Apprentissage supervisé et théorie de la décision

Chapitre 1

Marie Chavent

à partir des cours d'Adrien Todeschini et Francois Caron

Master MIMSE - Université de Bordeaux

2015-2016

Ressources

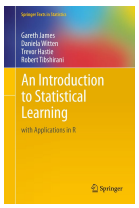
- Livres de référence

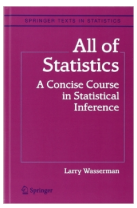
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013).
An Introduction to Statistical Learning
Springer. http://www-bcf.usc.edu/~gareth/ISL/

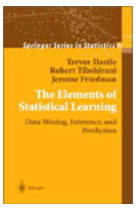
Larry Wasserman (2004).
All of statistics - A concise course in Statistical Inference
Springer Texts in Statistics.
http://www.stat.cmu.edu/~larry/all-of-statistics/

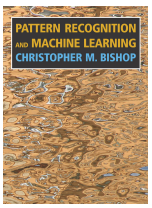
Trevor Hastie, Robert Tibshirani et Jerome Friedman (2009).
The Elements of Statistical Learning
Springer Series in Statistics.
http://statweb.stanford.edu/~tibs/ElemStatLearn/

Christopher M. Bishop (2009).
Pattern recognition and machine learning.
Springer. http://research.microsoft.com/en-us/um/people/cmbishop/prml/









Ressources

- Cours en ligne
  - Machine Learning, Andrew Ng (Stanford University) : `https://www.coursera.org/course/ml`
  - WikiStat : `http://wikistat.fr/`
- Logiciels
  - R + RStudio : `http://www.rstudio.com/`
  - Python + scikit-learn : `http://scikit-learn.org/`
- Technologies big data
  - MapReduce avec RHadoop
  - Spark+MLlib : `http://spark.apache.org/mllib/`
- Jeux de données
  - UC Irvine Machine Learning Repository : `http://archive.ics.uci.edu/ml/`
- Challenges industriels
  - Kaggle : `https://www.kaggle.com/`
  - Datascience.net : `https://datascience.net/`

Introduction

Exemples

- L'apprentissage statistique joue un rôle important dans de nombreux problèmes
  - Prédire si un patient, souffrant d'une attaque cardiaque, aura à nouveau une attaque cardiaque, en se basant sur des mesures cliniques et démographiques sur ce patient
  - Prédire la consommation en électricité d'un client à partir de données sur l'habitation et les habitudes de consommation
  - Identifier de façon automatique les chiffres du code postal sur une enveloppe
  - Identifier des groupes de sujets similaires dans un ensemble d'articles sur le web

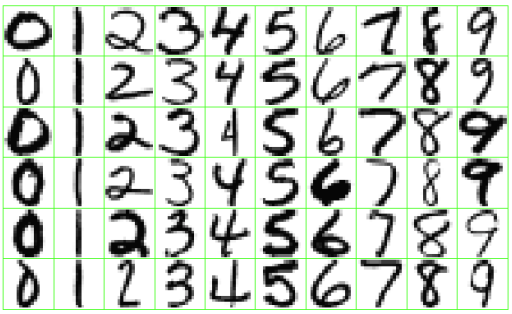
Introduction

Exemples

- Variable de sortie
  - quantitative (e.g. consommation d'électricité)
  - qualitative (attaque cardiaque: oui/non)
- Objectif : Prédire la valeur de cette variable de sortie en fonction d'un ensemble de caractéristiques (mesures cliniques du patient, données sur l'habitation, etc.)
- Ensemble d'apprentissage: couples de valeurs entrées/sorties permettant de créer un modèle de prédiction, utilisé pour prédire de nouveaux cas

Exemples

Reconnaissance de caractères manuscripts



- Objectif : Prédire la classe de chaque image (0,. . . ,9) à partir d'une matrice de 16 $\times$ 16 pixel, chaque pixel ayant une intensité de 0 à 255
- Taux d'erreur doit être faible afin d'éviter les mauvaises attributions du courrier

