Apprentissage automatique / Statistique

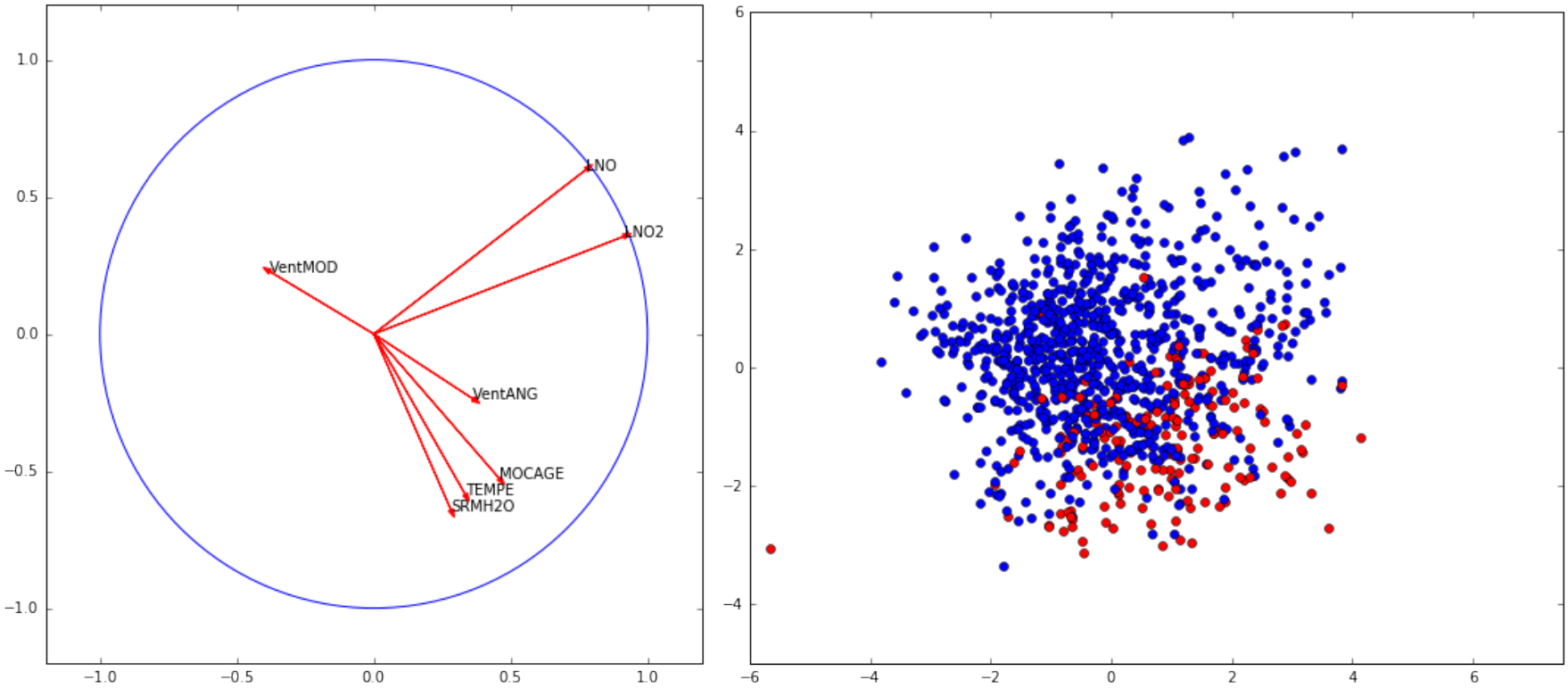
Introduction

PHILIPPE BESSE

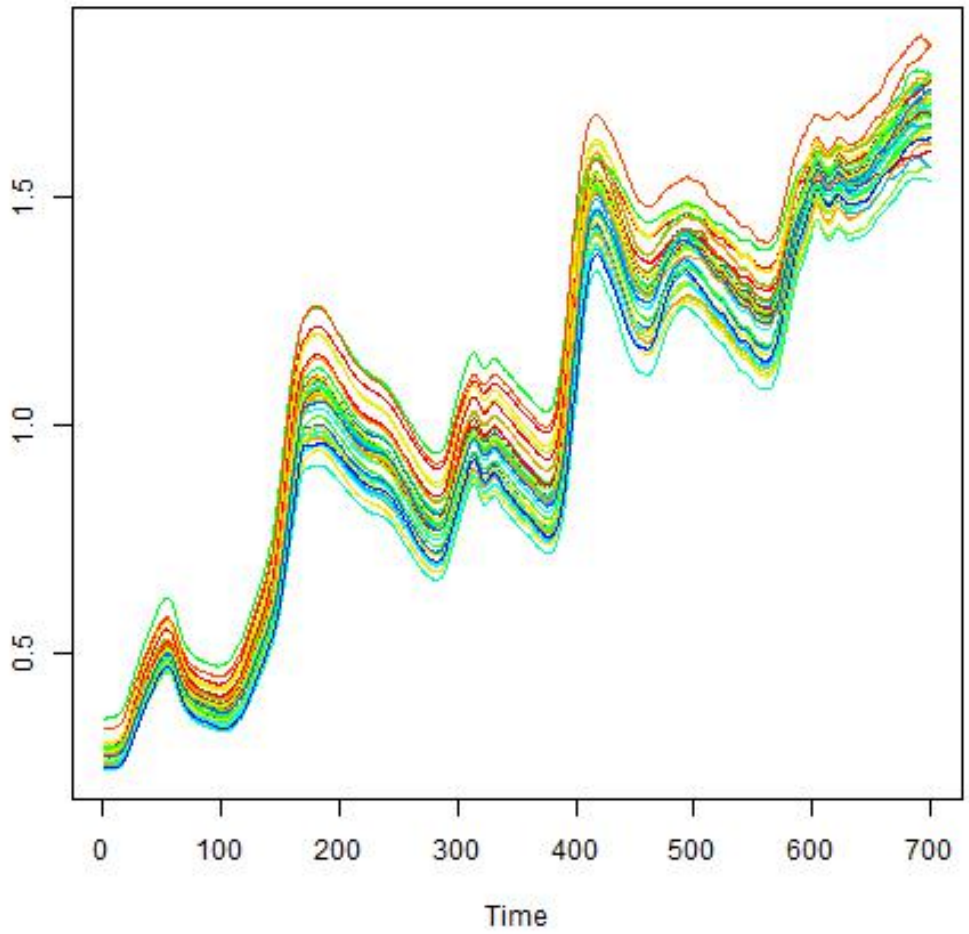
INSA de Toulouse
Institut de Mathématiques

Questions ?

- Facteurs de risque épidémiologiques
- Facteur génétique ou biomarqueurs
- Reconnaissance de forme (caractères)
- Adaptation statistique en prévision météo (pic d'ozone)
- Score d'appétence ou d'attrition en GRC
- Méta modèle ou réduction de modèle physique
- Détection défaillance ou fraude (atypique)
- ...
- Estimer un modèle apprendre un algorihtme
- Minimiser une erreur de prévision ou risque



Ozone : premier plan de l'ACP réduite (47%)



Cookies : Spectres NIR de pâte à gâteaux ($n = 72, p = 700$)

