Apprentissage statistique

Apprentissage supervisé: introduction

Cours pour non-spécialistes

Aurélien Garivier

Algorithme de d'apprentissage

Cadre classique (batch) :

Données : *échantillon d'apprentissage* $(x_k, y_k)_{1 \leq k \leq n}$ constitué d'observations que l'on suppose représentatives et sans lien entre elles.

Objectif : prédire les valeurs de y associées à chaque valeur possible de $x \in \mathcal{X}$.

Classification : $\mathcal{Y}$ discret (typiquement, binaire) pour chaque valeur de $x \in \mathcal{X}$, il faut prédire la classe la plus souvent associée.

Régression : $\mathcal{Y}$ continu, voire plus (fonctionnel).

Règle de classification : à partir de l'échantillon d'apprentissage, construire $f_n : \mathcal{X} \to \mathcal{Y}$ associant, à chaque entrée possible x, la classe y prédite pour elle.

Apprentissage vs. Statistiques ?

- Les données sont des réalisations de v.a. iid de même loi que

$$(X, Y) \sim P_{(X,Y)} \in \mathcal{M}$$

- Modèle *instrumental* : on ne pense pas que ça soit vrai, ni que $P \in \mathcal{M}$
- Pas de "vrai modèle", de "vraies valeurs du paramètre", etc.
- Consistance statistique sans intérêt
- Souvent ($\neq$ étude statistique) données disponibles avant intervention du statisticien (malheureusement)
- Tous les coups sont permis, seul critère = efficacité prédictive
- Classification :

$$R_n = \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[P_{(X,Y)}(f_n(X) \neq Y)\right] .$$

- Régression : typiquement

$$R_n = \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[E_{(X,Y)}\left((Y - f_n(X))^2\right)\right] .$$

Paramétrique vs. Non-paramétrique !

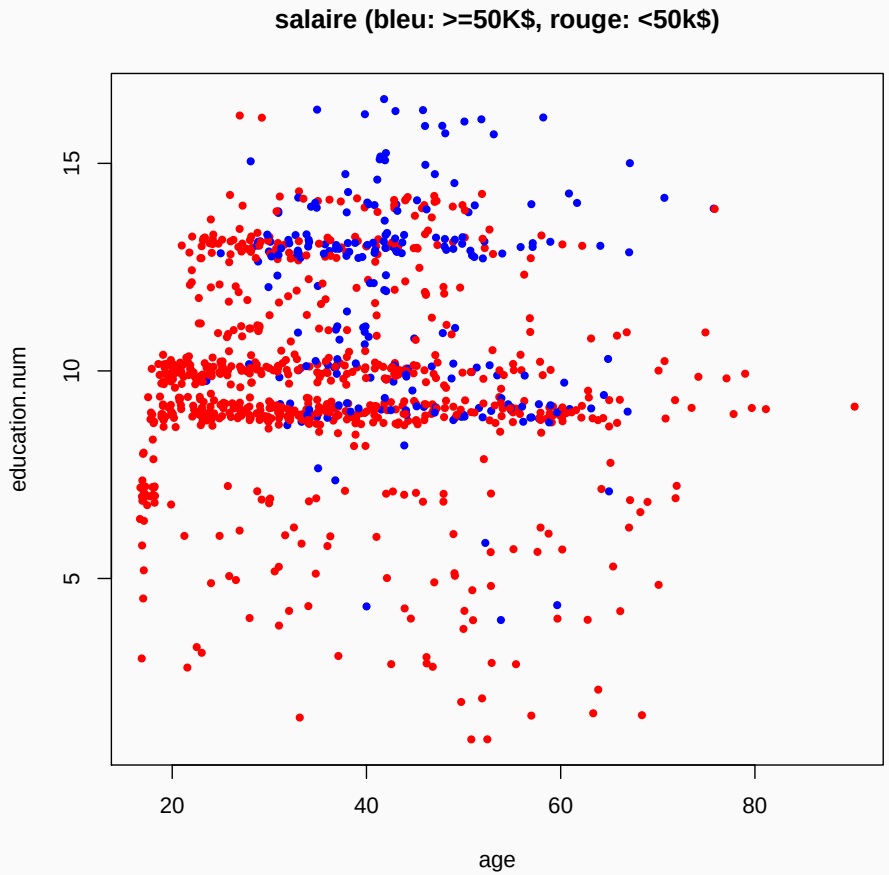
- Théoriquement, quand un modèle est vrai il est optimal de l'utiliser :
  - Théorème de Gauss-Markov : parmi les estimateurs sans biais, celui des moindres carrés est de variance minimale
  - MAIS on peut avoir intérêt à sacriier du biais contre de la variance !