[Classification supervisée]

Exemples

Reconnaissance automatique de spams

Spam

```
WINNING NOTIFICATION
We are pleased to inform you of the result
of the Lottery Winners International
programs held on the 30th january 2005.
[...]  You have been approved for a lump sum
pay out of 175,000.00 euros.
CONGRATULATIONS!!!
```

No Spam

```
Dear George,
Could you please send me the report #1248 on
the project advancement?
Thanks in advance.

Regards,
Cathia
```

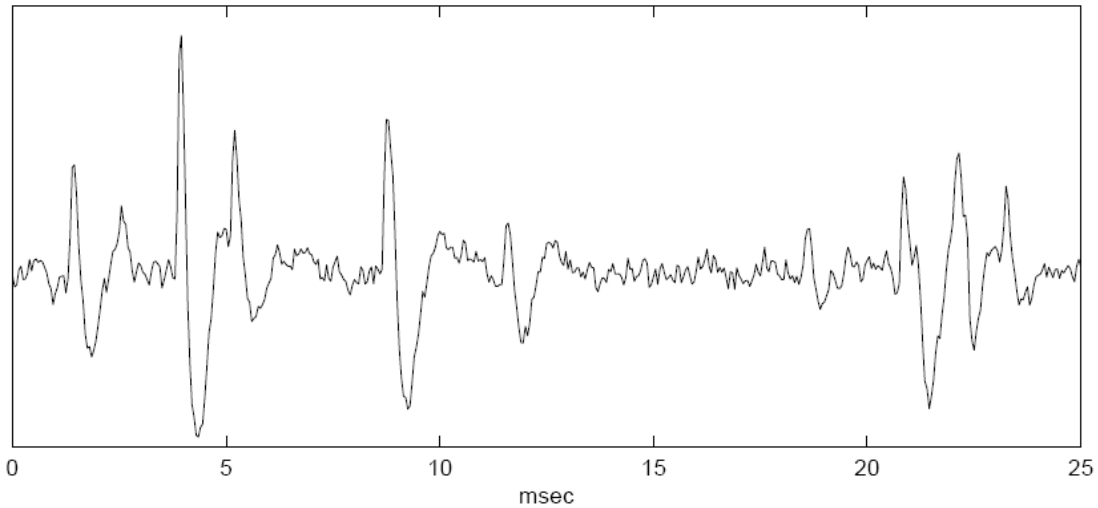
- Base de données de courriel identifiés ou non comme spam
- Objectif : Définir un modèle de prédiction permettant, à l'arrivée d'un courriel de prédire si celui-ci est un spam ou non
- Taux d'erreur doit être faible afin d'éviter de supprimer des messages importants ou d'inonder la boite au lettre de courriers inutiles

[Classification supervisée]

Exemples

Classification de signaux neuronaux

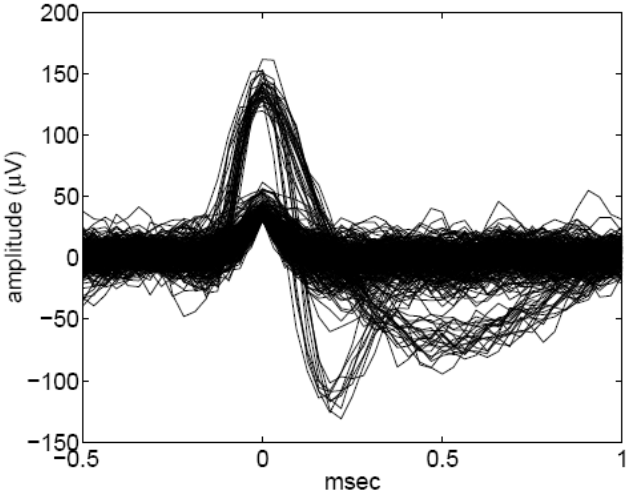
- Brefs potentiels d'action enregistrés par une micro-électrode
- Objectif : classer les signaux afin d'attribuer chaque potentiel à un neurone particulier



