Classifieurs linéaires

Guillaume Wisniewski
guillaume.wisniewski@limsi.fr

janvier 2016

Première partie I

Contexte / rappel

Classification supervisée

La tâche

- observations $\mathbf{x} \in \mathcal{X}$ associées à des étiquettes $y \in \mathcal{Y}$
- ensemble fini d'exemples : $\left(\mathbf{x}^{(i)}, y^{(i)}\right)_{i=1}^{N}$
- peut-on déterminer une fonction généralisant les exemples à l'ensemble des observations possibles $\mathcal{X}$

Représentation des exemples

- les exemples $\mathbf{x}$ sont représentés par des éléments de $\mathbb{R}^n$
- chaque dimension correspond à une caractéristiques (*feature*, attribut)
- les représentations sont choisies arbitrairement

Deuxième partie II

Intérêt

1) Interprétation graphique

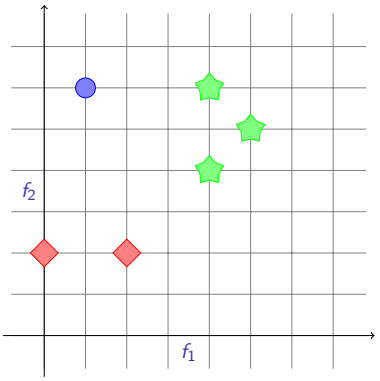
Principe

- les exemples peuvent être représentés dans un espace euclidien
  - chaque axe correspond à une caractéristique
  - exemple = point / vecteur

Exemple

- classe ω_1 : $\begin{pmatrix}0\\1\end{pmatrix}$, $\begin{pmatrix}1\\1\end{pmatrix}$
- classe ω_2 : $\begin{pmatrix}2\\2\end{pmatrix}$, $\begin{pmatrix}2\\3\end{pmatrix}$, $\begin{pmatrix}2,5\\2,5\end{pmatrix}$
- classe ω_3 : $\begin{pmatrix}0,5\\3\end{pmatrix}$

Représentation



2) Distance entre points

- espace euclidien $\Rightarrow$ on peut définir une « distance » entre point
- infinité de distances avec des propriétés différentes
- exemple :

$$||\mathbf{x} - \mathbf{x}'||_2 = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2} \qquad ||\mathbf{x} - \mathbf{x}'||_1 = \sum_{i=1}^{n} |x_i - x'_i|$$

- plus des points sont proches plus ils sont similaires

Conséquence

Comment classifier un point ?

- on attribut à l'exemple non étiqueté l'étiquette de l'exemple le plus proche
- plus « robuste » : regarder l'éttiquette des k plus proches voisins $\oplus$ vote majoritaire

k-ppv / k-nn

Autre méthode

Classifieur à distance minimale

1. on détermine un représentant de chaque classe, typiquement la moyenne :

$$\tilde{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}$$

2. on attribut l'étiquette de la classe dont le représentant est le plus proche

avantages / inconvénients par rapport aux k-ppv ?

Limite des classifieurs à base de distance

Complexité

- complexité des $k$ppv : $\mathcal{O}(N \times d)$ (N : nombre d'exemples, d : dimension)
- Problème pour les problèmes en grande dimension

Curse of dimensionnality

- quand la dimension augmente, les distances entre deux paires d'objets arbitraires tendent vers la même valeur :

$$\lim_{d \to \infty} \frac{\text{dist}_{\min} - \text{dist}_{\max}}{\text{dist}_{\min}} = 0$$

- intuition des $k$ppv « moins » valide

Troisième partie III

Principe des classifieurs linéaires

Avertissement

On se limitera pour aujourd'hui à la classification binaire : $y = \pm 1$

Rappel

Élément d'une approche d'apprentissage

1. des exemples étiquetés
   - ensemble de vecteurs $\oplus$ étiquettes
2. une mesure d'évaluation
3. une classe d'hypothèses $\oplus$ un biais inductif