$\Longrightarrow$ Même quand il y en a un 'vrai' modèle, on n'a pas forcément intérêt à l'utiliser

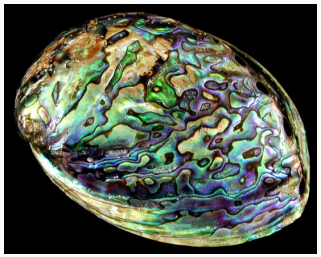
- Des approches non-paramétriques peuvent avoir une efficacité proche :
  - cf Test de Student vs Mann-Whitney
  - exemple : k-NN versus régression polynomiale
- ... et ils sont beaucoup plus robustes !

Exemple de problème de classification

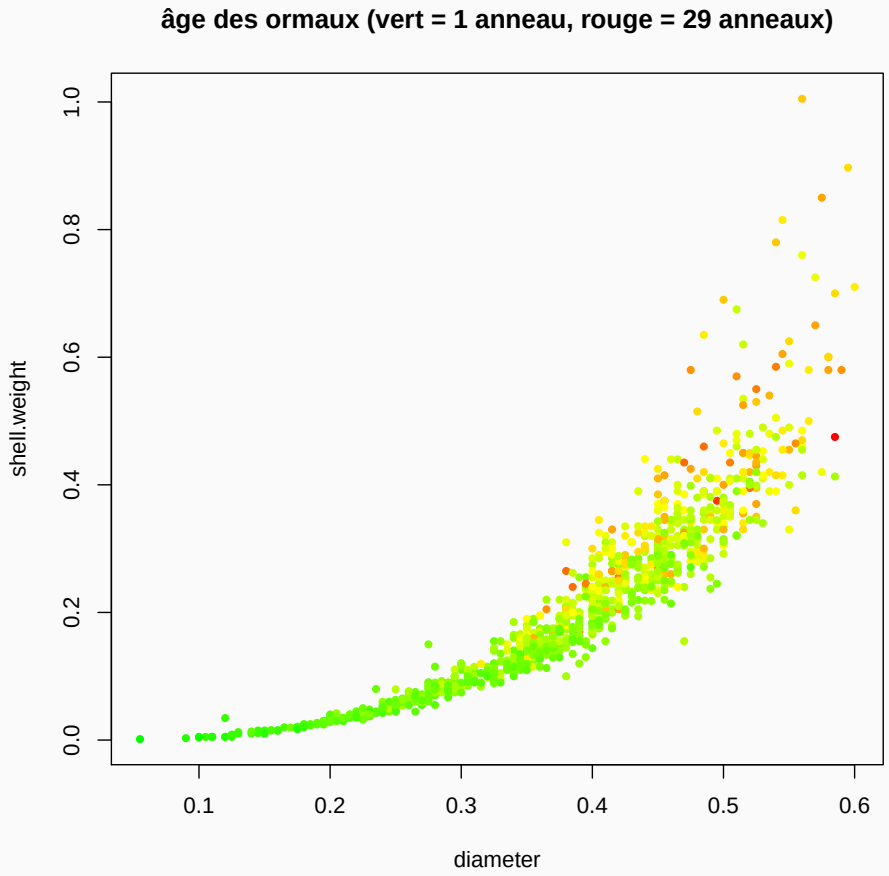
Objectif : prédire qui gagne plus de 50k$ à partir de données de recensement.