Exemples

Classification de signaux neuronaux

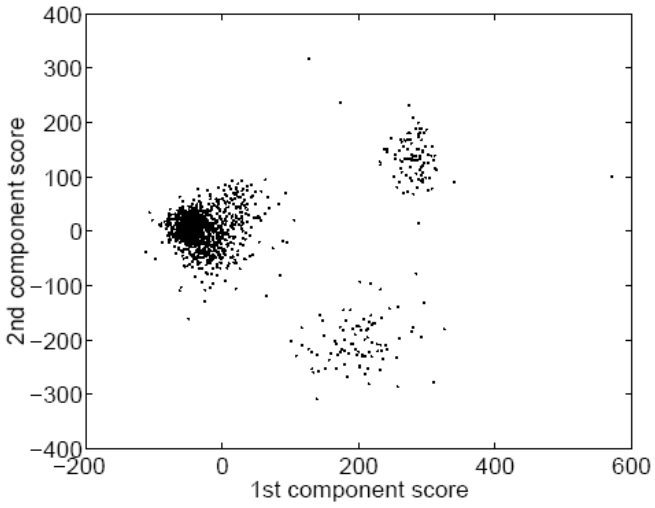
- Brefs potentiels d'action enregistrés par une micro-électrode
- Objectif : classer les signaux afin d'attribuer chaque potentiel à un neurone particulier



Exemples

Classification de signaux neuronaux

- Brefs potentiels d'action enregistrés par une micro-électrode
- Objectif : classer les signaux afin d'attribuer chaque potentiel à un neurone particulier

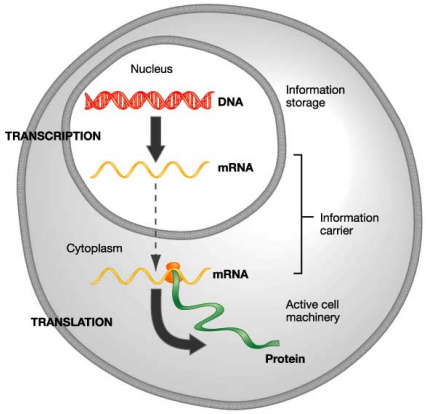


[Classification non supervisée]

Exemples

Classification de séquences d'expression de gènes

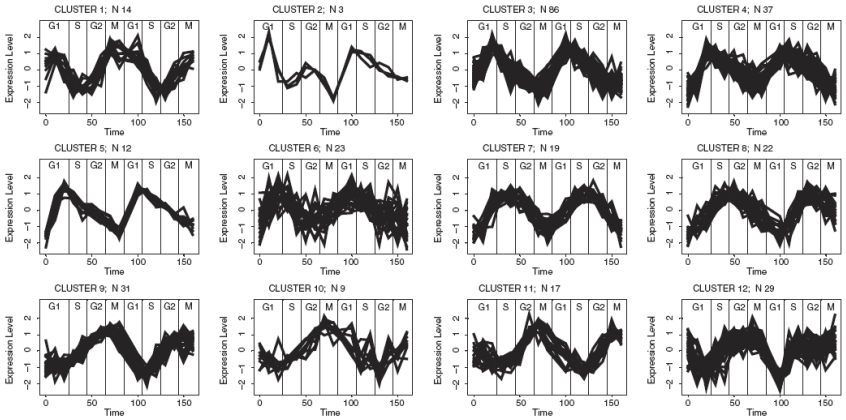
- Puces à ADN : données d'expression de gène pour plusieurs gènes et plusieurs conditions expérimentales
- Classification de gènes avec des séquences d'expression similaires



Exemples

Classification de séquences d'expression de gènes

- Puces à ADN : données d'expression de gène pour plusieurs gènes et plusieurs conditions expérimentales
- Classification de gènes avec des séquences d'expression similaires



[Classification non supervisée]

