Classifieur linéaire

Introduction à l'apprentissage automatique
Master Sciences Cognitives
Aix Marseille Université

Alexis Nasr

Motivations

- Les modèles d'apprentissage sont souvent appris sur un **grand** nombre de données de **dimensions** importantes.
- Cela rend difficile la **visualisation** des données et l'**interprétation** des modèles.
- Nous verrons ici un exemple de **classification** simplifié à l'extrême (classification binaire en dimension 2) pour lequel un solution peut être trouvée géométriquement.
- Cet exemple permettra d'introduire des notions fondamentales qui restent valables pour des problèmes complexes.
- Nous verrons plus tard comment les opérations que nous effectuerons à la main ici peuvent être réalisées **automatiquement** par un ordinateur.

Objectifs

- Introduire la notion de **séparabilité** des données.
- Distinguer les données **linéairement séparables** des données non linéairement séparables en dimension 2
- Introduire la notion de **classifieur linéaire**.
- Introduire la notion de **probabilité d'appartenir à une classe**.

Langue d'un document

- On dispose de documents en anglais et en français et on désire construire un modèle qui, étant donné un nouveau document, permet de **prédire** s'il est en français ou en anglais.
- On collecte 20 documents pour lesquels **on connaît la langue** et, pour chacun d'entre eux, on calcule la fréquence des deux lettres u et w.
- Nos données d'apprentissage se présentent donc sous la forme suivante :

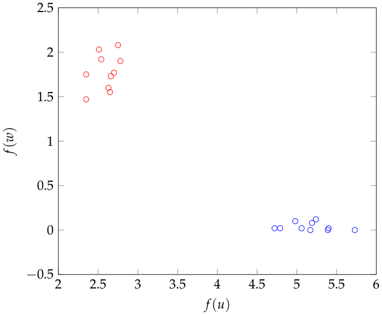
$$\mathcal{D} = \{((f_i(u), f_i(w)), c_i)\}_{i=1}^{20}$$

  - $f_i(u)$: fréquence de la lettre u dans le document i,
  - $f_i(w)$: fréquence de la lettre v,
  - c_i la langue du document.
- Les données ont été collectés sur des document de 1000 mots.

Représentation tabulaire

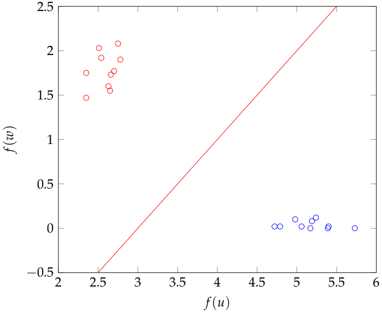
i	x_1 $f_i(u)$	x_2 $f_i(w)$	y c_i	i	x_1 $f_i(u)$	x_2 $f_i(w)$	y c_i
1	4.79	0.02	fr	11	2.78	1.90	en
2	4.98	0.10	fr	12	2.51	2.03	en
3	5.24	0.12	fr	13	2.63	1.60	en
4	5.73	0.00	fr	14	2.75	2.08	en
5	4.72	0.02	fr	15	2.54	1.92	en
6	5.39	0.00	fr	16	2.65	1.55	en
7	5.19	0.08	fr	17	2.35	1.75	en
8	5.17	0.00	fr	18	2.70	1.77	en
9	5.06	0.02	fr	19	2.66	1.73	en
10	5.40	0.02	fr	20	2.35	1.47	en

Représentation dans le plan



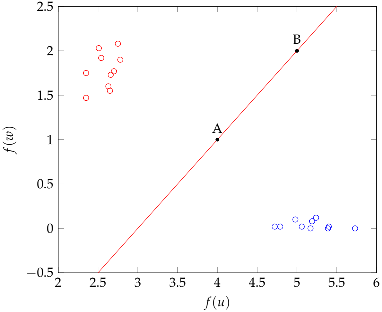
- Un point représente un document
  - Les documents en français sont en bleu
  - Les documents en anglais sont en rouge

Séparabilité



- Problème de classification **simple**
- Il est facile de trouver une droite D qui **sépare** les données :
  - les documents en français se situent d'un côté de la droite
  - les documents en anglais se situent de l'autre côté.
- Les données sont dites **linéairement séparables**.

Droite de séparation



- Pour construire D, il suffit de choisir deux points dans le plan par lesquels passe D.
- On peut choisir le point $A = (4, 1)$ et $B = (5, 2)$.

Equation d'une droite

- Une équation de droite est une égalité caractérisant tous les points d'une même droite.
- Toute droite du plan a une équation d'inconnues x et y du type

$$ax + by + c = 0$$

  appelée **équation cartésienne** de la droite (où a, b et c sont des nombres réels).
- Un point p de coordonnées (x_p, y_p) appartient à la droite si et seulement si ses coordonnées vérifient l'équation de droite.
- Toute droite non parallèle à l'axe des ordonnées admet une équation dite **réduite** du type

$$y = mx + p$$

  m et p sont des réels.
- m est appelé **coefficient directeur**, et p **ordonnée à l'origine**.
- Toute droite admet une seule équation réduite mais une infinité d'équations cartésiennes.

Détermination de l'équation d'une droite à partir de deux points

Si A (x_A, y_A) et B (x_B, y_B) sont deux points d'abscisses différentes, alors la droite (AB)

- a pour coefficient directeur :

$$m = \frac{y_B - y_A}{x_B - x_A}$$

- et pour ordonnée à l'origine :

$$p = y_A - m x_A$$

Détermination des paramètres de D

- $A = (4, 1)$ et $B = (5, 2)$
- $m = \frac{y_B - y_A}{x_B - x_A} = \frac{1}{1} = 1$
- $p = y_A - m x_A = 1 - 4 = -3$
- D admet pour équation réduite :

$$y = x - 3$$

- et pour équation cartésienne :

$$y - x + 3 = 0$$

Position d'un point par rapport à une droite

Pour savoir si un point $p = (x_p, y_p)$ se trouve au dessus ou en dessous de la droite d'équation $ax + by + c = 0$, il suffit de calculer $d = ax_p + by_p + c$.

