Régression logistique

Alexandre Allauzen

Université Paris-Sud / LIMSI-CNRS

Plan du TER

Cours

- Régression logistique, le neurone artificiel
  - Classification binaire de texte
  - Modèle probabiliste
  - Fonction objectif
  - Algorithme d'apprentissage
- Représentation de mots
  - Un premier réseau de neurones
  - Word embeddings
- Réprésentation des séquences de mots
  - Réseau récurrents
  - Apprentissage

TP

- Prise en main
  - python, numpy, matplotlib
  - les textes et les données
- Régression logistique
  - Introduction à KERAS
  - Apprentissage, évaluation
- Réseau de neurones (feed-forward)
  - couche cachée
  - word embeddings
- Réseaux récurrents
- Projets

Analyse d'opinion dans les textes

Les applications

- Online customer reviews
- Advertisement targeting
- Public relations/marketing
- Analytics/reputation mining
- Web content filtering ...

Cas d'étude: la critique/review de film

à partir d'un texte, l'avis est-il positif / négatif

Quelques difficultés

La polarité d'un mot est contextuelle

- The movie was unpredictable
- The car steering is unpredictable

Des expressions ...

How can anyone sit through this movie?

Différents types d'information

My wonderful boyfriend took me to see this movie for our anniversary. It was terrible.

Quand le négatif est positif

The slow, methodical way he spoke. I loved it! It made him seem more arrogant and even more evil.

Plan

Apprentissage automatique : classification binaire

Soit un ensemble de données d'apprentissage :

$$\mathcal{D} = (\boldsymbol{x}_i, \tilde{y}_i)_{i=1}^{n}$$

Choix d'un modèle

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$$

$\boldsymbol{\theta}$: ensemble des paramètres qui définissent la fonction de décision

Apprentissage

- Déterminer $\boldsymbol{\theta}$ grâce à $\mathcal{D}$,
- Minimisation d'une fonction de pertes (*loss function*) mesurée sur $\mathcal{D}$

Exemple de classification binaire

But:

Prédire si un étudiant valide ($y = 1$) ou non ($y = 0$) connaissant $\boldsymbol{x}$ Un étudiant = 2 notes(examen et C.C) $\Rightarrow \boldsymbol{x} = (x_1, x_2)$

Un modèle simple

La validation du module :

$$y = 1 \text{ si } w_1 x_1 + w_2 x_2 > \alpha \text{ sinon } y = 0$$
$$y = 1 \text{ si } w_0 + w_1 x_1 + w_2 x_2 > 0 \text{ sinon } y = 0$$
$$y = 1 \text{ si } w_0 + \mathbf{w}^T \boldsymbol{x} > 0 \text{ sinon } y = 0$$

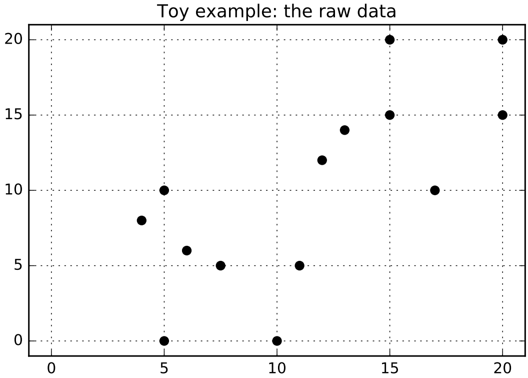
Apprentissage

Déterminer $\boldsymbol{\theta} = (w_0, \mathbf{w})$ à partir d'un ensemble d'apprentissage $\mathcal{D} = (\boldsymbol{x}_i, \tilde{y}_i)_{i=1}^n$
$\tilde{y}$ la réponse attendue, la référence à reproduire

Exemple de classification binaire

Les données

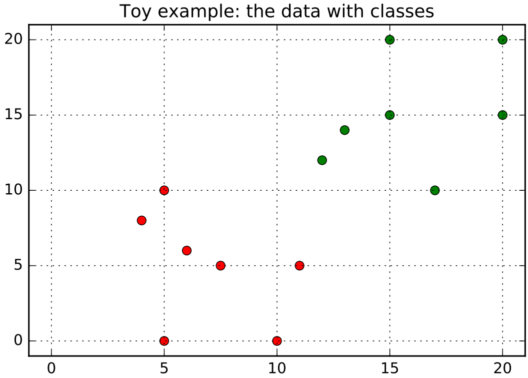
$$\mathcal{D} = \begin{pmatrix} x_1 = & 17 & 12 & 13 & 15 & 15 & 20 & 20 & 4 & 7.5 & 10 & 11 & 5 & 5 & 6 \\ x_2 = & 10 & 12 & 14 & 15 & 20 & 15 & 20 & 8 & 5 & 0 & 5 & 0 & 10 & 6 \\ y = & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