Exemple de problème de régression



Prédire l'âge d'un ormeau (abalone) à partir de sa taille, son poids, etc.



Nous disposons d'un ensemble d'observations. Les caractéristiques ou variables $\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^p)$ dites explicatives ont été observées sur un ensemble de n objets, individus ou unités statistiques.

- Premier travail : mener une exploration statistique des données.
  - allure des distributions,
  - présence de données atypiques,
  - corrélations et cohérence,
  - transformations éventuelles des données,
  - description multidimensionnelle,
  - classification.
- Deuxième travail : modélisation statistique ou encore d'apprentissage pour la prédiction d'un variable cible Y par les variables explicatives $(\mathbf{X}^1, \ldots, \mathbf{X}^p)$.

L'enchaînement de ces étapes (exploration puis apprentissage) constitue le fondement de la fouille de données.

But : Déterminer la stratégie à mettre en oeuvre pour aboutir au bon apprentissage ou au bon modèle prédictif à partir des données observées.

Contrairement à une démarche statistique traditionnelle dans laquelle l'observation des données est intégrée à la méthodologie (plannification expérimentale), les données sont ici préalable à l'analyse.

Néanmoins, il est clair que les préoccupations liées à leur analyse et à son objectif doivent intervenir le plus en amont possible pour s'assurer quelques chances de succès.

Étapes de la fouille de données

1. Extraction des données avec ou sans apprentissage : techniques de sondage appliquées ou applicables à des bases de données.

2. Exploration des données
   - pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences,
   - pour l'étude des distributions, des structures de corrélation, recherche de typologies,
   - pour des transformations de données.

3. Partition aléatoire de l'échantillon (apprentissage, validation, test) en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prédiction en vue des choix de modèles, choix et certification de méthode.

Étapes de la fouille de données (suite)

4. Pour chacune des méthodes considérées : modèle linéaire général (gaussien, binomial ou poissonien), discrimination paramétrique (linéaire ou quadratique) ou non-paramétrique, *k* plus proches voisins, arbre, réseau de neurones (perceptron), support vecteur machine, combinaison de modèles (bagging, boosting)
   - estimer le modèle pour une valeur donnée d'un paramètre de complexité : nombre de variables, de voisins, de feuilles, de neurones, durée d'apprentissage, largeur de fenêtre...
   - optimiser ce paramètre (sauf pour les combinaisons de modèles affranchies des problèmes de sur-apprentissage) en fonction de la technique d'estimation de l'erreur retenue : échantillon de validation, validation croisée, approximation par pénalisation de l'erreur d'ajustement.

Étapes de la fouille de données (suite et fin)

5. Comparaison des modèles optimaux obtenus (un par méthode) par estimation de l'erreur de prédiction sur l'échantillon test ou, si la présence d'un échantillon test est impossible, sur le critère de pénalisation de l'erreur (Akaike par exemple) s'il en existe une version pour chacune des méthodes considérées.

6. Itération éventuelle de la démarche précédente (validation croisée), si l'échantillon test est trop réduit, depuis l'étape 3. Partitions aléatoires successives de l'échantillon pour moyenner sur plusieurs cas l'estimation finale de l'erreur de prédiction et s'assurer de la robustesse du modèle obtenu.

7. Choix de la méthode retenue en fonction de ses capacités de prédiction, de sa robustesse mais aussi, éventuellement, de l'interprétabilité du modèle obtenu.