Définitions

Régression vs classification

- Variables d'entrées $X = (X_1, \ldots, X_p)$: variables explicatives
- Variable de sortie Y : variable à expliquer ou de réponse
  - Y quantitative $\rightsquigarrow$ Problème de régression
  - Y qualitative $\rightsquigarrow$ Problème de classification
- On s'intéressera principalement dans ce cours à des problèmes de classification où Y est une variable qualitative à K modalités et on parle de problème de discrimination en K classes.
- Vocabulaire : apprentissage = machine learning
  - Classification supervisée = discrimination
  - Classification non supervisée = clustering

Définitions

Classification

- Couples de variables aléatoires i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$
- $X_i = (X_{i,1}, \ldots, X_{i,p})$ prend ses valeurs dans $\mathcal{X}$, par exemple $\mathcal{X} = \mathbb{R}^p$
- Y_i prend ses valeurs dans un ensemble fini $\mathcal{Y} = \{1, \ldots, K\}$

Définition

Une règle de classification est une fonction $g : \mathcal{X} \to \mathcal{Y}$.

- Ensemble d'apprentissage : ensemble de données permettant l'apprentissage de la règle de classification g
  - Classification supervisée : on dispose d'un ensemble de couples (X_i, Y_i) pour l'apprentissage de g.
  - Classification non supervisée : on ne dispose que de données (X_i) non classées

Définitions

Classification supervisée

- Objectif : A partir de l'ensemble d'apprentissage, définir une règle de classification g permettant des prédictions précises sur de nouveau éléments
- Erreur de prédiction est quantifiée par une fonction de coût $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- En classification on considère souvent la fonction de coût 0-1

$$\begin{aligned} L(i,j) &= 0 \text{ si } i = j \\ &= 1 \text{ sinon} \end{aligned}$$

- Combien cela me coûte-t-il de faire des mauvaises prédictions?

Définitions

Classification supervisée

- Le coût moyen d'une règle de classification g est

$$\begin{aligned} E(g) &= \mathbb{E}_{X,Y}[L(Y, g(X)] \\ &= \mathbb{E}_X[\mathbb{E}_Y[L(g(X, Y)|X]] \end{aligned}$$

  - $\rightsquigarrow$ Ce coût moyen est appelé risque moyen théorique
  - $\rightsquigarrow$ Lorsque la fonction de coût 0-1 est utilisée, on parle de taux d'erreur théorique et $E(g) = \mathbb{P}(g(X) \neq Y)$

- Le coût moyen empirique associé à l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ est

$$\widehat{E}_n(g) = \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

  - $\rightsquigarrow$ On parle de risque moyen empirique
  - $\rightsquigarrow$ Lorsque la fonction de coût 0-1 est utilisée, on parle de taux d'erreur empirique et $\widehat{E}_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(g(X_i) \neq Y_i)$

Théorie de la décision statistique

Classification

- On cherche la règle de décision g qui minimise le risque moyen
- Obtenue en prenant pour chaque $x \in \mathcal{X}$

$$g^\star(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\min}\ \mathbb{E}\left[L(Y,\ell)|X=x\right]$$

  - $\rightsquigarrow$ Cette règle est appelée la règle de classification de Bayes
  - $\rightsquigarrow$ Le risque associé $E(g^\star)$ est appelé risque de Bayes
- Lorsque la fonction de coût 0-1 est utilisée cette règle s'écrit

$$g^\star(x) = \underset{\ell \in \{1,\dots,K\}}{\arg\max}\ \mathbb{P}(Y=\ell|X=x)$$

  - $\rightsquigarrow$ Le risque associé $E(g^\star)$ est appelé taux d'erreur de Bayes.