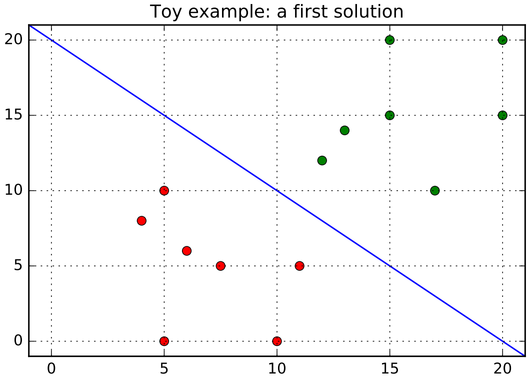
Exemple de classification binaire

Les données

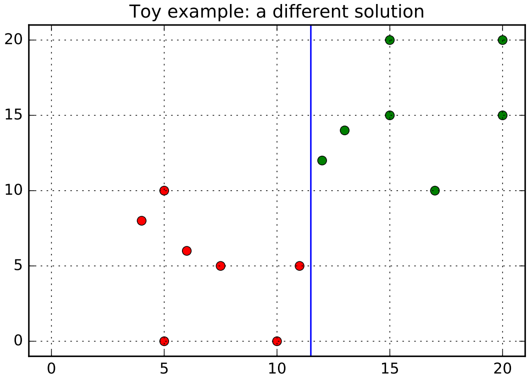
$$\mathcal{D} = \begin{pmatrix} x_1 = & 17 & 12 & 13 & 15 & 15 & 20 & 20 & 4 & 7.5 & 10 & 11 & 5 & 5 & 6 \\ x_2 = & 10 & 12 & 14 & 15 & 20 & 15 & 20 & 8 & 5 & 0 & 5 & 0 & 10 & 6 \\ y = & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