Explicatives L'ensemble des p variables explicatives ou prédictives est noté X, il est constitué de variables

- $X_{\mathbb{R}}$ toutes quantitatives (rq : variable explicative qualitative à 2 modalités (0,1) peut être considérée comme quantitative)
- X_E toutes qualitatives,
- $X_{\mathbb{R} \cup E}$ un mélange de qualitatives et quantitatives.

À expliquer La variable à expliquer ou à prédire ou *cible* (target) peut être

- Y quantitative,
- Z qualitative à 2 modalités,
- T qualitative.

1. Modèle linéaire généralisé

   `RLM` $X_{\mathbb{R}}$ et Y
   `ANOVA` X_E et Y
   `ACOVA` $X_{\mathbb{R} \cup E}$ et Y
   `Rlogi` $X_{\mathbb{R} \cup E}$ et Z
   `Lglin` X_T et T

2. Analyse discriminante

   `ADpar/nopar` $X_{\mathbb{R}}$ et T

3. Classification and regression Tree

   `ArbReg` $X_{\mathbb{R} \cup E}$ et Y
   `ArbCla` $X_{\mathbb{R} \cup E}$ et T

4. Réseaux neuronaux

   `percep` $X_{\mathbb{R} \cup E}$ et Y ou T

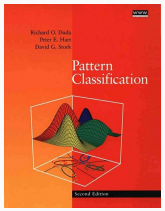
5. Agrégation de modèles

   Bagging $X_{\mathbb{R} \cup E}$ et Y ou T
   RandFor $X_{\mathbb{R} \cup E}$ et Y ou T
   Boosting $X_{\mathbb{R} \cup E}$ et Y ou T

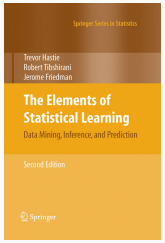
6. Support Vector Machine

   SVM-R $X_{\mathbb{R} \cup E}$ et Y
   SVM-C $X_{\mathbb{R} \cup E}$ et T

Bibliographie - Ressources

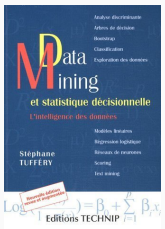


Pattern Classification (2001) - Wiley Interscience, *R. Duda, P. Hart, D. Stork*



The Elements of Statistical Learning (2001) - Springer, *T. Hastie, R. Tibshirani, J. Friedman*

Disponible en ligne : `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`



Data Mining - Technip, *S. Tufféry*

Cours en ligne de Andrew Ng (Stanford) : `https://www.coursera.org/course/ml`

`http://wikistat.fr/`

dont sont issus certains de ces slides !

`http://scikit-learn.org`

Base de données de benchmarking : `http://archive.ics.uci.edu/ml/`

Roadmap

Les règles de classification/régression

La méthodes des k plus proches voisins

Choix de modèle et fléau de la dimension

Régression logistique

Classification multi-classe

- Pour certaines règles de classification, extension possible directement (exemple : kNN, régression logistique).
- Sinon, deux possibilités de se ramener à la classification binaire :

  OvA (One vs. All) Pour chaque classe, construire un classifieur pour discriminer entre cette classe et tout le reste.

  Défaut : classes souvent très déséquilibrées (à prendre en compte)

  AvA (All vs. All) Pour chaque paire de classes (C_1, C_2), construire un classifieur pour discriminer entre C_1 et C_2.

  A priori, $\approx n^2/2$ classifieurs MAIS classifs entre paires plus rapides ET classes équilibrées $\implies$ souvent plus rapide.

Puis Error-Correcting Output Codes, typiquement : vote majoritaire

Règles de classification binaires aléatoires

- La règle randomisée : $f_n(x) = 1$ avec probabilité $1/2$ donne la bonne réponse une fois sur 2 !
- La règle constante classifiant toujours dans la classe qui a la probabilité p la plus grande se trompe avec probabilité $1 - p < 1/2$.