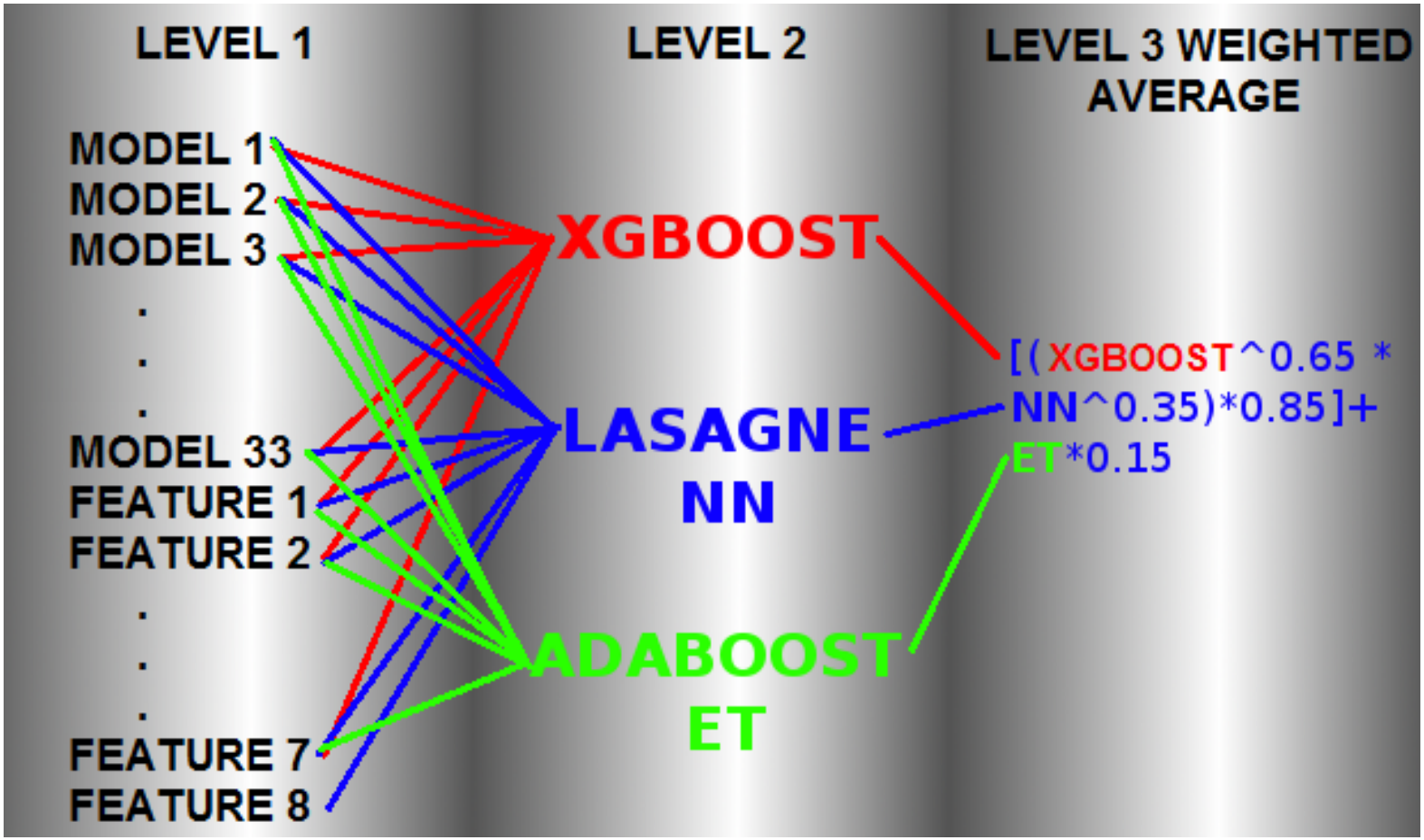
Objectif ?

- Explorer ou, représenter, décrire
- Expliquer ou tester, vérifier une influence
- Prévoir et sélectionner, interpréter
- Prévision "brute"

Statistique *vs.* Apprentissage Statistique *vs.* Machine

But ?

- Publication académique (*Benchmarks — UCI repository*)
- Solution industrielle peu *glamour*
- Concours de type Kaggle



Concours Kaggle : Identify people who have a high degree of Psychopathy based on Twitter usage.

Apprentissage Supervisé *vs.* non-supervisé

- Observation ou non d'une variable Y à expliquer
- Y : variable cible (*target*) ou sortie (*output*)
- Modélisation : $Y = \widehat{f}(X) + \varepsilon$
- Sinon : apprentissage (ou classification) non supervisé
- Faux amis : Discrimination *classification*) *vs.* classification (*clustering*

Les données

- p variables (*features*) explicatives ou prédictives $X = (X^1, \ldots, X^p)$
- n observations, individus, unités statistiques, *instances*
- *Attention*, données préalables ou *planifiées*
- Ensemble d'apprentissage $D_1^n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- $\mathbf{x}_i \in \mathcal{X}$ $(= \mathbb{R}^p)$, $y_i \in \mathcal{Y}$ pour $i = 1 \ldots n$
- Choix d'un ensemble de modèles, méthodes, algorithmes

$$Y = f(\boldsymbol{X}), \quad f \in \mathcal{F}$$

Régression *vs.* Discrimination

sorties quantitatives	sorties qualitatives
$Y \in \mathcal{Y} \subset \mathbb{R}^p$	$Y \in \mathcal{Y}$ fini
$\downarrow$	$\downarrow$
régression	discrimination, classement reconnaissance de forme

Estimation *vs.* apprentissage

- Statistique classique
  - Modèle "vrai"
  - Estimer les paramètres ou ajuster un modèle
  - Tester l'influence
  - Expliquer l'effet de facteurs
- Apprentissage automatique / statistique
  - Apprendre un algorithme
  - Prévision (et interprétation)
  - Modèle ou algorithme parcimonieux


Les devises Shadok
ROUXEL
POURQUOI FAIRE SIMPLE
QUAND ON PEUT FAIRE
COMPLIQUÉ ?!