Classe d'hypothèses

- règle de décision : comment choisir la sortie associée à une entrée
- méthode d'apprentissage : comment choisir un élément dans cette classe de fonctions (biais inductif)

Fonction de décision

Fonction de score

$$F(\mathbf{x}\,;\mathbf{w}) = \mathbf{w}\cdot\mathbf{x}$$

- rappel : la représentation inclus un terme constant
- fonction paramétrée par un vecteur $\mathbf{w}$
- combinaison linéaire des caractéristiques : chaque caractéristique a un poids par rapport à son importance

Fonction de décision

$$y^* = \text{sign}\, F(\mathbf{x};\mathbf{w}) = \begin{cases} 1 & \text{si } F(\mathbf{x};\mathbf{w}) \geq 0 \\ -1 & \text{sinon} \end{cases}$$

Frontière de décision

Définition

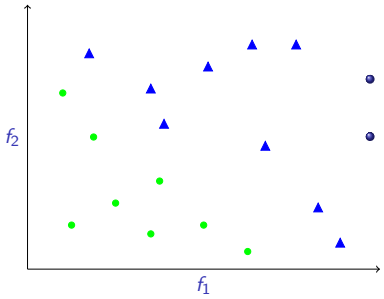
- caractéristique d'une classe de fonctions
- = quand le critère de décision « change » / quand on ne sait pas prendre de décision

Dans notre cas...

$$F(\mathbf{x},\mathbf{w}) = 0$$

$\Rightarrow$ équation d'un hyperplan = hyperplan séparateur

Graphiquement

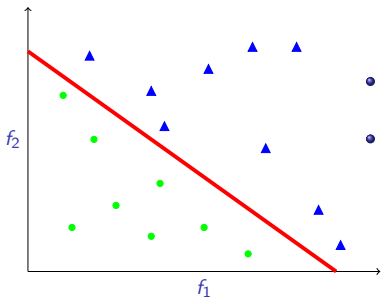


- en dimension 2 : $\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$
- frontière de décision :

$$\mathbf{x}\cdot\mathbf{w} = 0$$
$$x_1\cdot w_1 + x_2\cdot w_2 + w_0 = 0$$

$\Rightarrow$ équation d'une droite

Graphiquement

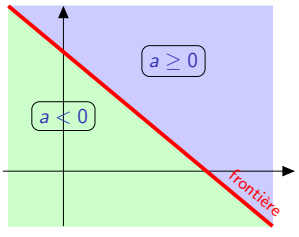


- en dimension 2 : $\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$
- frontière de décision :

$$\mathbf{x}\cdot\mathbf{w} = 0$$
$$x_1\cdot w_1 + x_2\cdot w_2 + w_0 = 0$$

$\Rightarrow$ équation d'une droite

Graphiquement



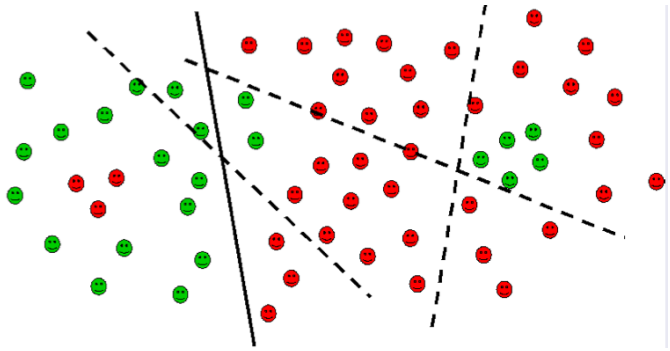
- en dimension 2 : $\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$
- frontière de décision :