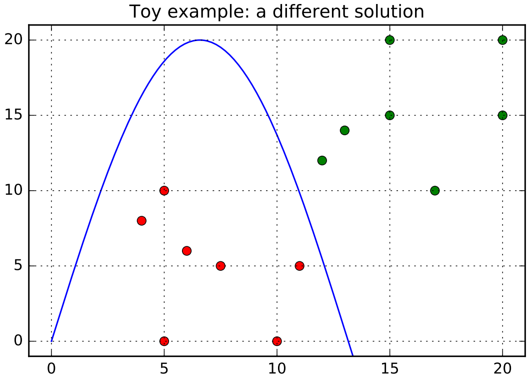
Classification binaire ou séparation



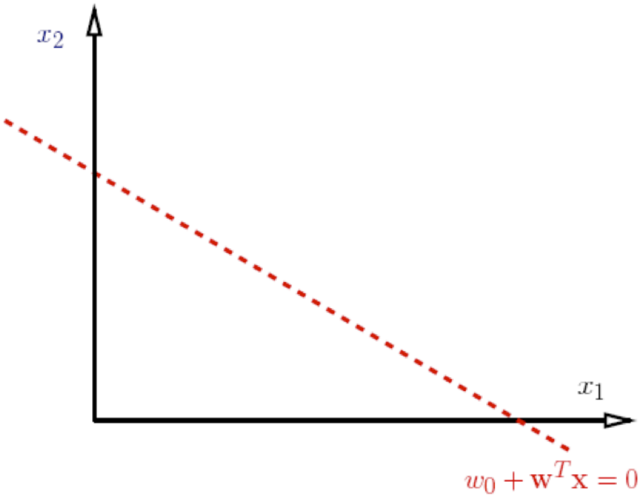
Classification binaire ou séparation



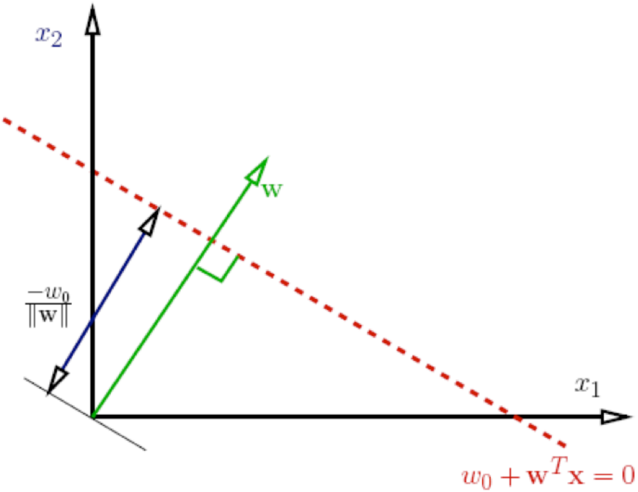
Classification binaire ou séparation



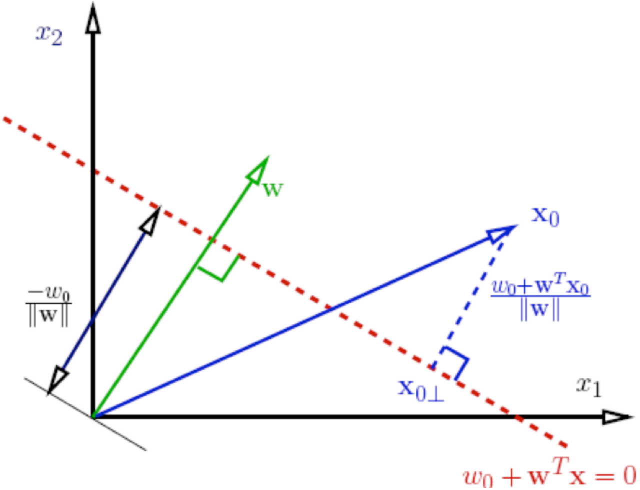
Séparation linéaire: un peu de géométrie



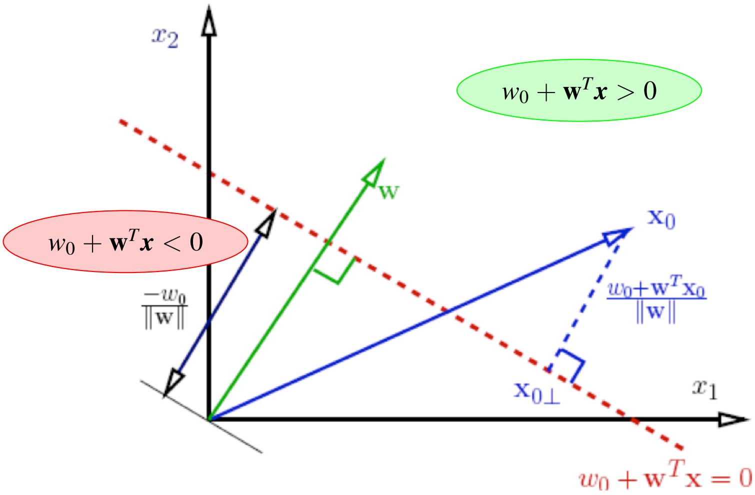
Séparation linéaire: un peu de géométrie



Séparation linéaire: un peu de géométrie



Séparation linéaire: un peu de géométrie



Représentation du réel

Le sac de mots

Comment représenter un texte ?

this movie is just great , with a great music , while a bit long

vocabulaire	sac binaire	compte	tfidf	...
the	0	0	0.01	...
awesome	0	0	1.2	...
this	1	1	0.1	...
long	1	1	2.5	...
great	1	2	0.9	...
...	...	...	...	...

Représentation en sac binaire

Le texte : ***this*** *movie is just* ***great*** *, with a* ***great*** *music , while a bit* ***long***
Le vocabulare : (the, **this**, awesome, **long**, **great**)

$$\Rightarrow \boldsymbol{x} = \begin{array}{c} \\ \end{array}\begin{pmatrix} 0 \\ \mathbf{1} \\ 0 \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \begin{array}{l} \textit{the} \\ \textbf{this} \\ \textit{awesome} \\ \textbf{long} \\ \textbf{great} \end{array}$$

Séparation linéaire et sac binaire

$$w_0 + \mathbf{w}^T \boldsymbol{x} = w_0 + \begin{pmatrix} 0 \\ \mathbf{1} \\ 0 \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{pmatrix} = w_0 + w_2 + w_4 + w_5$$

La classe est positive ($y = 1$) si

$$\begin{aligned} w_0 + w_2 + w_4 + w_5 &> 0 \\ w_2 + w_4 + w_5 &> -w_0 \\ w_{this} + w_{long} + w_{great} &> \text{seuil} \end{aligned}$$

Modèles de séparation linéaire

Comment calculer la frontière de séparation (critère d'optimisation):

- Perceptron
- SVM
- Bayésien Naïf
- ...

