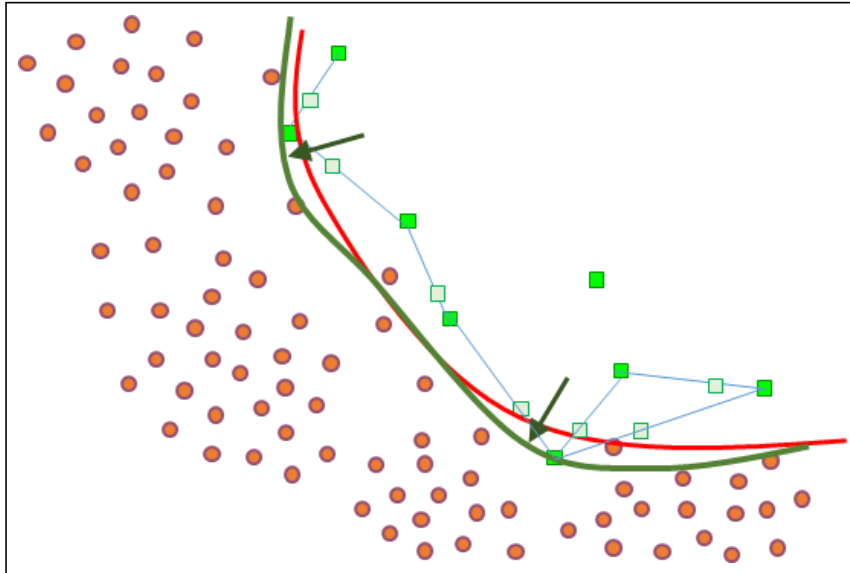
Classes déséquilibrées (5)

4. Employer un **ensemble** de classifieurs, chacun construit sur un **échantillon** de la classe majoritaire et **toutes** les données de la classe minoritaire



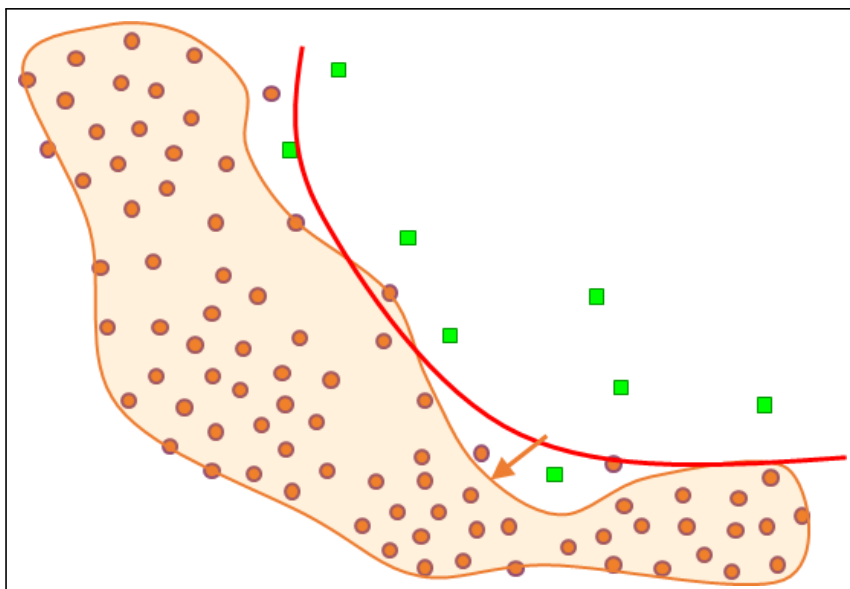
Classes déséquilibrées (6)

5. Générer des **observations synthétiques** pour la classe minoritaire (par ex. avec *Synthetic Minority Over-sampling TEchnique*, SMOTE)



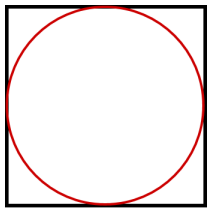
Classes déséquilibrées (7)