$$\mathbf{x}\cdot\mathbf{w} = 0$$
$$x_1\cdot w_1 + x_2\cdot w_2 + w_0 = 0$$

$\Rightarrow$ équation d'une droite

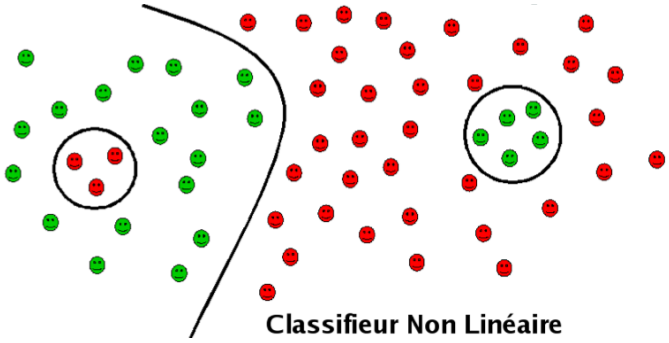
- fonction de décision

Pouvoir d'expression



Tous les ensembles de données ne sont pas linéairement séparables

Classification non linéaire



cf. cours sur la classification non linéaire

Apprentissage

Comment choisir le paramètre $\mathbf{w}$

Objectifs

- faire le moins d'erreurs sur des données nouvelles
- à choisir à partir d'un ensemble fini d'exemples (ensemble d'apprentissage)

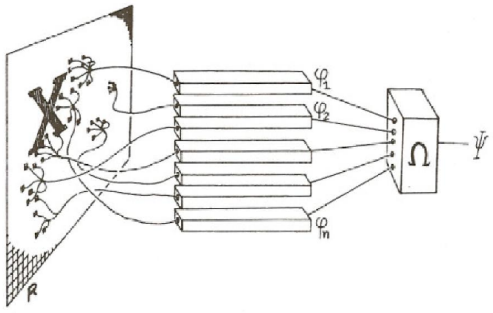
Méthodes

- perceptron
- modèle à maximum d'entropie (Maxent)
- machines à vecteurs de support (SVM)
- Winnow
- ...

Quatrième partie IV

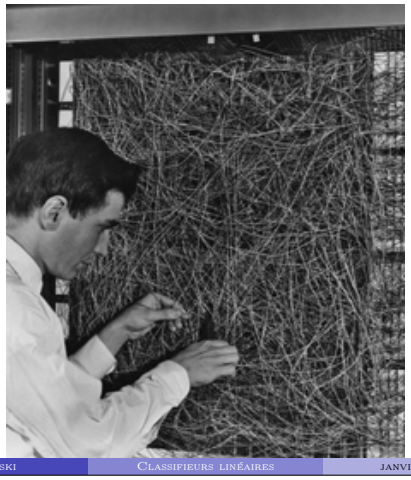
Le perceptron

Aspect historique (1)



- Rosenblatt, 1957
- objectif = reproduire les capacités visuelles d'un cerveau humain

Le perceptron est une machine !



Apprentissage

Apprentissage par correction d'erreurs (Rosenblatt, 57)

- pour chaque exemple de l'ensemble d'apprentissage
- classer l'exemple
- si la sortie est correcte : ne rien faire
- si la sortie est fausse : corriger les paramètres pour ne plus faire d'erreur

Plus formellement

Require: a training set $\left(\mathbf{x}^{(i)}, y^{(i)}\right)_{i=1}^{n}$ and a learning rate η

$\mathbf{w} \leftarrow 0$
while there are classification errors **do**
  for $i = 1$ to n **do**
    $y = \text{sign}\left(\mathbf{w} \cdot \mathbf{x}^{(i)}\right)$
    if $y \neq y^{(i)}$ **then**
      $\mathbf{w} \leftarrow \mathbf{w} + \eta \cdot y^{(i)} \cdot \mathbf{x}^{(i)}$
    end if
  end for
end while

Propriétés « théoriques »

- si l'ensemble d'apprentissage est linéairement séparable :
  - converge en un nombre fini d'opérations (ni $\eta = 1$)
  - aucune garantie en généralisation
- si l'ensemble d'apprentissage n'est pas linéairement séparable :
  - algorithme ne converge pas
  - plusieurs « astuces », aucune preuve
- comment savoir si l'ensemble d'apprentissage est linéairement séparable ? $\Rightarrow$ réponse uniquement empirique