Choix de méthode

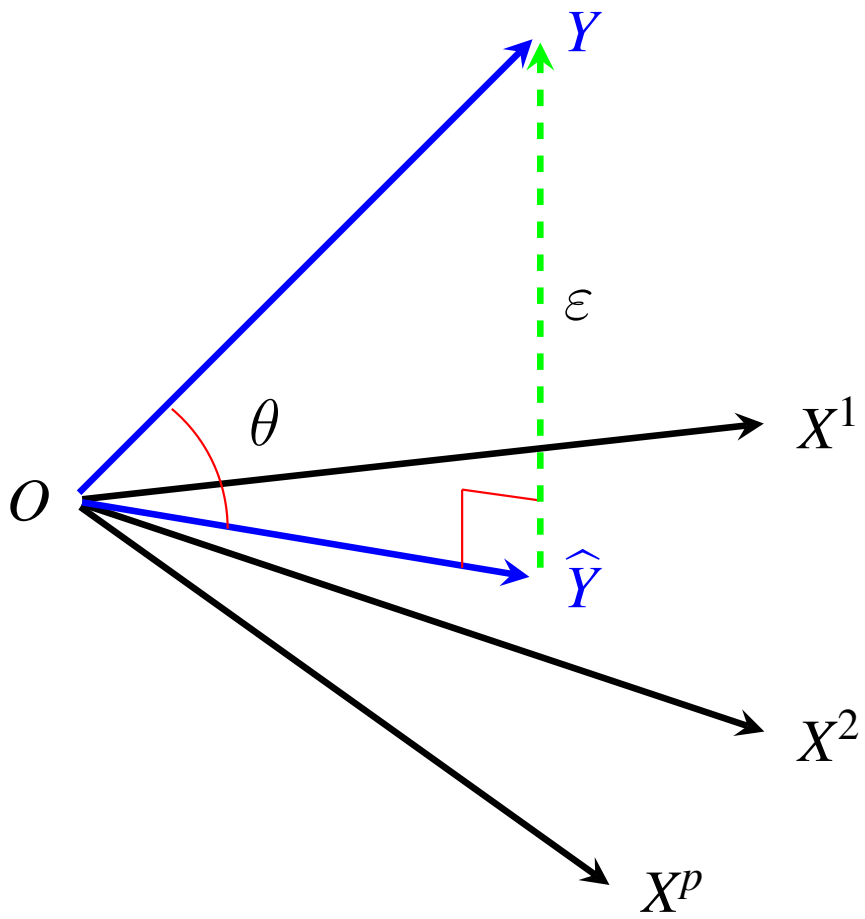
- Bibliographie explosive
- Pas de méthode universellement meilleure
- Adaptation de la méthode aux données
- Qualité : erreur de prévision

Choix de modèle

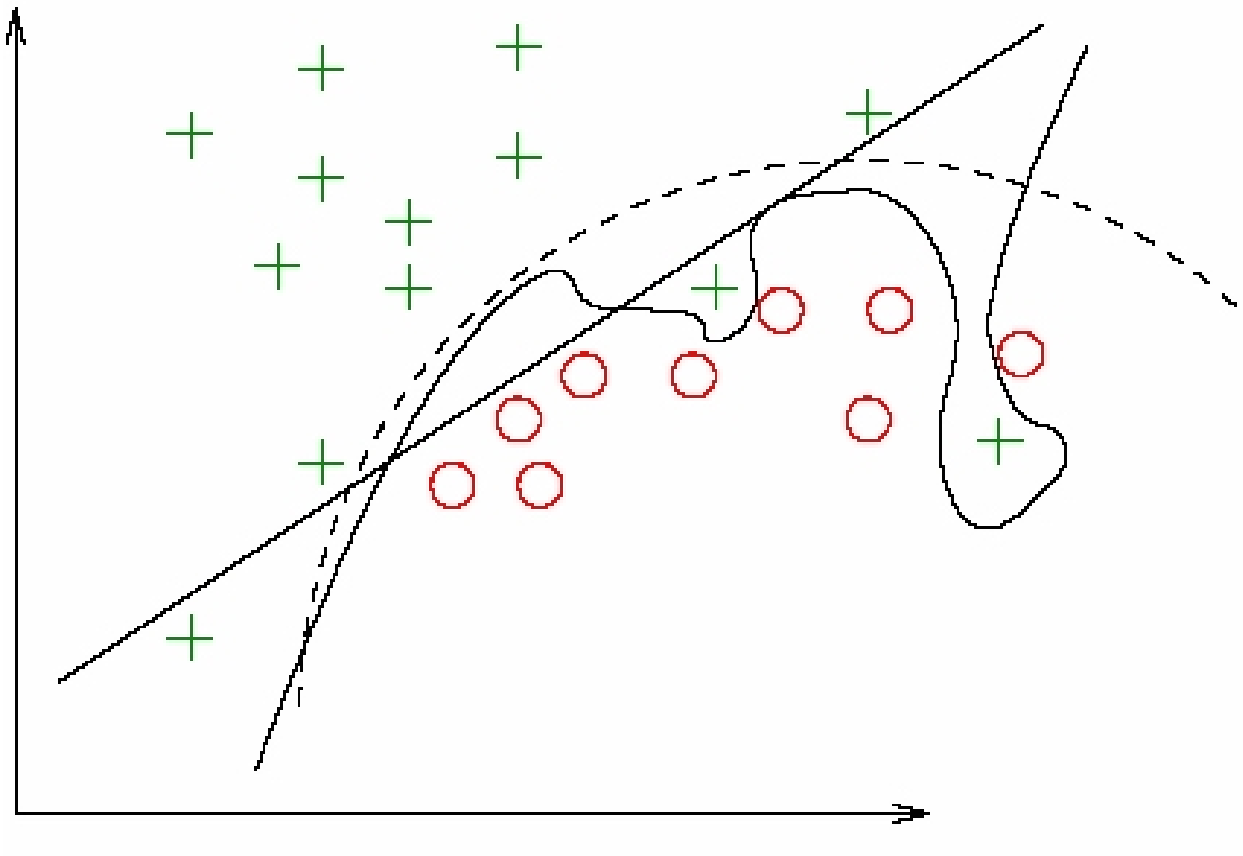
- Contrôle de la complexité / flexibilité
- Flexibilité et ajustement
- Robustesse et prévision
- Équilibre biais-variance