6. **Reformuler** le problème comme une détection d'observations aberrantes (celles de la classe minoritaire)



Dimension très élevée des données

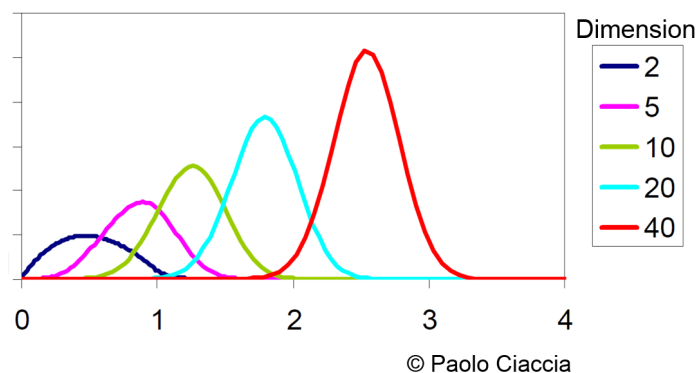
- Dimension d des données : nombre de variables unidimensionnelles qui caractérisent les données
- d élevé \Rightarrow « malédiction de la dimension » (*curse of dimensionality*)
 1. A nombre de données fixé, la densité diminue exponentiellement avec la dimension \Rightarrow problèmes pour l'estimation de densités, tests statistiques
 2. Les données uniformément distribuées dans des volumes en dimension d sont proches des hypersurfaces externes (de dimension $d - 1$)



Dimension	Vol. sphère / vol. cube englobant
1	1
2	0,78732
4	0,329707
6	0,141367
8	0,0196735
10	0,00399038

Dimension très élevée des données (2)

3. Augmentation du bruit si nombreuses variables non pertinentes
 4. Augmentation du risque de trouver explicatives des variables qui ne le sont pas
 5. La variance de la distribution des distances entre observations diminue avec l'augmentation de la dimension (« concentration des mesures »)
 - \Rightarrow problèmes pour l'exploitation des distances dans kppv ou classification automatique
- Exemple pour données issues d'une distribution uniforme :



6. Coût d'analyse / modélisation / décision proportionnel à d^2 ou d^3

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

Nature des données

- Observation (*individu, entité, ...*) = valeurs que prennent simultanément les variables (*attributs, traits, ...*) qui interviennent dans le problème
 - Par ex., une région d'image correspondant (ou non) à un véhicule

Observation	X_1	X_2	...	X_d
e_1
e_2
...
e_n

Nature des données (2)

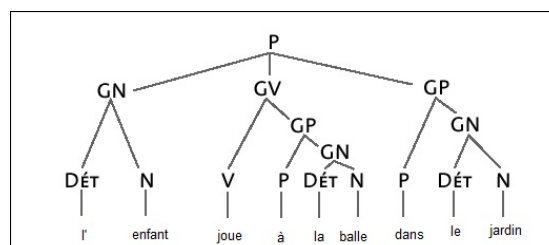
Type d'une variable \leftarrow nature des valeurs qu'une variable peut prendre

- Quantitative (ou numérique)
 - Continue : longueur, durée, température, autres mesures physiques, etc.
 - Discrète : population, etc.
- Qualitative (ou catégorielle)
 - Ordinale (un ordre total est présent entre les valeurs possibles) : classement, échelle de Lickert (ex. *Pas du tout d'accord* / *Pas d'accord* / *Ni en désaccord ni d'accord* / *D'accord* / *Tout à fait d'accord*), etc.
 - Nominale : catégorie socio-professionnelle, nom de marque, etc.

Nature des données (3)

Autres types ou distinctions utiles :

- Structurées (\leftrightarrow non structurées) : chaque valeur possède une structure interne
 - Par ex., une phrase possède une structure grammaticale (arbre syntaxique) qui est en général importante pour le traitement de la phrase



- Valeurs-ensembles : chaque valeur est un sous-ensemble d'un très grand ensemble
 - Par ex., un texte est souvent traité comme un ensemble de mots
- Observations dépendantes : un « voisinage » entre observations \rightarrow dépendances
 - Dépendances temporelles (entre observations successives) : taux de change, débit rivière, etc.
 - Dépendances spatiales (entre observations spatialement proches) : fertilité du sol, pollution atmosphérique, etc.