En pratique

Fonction de coût

- Soit $\hat{Y} = g(X) \in \mathcal{Y}$ la sortie prédite d'un individu et $Y \in \mathcal{Y}$ sa vraie valeur.
- Soit $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ une fonction de coût telle que $L(Y, \hat{Y})$ est le coût associé à la décision : "prédire $\hat{Y}$ alors que la vraie valeur est Y".
- En classification, $\mathcal{Y} = \{1, ..., K\}$ et la fonction de coût peut être résumée par une matrice de coûts C de taille $K \times K$ telle que $C_{k\ell} = L(k, \ell)$.

vraie classe \ classe prédite	$\ell = 1$	$\ell = 2$	...	$\ell = K$
$k = 1$	$C_{11} \leq 0$	$C_{12} \geq 0$	...	$C_{1K} \geq 0$
$k = 2$	$C_{21} \geq 0$	$C_{22} \leq 0$	...	$C_{2K} \geq 0$
...	⋮	⋮	⋱	⋮
$k = K$	$C_{K1} \geq 0$	$C_{K2} \geq 0$	...	$C_{KK} \leq 0$

- La fonction de coût se réecrit :

$$L(Y, \hat{Y}) = C_{Y\hat{Y}} = \sum_{k=1}^{K} \sum_{\ell=1}^{K} C_{k\ell} \mathbb{1}(Y = k \text{ et } \hat{Y} = \ell)$$

En pratique

Matrice de confusion

- Soit un échantillon $(Y_i, \hat{Y}_i)$, $i = 1, ..., n$ de couples classe réelle, classe prédite.
- La matrice de confusion comptabilise les occurences des prédictions en fonction des vraies valeurs

vraie classe \ classe prédite	$\ell = 1$	$\ell = 2$	...	$\ell = K$
$k = 1$	$p_{11} \times n$	$p_{12} \times n$	...	$p_{1K} \times n$
$k = 2$	$p_{21} \times n$	$p_{22} \times n$	...	$p_{2K} \times n$
...	⋮	⋮	⋱	⋮
$k = K$	$p_{K1} \times n$	$p_{K2} \times n$	...	$p_{KK} \times n$

où $p_{k\ell}$ est la proportion d'individus de classe k auxquels on a prédit la classe ℓ

- On a alors

$$p_{k\ell} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = \ell)$$

En pratique

Risque de prédiction

- Le risque théorique ou encore taux d'erreur théorique est l'espérance de la fonction de coût $E = \mathbb{E}\left[L(Y, \hat{Y})\right]$
- Il peut être approché par le risque empirique ou encore taux d'erreur empirique

$$\begin{aligned} \hat{E}_n &= \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{Y}_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{\ell=1}^{K} C_{kl} \mathbb{1}(Y_i = k \text{ et } \hat{Y}_i = \ell) \\ &= \sum_{k=1}^{K} \sum_{\ell=1}^{K} C_{k\ell} p_{k\ell} \end{aligned}$$

Il s'agit du coût moyen sur l'échantillon.

En pratique

Exemples de risques empiriques

- On prend l'exemple de la classification binaire
- $K = 2$ et 1=positif et 2=négatif
- On note $FP = p_{21} \times n$ le nombre de faux positifs
- On note $FN = p_{12} \times n$ le nombre de faux négatifs.
- Les deux fonctions de coût suivantes sont considérées.
  - La fonction de coût 0-1 avec $C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
  - La fonction de coût pondérée avec $C = \begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix}$

En pratique

Exemples de risques empiriques

- Fonction de coût 0-1
- La matrice de coût ne distingue pas les deux types d'erreurs FP et FN
- L'expression du coût moyen devient

$$\begin{aligned} \hat{E}_n &= \sum_{k=1}^{2}\sum_{\ell=1}^{2} C_{k\ell} p_{k\ell} \\ &= C_{11}p_{11} + C_{21}p_{21} + C_{12}p_{12} + C_{22}p_{22} \\ &= p_{21} + p_{12} \\ &= \frac{FP + FN}{n} \end{aligned}$$

ce qui correspond au taux d'erreur standard.