Règle optimale : Erreur de Bayes

Théorème :
La meilleure règle possible est la *règle de Bayes* :

- en classification : avec $\mathcal{Y} = \{0, 1\}$ et $\eta(x) = P(Y = 1|X = x)$:

$$f^*(x) = \mathbb{1}\{\eta(x) > 1/2\} \, .$$

- en régression : avec $\mathcal{Y} = \mathbb{R}$ et la perte quadratique :

$$f^*(x) = E[Y|X = x] \, .$$

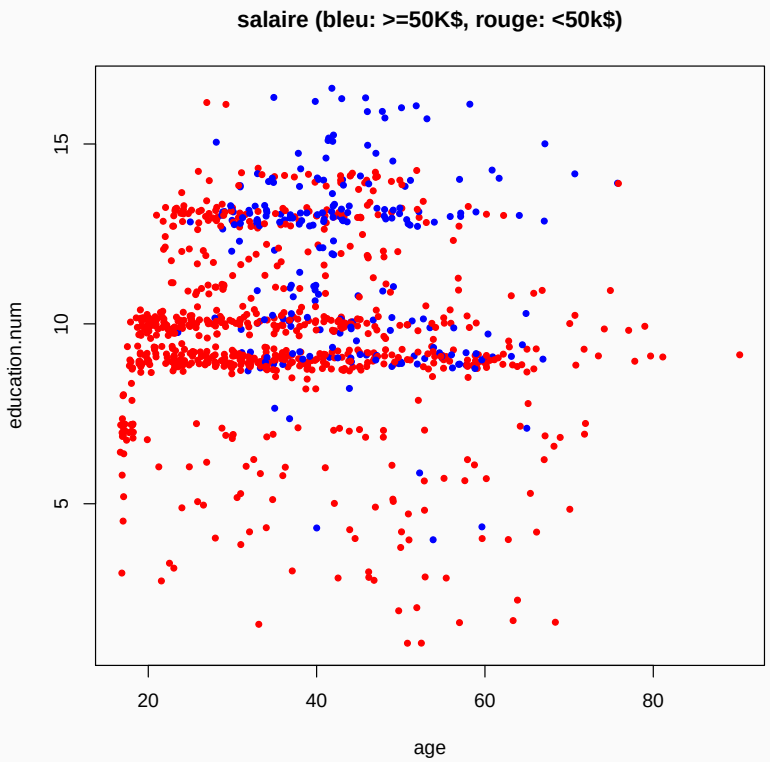
Problème : on ne connaît *jamais* P, donc on ne peut pas la calculer.

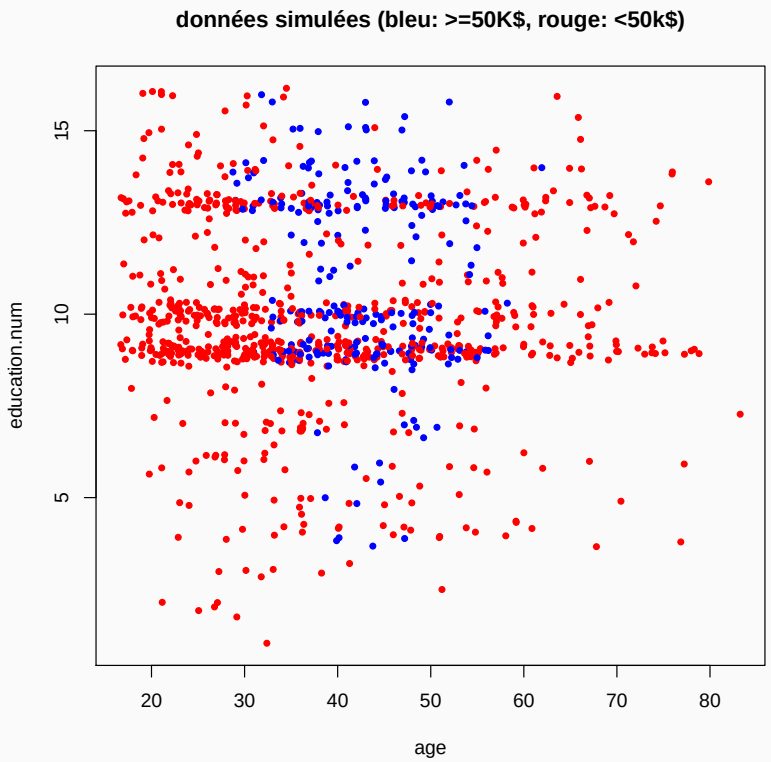
Attention : le risque de la règle de Bayes n'est pas nul !