Stratégies de choix de modèle

- Contrôle de la complexité
- Choix de modèle : sélection *vs.* régularisation
  - Sélection de variables et nombre de paramètres
  - Sélection et pénalisation en $||.||_{l_1}$
  - Régularisation et pénalisation en $||.||_{l_2}$ (*ridge, schrinkage*)



Régression : Projection $\widehat{Y}$ de Y sur l'espace vectoriel Vect$\{\mathbf{1}, X^1, \ldots, X^p\}$



Classification supervisée : Complexité des modèles

Première phase : *Data munging*

1. Extraction avec ou sans sondage
2. Exploration, visualisation
3. Nettoyage, transformations des données, choix d'une base (spline, ondelettes, Fourier)...
4. Nouvelles variables (*features*)
5. Gestion des données manquantes

Deuxième phase : *Apprentissage*

1. Partition aléatoire de l'échantillon : apprentissage, (validation), test
2. **Pour** chacune des méthodes considérées :
   - Apprentissage (estimation) fonction de θ (complexité)
   - Optimisation de θ : validation ou validation croisée
3. Comparaison des méthodes : erreur de prévision sur échantillon test
4. Itération éventuelle (*Monte Carlo*)
5. Choix de la méthode (prévision *vs.* interprétabilité).
6. Ré-estimation du modèle, exploitation

Possible : combinaison de modèles

Question : Où faire porter l'effort

- *Data munging*
- Sélection des méthodes à comparer
- Optimisation des paramètres
- Combinaison optimale de modèles

En fonction de :

- But (temps imparti)
- Régularité du problème sous-jacent
- Structure et propriétés des données

Méthodes et algorithmes d'apprentissage

- Modèle linéaire avec sélection ou régularisation
- Régression PLS avec pénalisation
- Modèle linéaire binomial (logistique) avec sélection ou régularisation
- Analyse discriminante, k plus proches voisins
- Arbres binaires de décision (CART)
- Réseaux de neurones, perceptron, apprentissage profond
- Agrégation de modèles : *random forest, boosting...*
- SVM ou séparateurs à vaste marge
- ...
- Imputation de données manquantes
- Détection d'anomalies ou d'atypiques

Exemples & tutoriels

- R et/ou Python
- ▸ wikistat.fr ou ▸ github.com/wikistat
  - Fil rouge : Adaptation statistique de prévision d'ozone
  - Exemples jouet : discrimination de nuages dans $\mathbb{R}^2$
  - Diagnostic d'une maladie coronarienne
  - Spectrographie en proche infra-rouge (*cookies, tecator*)
  - QSAR ou criblage virtuelle de molécules
  - Sélection de gènes
  - GRC Score d'appétence ou d'attrition (*churn*)
  - Diagnostic de tumeurs
  - Données d'enquête (*adult census*)
  - Détection de pourriels (spam)
  - ...

Rappel :

Objectif pédagogique :

Apprendre à auto-apprendre...