En pratique

Exemples de risques empiriques

- Fonction de coût pondérée
- On associe un coût cinq fois plus important aux faux positifs
- L'expression du coût moyen devient L'expression du coût moyen devient

$$\begin{aligned} \hat{E}_n &= C_{11}p_{11} + C_{21}p_{21} + C_{12}p_{12} + C_{22}p_{22} \\ &= 5 \times p_{21} + p_{12} \\ &= \frac{5 \times FP + FN}{n} \end{aligned}$$

ce qui correspond à un taux d'erreur pondéré.

En pratique

Règle de décision

- Soit $x \in \mathcal{X}$ la variable d'entrée
- La règle optimale de classification de Bayes est donnée par

$$\begin{aligned} g(x) &= \underset{\ell \in \{1,\dots,K\}}{\arg\min}\ \mathbb{E}\left[L(Y,\ell)|X=x\right] \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\min}\ \sum_{k=1}^{K} L(k,\ell)\mathbb{P}(Y=k|X=x) \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\min}\ \sum_{k=1}^{K} C_{k\ell}\,\mathbb{P}(Y=k|X=x) \quad \text{(approche discriminante)} \\ &= \underset{\ell \in \{1,\dots,K\}}{\arg\min}\ \sum_{k=1}^{K} C_{k\ell}\,f(x|Y=k)\mathbb{P}(Y=k) \quad \text{(modèle génératif)} \end{aligned}$$

où $f(x|Y=k)$ est la densité de X dans la classe k.

En pratique

Règle de décision

- Théorème de Bayes

$$\mathbb{P}(Y = k|X = x) = \frac{f(x|Y = k)\mathbb{P}(Y = k)}{\sum_{j=1}^{K} f(x|Y = j)\mathbb{P}(Y = j)}$$

- Cette règle de de classification de Bayes est applicable si l'on dispose
  - dans le cas discriminatif :
    - ⇝ des probabilité à posteriori $\mathbb{P}(Y = k|X = x)$
    - ⇝ exemple : régression logistique
  - dans le cas génératif :
    - ⇝ des fonctions de densité $f(x|Y = k)$ aussi notées $f_k(x)$
    - ⇝ des probabilités à priori $\mathbb{P}(Y = k)$ aussi notées π_k
    - ⇝ exemples : analyse discriminante, bayésien naïf

En pratique

Règle de décision binaire

- On considère le cas de la classification binaire
- $K = 2$ classes et la matrice de coûts est $C = \begin{pmatrix} 0 & C_{12} \\ C_{21} & 0 \end{pmatrix}$
- Deux cas sont à considérer :
  - Si la règle prédit $\ell = 1$:

$$\begin{aligned}\mathbb{E}\left[L(Y, \ell)|X = x\right] &= \sum_{k=1}^{2} C_{k1}\mathbb{P}(Y = k|X = x) \\ &= C_{21}\mathbb{P}(Y = 2|X = x)\end{aligned}$$

  - Si la règle prédit $\ell = 2$:

$$\begin{aligned}\mathbb{E}\left[L(Y, \ell)|X = x\right] &= \sum_{k=1}^{2} C_{k2}\mathrm{Pr}(Y = k|X = x) \\ &= C_{12}\mathrm{Pr}(Y = 1|X = x)\end{aligned}$$

En pratique

Règle de décision binaire

- On en déduit la règle de décision suivante :
  - cas discriminant : si $C_{21}\mathbb{P}(Y=2|X=x) \leq C_{12}\mathbb{P}(Y=1|X=x)$ i.e.

$$\mathbb{P}(Y=1|X=x) \geq \frac{C_{21}}{C_{12}}\mathbb{P}(Y=2|X=x)$$

    alors on affectera x à la classe $g(x)=1$, sinon on affectera x à la classe $g(x)=2$.

  - cas génératif : si $C_{21}f(x|Y=2)\mathbb{P}(Y=2) \leq C_{12}f(x|Y=1)\mathbb{P}(Y=1)$ i.e.

$$f(x|Y=1) \geq \frac{C_{21}\pi_2}{C_{12}\pi_1}f(x|Y=2)$$

    où $\mathbb{P}(Y=k)=\pi_k$, alors on affectera x à la classe $g(x)=1$, sinon on affectera x à la classe $g(x)=2$.