- si $d < 0$ alors p se trouve sous la droite
- si $d = 0$ alors p se trouve sur la droite
- si $d > 0$ alors p se trouve au dessus la droite

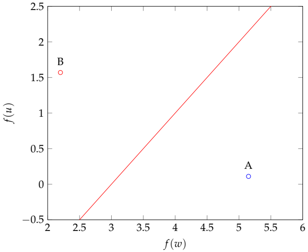
Prédiction

- Etant donné un nouveau document correspondant au point (x, y), si ce point se situe au dessus de la droite alors il s'agit d'un document en anglais, sinon, il s'agit d'un document en français.
- Le classifieur s'écrit donc ainsi : $signe(y - x + 3)$ où signe est la fonction :

$$signe(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{sinon} \end{cases}$$

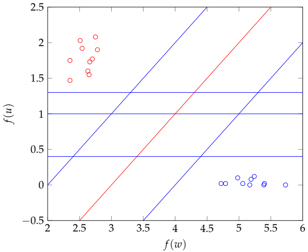
- où 1 est la classe correspondant à l'anglais et -1 la classe correspondant au français.

Prédicitions



- Nouveaux exemples :
  - $A = ((5.15, 0.11), fr)$
  - $B = ((2.20, 1.57), en)$
- On obtient dans le premier cas :
  - $signe(0.11 - 5.15 + 3) = signe(-2.04) = -1$
  - le modèle prédit qu'il s'agit d'un exemple français
- dans le second cas :
  - $signe(1.57 - 2.20 + 3) = signe(2.3) = 1$
  - le modèle prédit qu'il s'agit d'un exemple anglais.

Remarque 1



- Dans l'exemple précédent, une droite particulière a été choisie pour séparer les données,
- mais il existe une infinité de droites vérifiant cette propriété.
- la droite choisie possède une propriété intéressante, elle se trouve à peu près à mi-chemin des deux nuages de points.

Remarque 2

- La visualisation que nous avons utilisée était possible car nous n'avions que **deux features** en entrée :
  - x_1 : la fréquence de la lettre u
  - x_2 : la fréquence de la lettre w
- On a d'habitude plus de deux dimensions et il est difficile, dans ces conditions de visualiser les données.

Remarque 3

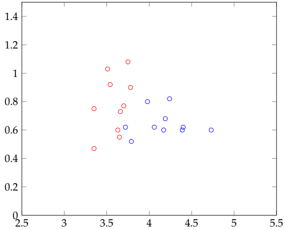
- Il est important de bien distinguer les deux étapes de l'exemple précédent.
- Dans une première étape, on a utilisé les données pour apprendre les paramètres θ du modèle (ici les coefficients a et b), il s'agit de l'étape d'**apprentissage** :

$$(x, y) \Rightarrow \theta$$

- Lors de la seconde étape, on a utilisé le modèle pour faire de la prédiction, il s'agit de l'étape d'**inférence** (qu'on appelle aussi décodage ou prédiction).

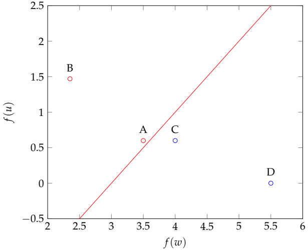
$$(x, \theta) \Rightarrow y$$

Remarque 4



- Dans l'exemple précédent, les données étaient linéairement séparables, mais cela aurait pu ne pas être le cas.
- On ne peut, dans un tel cas, aboutir à un classifieur linéaire qui permette une séparation parfaite des exemples des deux classes.
- Plusieurs choix sont alors possibles :
  - Modifier la représentation des exemples.
  - Recourir à un classifieur non linéaire
  - Autoriser un certain taux de mauvaise classification

Probabilité d'appartenir à une classe



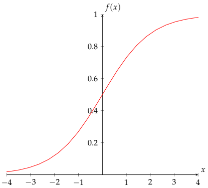
- La fonction de classification $f(x)$ prend ses valeurs dans l'intervalle $]-\infty,+\infty[$.
- que l'on réduit aux deux valeurs $\{0,1\}$ correspondant aux deux classes à l'aide de la fonction signe :

$$signe(ax + by + c)$$

- On peut aussi s'intéresser à la **confiance** de la décision prise ou la **probabilité** d'appartenance à une classe.
- On est plus sûrs de nous lorsqu'on classe B et D que lorsqu'on classe A et C.

Probabilité d'appartenir à une classe

- Pour cela on se ramène à l'intervalle $[0, 1]$ en utilisant une fonction écrasante ou *smashing function*.
- On peut utiliser par exemple la fonction sigmoïde $\sigma(x) = \frac{1}{1+e^{-x}}$



- Le classifieur devient : $\sigma(f(x)) = \frac{1}{1+e^{-(ax+by+c)}}$
- On peut interpréter la valeur calculée par le classifieur comme une probabilité :

$$\sigma(f(x)) = P(\hat{y} = 1|x)$$

  c'est la probabilité que x appartienne à la classe positive
- Plus la valeur calculée est proche de 0 ou de 1 plus on est sûr de notre choix.

Sources

- Tom Mitchell, *Machine Learning*, McGraw Hill, 1997.
- Yoav Goldberg, *Neural Network Methods for Natural Language Processing*, Morgan & Claypool, 2017.