Représentation des variables

■ Variables ordinales :

■ Représentation par valeurs numériques ?

<i>Pas du tout d'accord</i>	1
<i>Pas d'accord</i>	2
<i>Ni en désaccord ni d'accord</i>	3
...	...

→ introduction de distances **arbitraires** entre modalités ← **à éviter**

⇒ Représentation plutôt par des codes binaires :

<i>Pas du tout d'accord</i>	0 0 0 0 1
<i>Pas d'accord</i>	0 0 0 1 1
<i>Ni en désaccord ni d'accord</i>	0 0 1 1 1
...	...

Représentation des variables (2)

■ Variables nominales :

■ Représentation par valeurs numériques ?

<i>Enseignant</i>	1
<i>Médecin</i>	2
<i>Technicien</i>	3
...	...

→ introduction d'un ordre **arbitraire** et de distances **arbitraires** entre modalités ← **à éviter**

⇒ Représentation plutôt par un codage disjonctif :

<i>Enseignant</i>	1 0 0 ...0 0 0
<i>Médecin</i>	0 1 0 ...0 0 0
<i>Technicien</i>	0 0 1 ...0 0 0
...	...

Représentation des variables (3)

- Variables à valeurs-ensembles : représentation vectorielle issue de la fonction caractéristique du sous-ensemble correspondant à la valeur

	armée	calme	Clube	garder	gens	pouvoir					
	0	0,5	0	0,5	0,4	0,3	0,2	0	0	0,3	0
1															10...4

- Variables à valeurs structurées : utilisation de méthodes à noyaux (par ex. noyaux sur arbres, voir l'ingénierie des noyaux dans RCP209)

⇒ La représentation des variables doit être en accord avec la nature des caractéristiques auxquelles les variables correspondent !

Plan du cours

- 2 Objectifs et contenu de l'enseignement
- 3 Organisation de l'enseignement (RCP208)
- 4 Exemples d'applications et nature du problème de modélisation
- 5 Décision et généralisation
- 6 Problèmes posés par les données
- 7 Nature des données et leur représentation
- 8 Modélisation : critères de choix, approche générale

Critères de choix de méthode de modélisation

- Mesure de performance sur les données de test : erreur minimale, aire sous la courbe ROC, taux minimal de faux négatifs...
 - Cette performance ne doit pas être l'unique critère de choix !
- Lisibilité (\sim *explicability*) : résultats ou décisions interprétables
 - Pour des applications critiques (par ex. contrôle réaction chimique à risque) on ne peut se contenter d'une solution « boîte noire »
 - La lisibilité rend possible la vérification/validation *a priori*
 - Solutions pour rendre lisibles des modèles qui ne le sont pas *a priori* (par ex. extraction de règles d'un réseau de neurones)
- Rapidité de la construction du modèle, de la prise de décision
 - Contraintes de temps sur la (re)construction du modèle ou sur la prise de décision ?
- Eventuellement, autres critères spécifiques à l'application

Critères de choix de méthode de modélisation (2)

- *Usability* : facilité d'emploi
 - Exemple : un expert est indispensable pour mettre au point le modèle et pour toute évolution ultérieure ?
- *Embedability* : facilité d'introduction dans un système global
 - Exemple : la méthode impose des contraintes fortes sur l'échange de données (par ex. codage spécifique des données d'entrée, de sortie) ?
- Flexibilité : adaptation facile au changement de spécifications
 - Exemples : faut-il repartir de zéro si les conditions de mesure changent ou si un capteur est remplacé par un autre de courbe de réponse différente ?
- Passage à l'échelle (*scalability*)
 - Exemple : gros volumes / débits de données

Approche générale de la modélisation