En pratique

Règle de décision binaire

- Si $C_{21}=C_{12}$ (fonction de coût 0-1) on retrouve la règle d'affectation de la classe la plus probable à posteriori
  - cas discriminant :

$$g(x)=\begin{cases}1 & \text{si } \mathbb{P}(Y=1|X=x) \geq \mathbb{P}(Y=2|X=x)\\ 2 & \text{sinon}\end{cases}$$

  - cas génératif :

$$g(x)=\begin{cases}1 & \text{si } \pi_1 f(x|Y=1) \geq \pi_2 f(x|Y=2)\\ 2 & \text{sinon}\end{cases}$$

Les méthodes de classification supervisées

Modèle ou Prototype

- Deux approches possibles pour définir la règle de classification $Y = g(X)$
- Approche basée sur un modèle
  - Apprentissage de $Loi(X, Y) = Loi(X|Y)Loi(Y)$
    - ⋆ Approche dite générative
    - ⋆ Exemples : Analyse discriminante linéaire, bayésien naïf
  - Apprentissage direct de $Loi(Y|X)$
    - ⋆ Approche dite discriminative
    - ⋆ Exemple : Régression logistique
- Approche de type prototype
  - Apprentissage direct de la règle de classification $Y = g(X)$
  - Exemple : K-plus proches voisins, arbres de décisions, forêts aléatoires, etc.

Les méthodes de classification supervisées

Paramétrique ou non

- Définir la règle de classification $Y = g(X)$ revient à définir pour tout $k \in \mathcal{Y}$ les probabilités à posteriori

$$\mathbb{P}[Y = k|X = x]$$

- Approche paramétrique
  - Régression logistique: $\mathcal{Y} = \{0, 1\}$ et

$$\mathbb{P}[Y = 1|X = x] = \frac{exp(X^T\beta)}{1 + exp(X^T\beta)}$$

  - Analyse discriminante:

$$\mathbb{P}(Y = k|X = x) = \frac{f(x|Y = k)\mathbb{P}(Y = k)}{\sum_{j=1}^{K} f(x|Y = j)\mathbb{P}(Y = j)}$$

et

$$f(x|Y = k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp(-\frac{1}{2}(x - \mu_k)^T\Sigma_k{}^{-1}(x - \mu_k))$$

$\rightsquigarrow$ Estimation des paramètres sur l'échantillon d'apprentissage

Les méthodes de classification supervisées

Paramétrique ou non

- Approche non paramétrique
  - Approche de type modèle avec estimation non paramétrique de $f(x|Y=k)$.
    Par exemple estimée de Parzen

$$f(x_0|Y=k) = \frac{1}{n\lambda}\sum_{i=1}^{n} K_\lambda(x_0, x_i)$$

    où K_λ est un noyau donné, e.g. gaussien 1D

$$K_\lambda(x_0, x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2\lambda^2}(x_0 - x)^2\right]$$

  - Approche de type prototype
    - ⋆ Apprentissage direct de la règle de classification $Y = g(X)$
    - ⋆ Exemple : K-plus proches voisins, arbres de décisions, forêts aléatoires, etc.
  - ⇝ Toujours un un paramètre à choisir.

Une méthode simple

k plus proches voisins

- Approche de type prototype
- Variables d'entrée réelles $X \in \mathbb{R}^p$
- règle de classification dans le cas d'une réponse binaire $Y \in \{0, 1\}$