Données simulées

- Pour contrôler la qualité d'un algorithme, on est donc amené à utiliser des *modèles génératifs* et à manipuler des *données simulées*.
- Sert à comprendre ce qu'on peut attendre (ou pas) des algorithmes d'apprentissage.
- Exemple : Sachant $X = (age, etudes)$, Y suit une loi de Bernoulli de paramètre $\max(\frac{etudes}{20} - \frac{(age-45)^2}{500}, 0)$.
- Avantage : on peut calculer la règle de Bayes (et visualiser la *frontière de Bayes*).

Exemple : salaires





Il n'y a pas de meilleure méthode !

- Chacune est plus ou moins adaptée au problème considéré, à la nature des données, aux propriétés de la relation entre descripteurs et variable expliquée...
- Il faut apprendre les qualités et les défauts de chaque méthode
- Il faut apprendre à expérimenter pour trouver la plus pertinentes
- L'estimation de la qualité des méthodes est donc centrale (mais pas toujours évidente)

Qu'est-ce qu'un bon algorithme d'apprentissage ?

Interprétabilité : la règle de classification est 'compréhensible'

Critique : fournit un score en classification, un intervalle en régression

Consistance : convergence vers l'erreur bayésienne : quand n tend vers l'infini, f_n tend vers la règle de Bayes

Minimax : cette convergence est la plus rapide possible

Non-asymptotique : garanties de performance pour n donné

Parameter-free : Paramétrage automatique

Vitesse : complexité linéaire, possibilité de parallèliser

Online : mise à jour séquentielle

Roadmap

Les règles de classification/régression

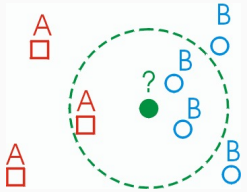
La méthodes des k plus proches voisins

Choix de modèle et fléau de la dimension

Régression logistique

Définition

Règle des k plus proches voisins : pour tout $x \in \mathcal{X}$, trouver ses plus proches voisins $x_{(1)}, \ldots, x_{(k)}$