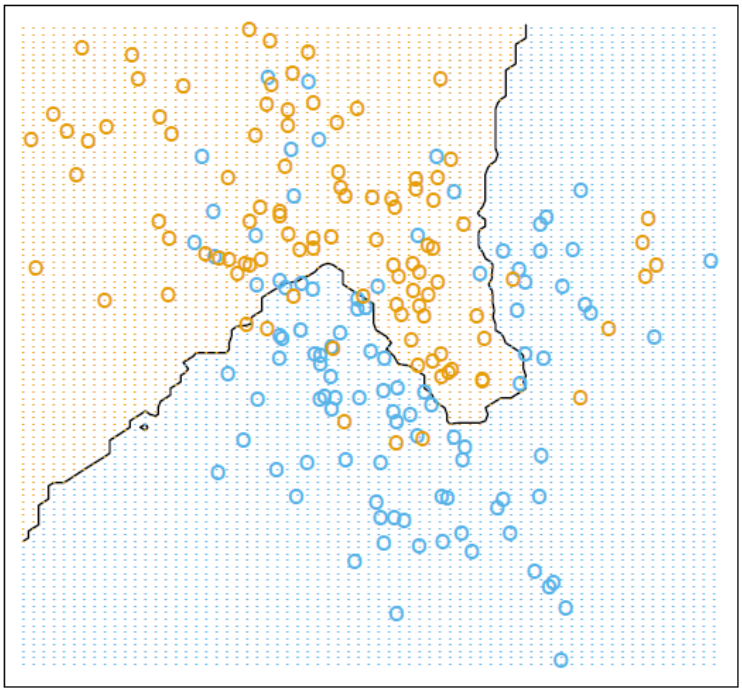
$$\begin{aligned} g(X) =& 1 \text{ si } \sum_{X_i \in N_k(X)} Y_i > \sum_{X_i \in N_k(X)} (1 - Y_i) \\ =& 0 \text{ sinon} \end{aligned}$$

  où $N_k(X)$ est un voisinage de X contenant uniquement les k plus proches voisins de l'ensemble d'apprentissage
- Vote à la majorité parmi les k plus proches voisins du point x à classer

Une méthode simple

k plus proches voisins

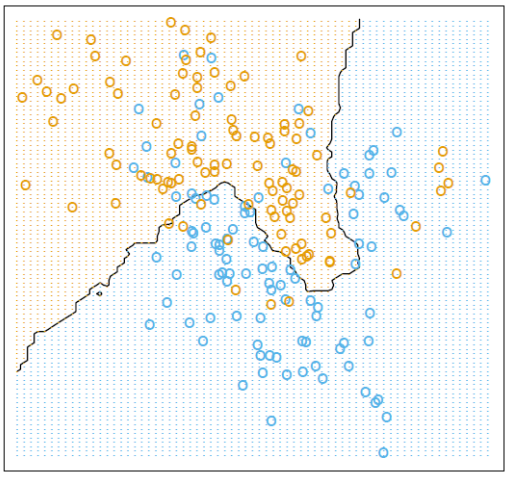
- Exemple avec $k = 15$, $p = 2$



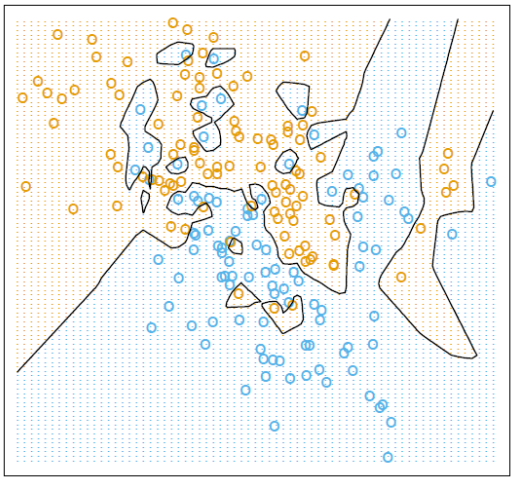
Une méthode simple

k plus proches voisins

k=15



k=1



Une méthode simple

k plus proches voisins

- Pour appliquer cette méthode il faut fixer un paramètre : le nombre de voisins k
- Une stratégie de choix du paramètre k
  - Choisir un échantillon d'apprentissage et un échantillon test
  - Faire varier k
  - Pour chaque valeur de k calculer le taux d'erreur empirique de prédiction des individus de l'échantillon test.

Une méthode simple

k plus proches voisins