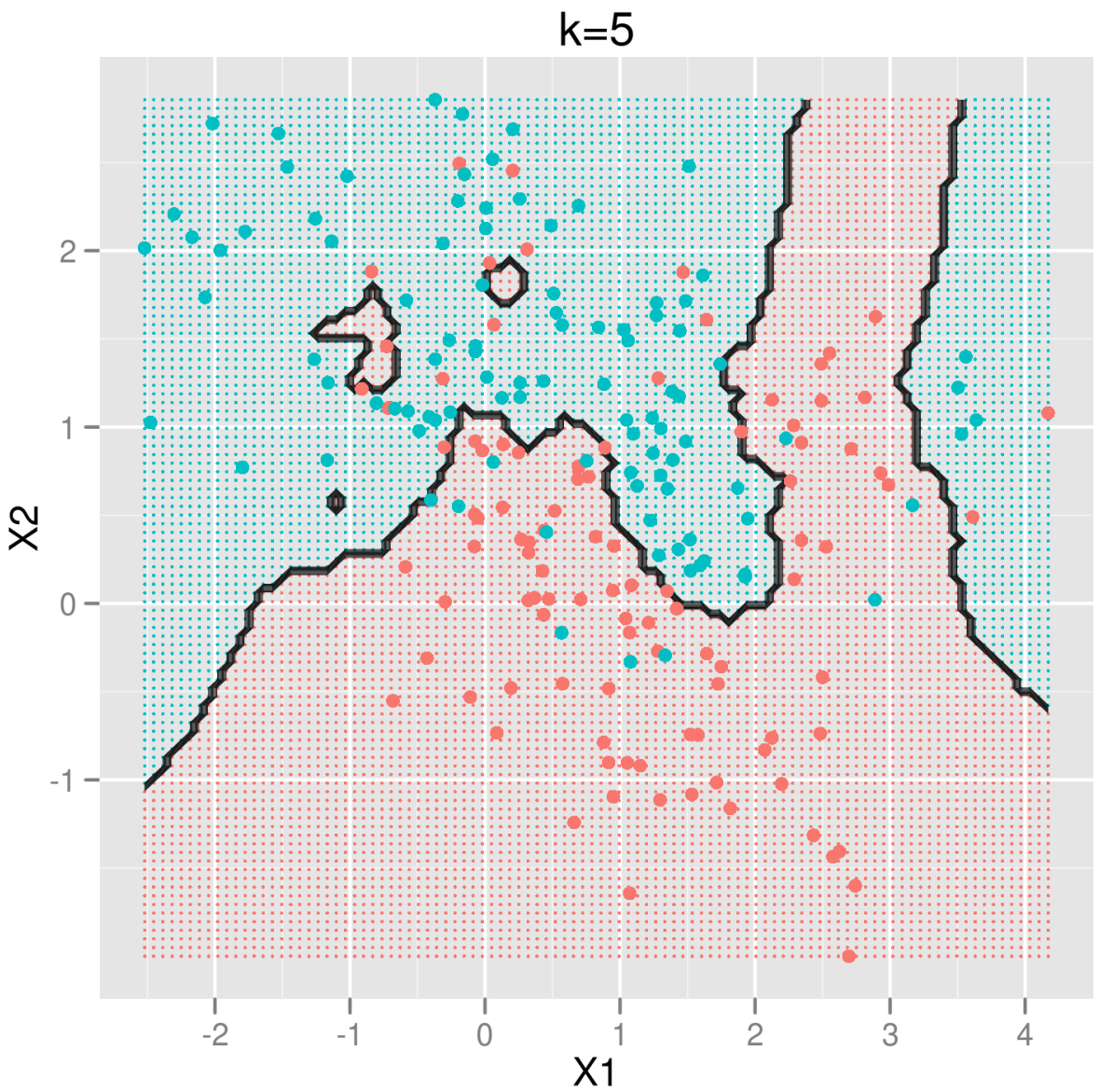
- classification :

$$f_n(x) = \arg\max_{y \in \mathcal{Y}} \sum_{j=1}^{k} \mathbb{1}\{y_{(j)} = y\}$$

- régression :

$$f_n(x) = \frac{1}{k} \sum_{j=1}^{k} y_{(j)}$$

Visualisation d'une règle k-NN



Propriétés de k-NN

Qualités :

- simplicité
- interprétabilité (?)
- pas de consistance (quel que soit k)
- MAIS asymptotiquement erreur au plus 2x supérieure à la règle de Bayes.
- possibilité de faire croître k avec n, consistance (théorique) par exemple pour $k = \log(n)$

Paramétrage :

- quelle métrique sur $\mathcal{X}$?

$\Longrightarrow$ au minimum, *normaliser* les variables (pb pour les qualitatives)

- Quelle valeur de k choisir ?

Qualités de kNN

Interprétabilité : OUI et NON

Critique : OUI mais pas très fiable

Consistance : NON mais possible si $k = \log(n)$ (par exemple)

Minimax : NON

Parameter-free : NON

Vitesse : OUI et NON, implémentation possible en $O(n \log n)$

Online : OUI