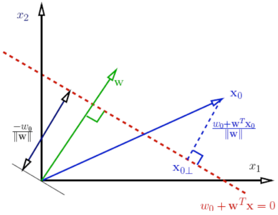
Régression logistique

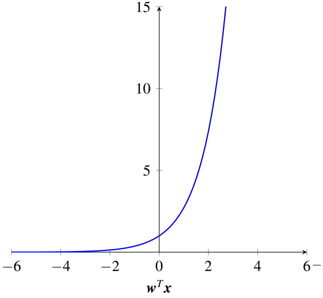
- Critère d'apprentissage probabiliste
- Interprétation probabiliste
- Le réseau de neurone le plus simple: un neurone

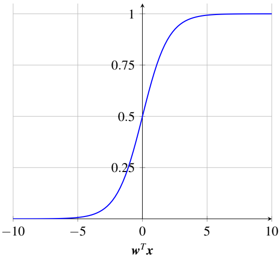
Plan

Interprétation probabiliste

$$-\infty < w_0 + \mathbf{w}^T\boldsymbol{x} < +\infty \qquad 0 < e^{w_0+\mathbf{w}^T\boldsymbol{x}} < +\infty \qquad 0 < \frac{e^{w_0+\mathbf{w}^T\boldsymbol{x}}}{1+e^{w_0+\mathbf{w}^T\boldsymbol{x}}} < 1$$







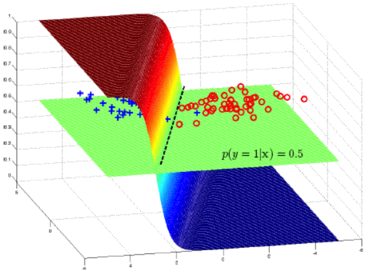
Régression logistique

Fonction logistique (sigmoïde)

$$P(y=1|\boldsymbol{x}) = \sigma(w_0 + \mathbf{w}^T\boldsymbol{x})$$

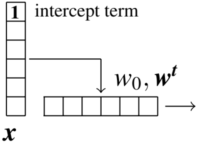
$$\sigma(a) = \frac{e^a}{1+e^a} = \frac{1}{1+e^{-a}}$$

$$a = w_0 + \mathbf{w}^T\boldsymbol{x} \in \mathbb{R}$$

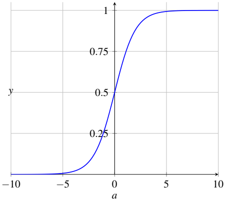


Régression logistique et neurone artificiel

Régression logistique



$\longrightarrow \quad y = \sigma(w_0 + \boldsymbol{w}^T\boldsymbol{x})$



$$\sigma(a = w_0 + \boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + e^{-a}}$$

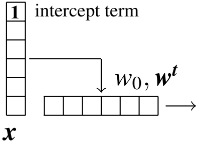
Le neurone artificiel

pre-activation : $a = w_0 + \boldsymbol{w}^T\boldsymbol{x}$

$y = \sigma(a)$, σ : la fonction d'activation

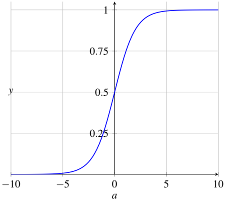
Régression logistique et neurone artificiel

Régression logistique

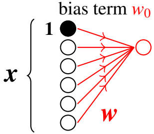


$$y = \sigma(w_0 + \boldsymbol{w}^T\boldsymbol{x})$$

$$\sigma(a = w_0 + \boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1 + e^{-a}}$$



Le neurone artificiel



pre-activation : $a = w_0 + \boldsymbol{w}^T\boldsymbol{x}$

$y = \sigma(a)$, σ : la fonction d'activation

Plan

Rappel: Loi de Bernouilli

Soit une variable aléatoire binaire y.

Loi de Bernouilli de paramètre π

$$\begin{aligned} P(y = 1|\pi) &= \pi \\ P(y = 0|\pi) &= 1 - \pi \\ P(y|\pi) &= \pi^y(1-\pi)^{1-y} \end{aligned}$$

Pour une série de tirage dont les résultats i.i.d sont : $(y_1, ..., y_n)$

$$P((y_1, ..., y_n)|\pi) = \prod_{i=1}^{n} P(y_i|\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

"Vraisemblance" des données $\mathcal{D}$

Soient $\mathcal{D} = (\boldsymbol{x}_i, \tilde{y}_i)_{i=1}^n = (\mathcal{X}, \tilde{\mathcal{Y}})$ et un modèle de regression logistique, $\boldsymbol{\theta} = (w_0, \mathbf{w})$.

$$P(\tilde{\mathcal{Y}}|\mathcal{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} P(\tilde{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$$

$$P(\tilde{\mathcal{Y}}|\mathcal{X}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_i^{\tilde{y}_i}(1 - \pi_i)^{1-\tilde{y}_i}$$

$$\pi_i = \sigma(w_0 + \mathbf{w}^T \boldsymbol{x}_i) = \frac{1}{1 + e^{-(w_0 + \mathbf{w}^T \boldsymbol{x}_i)}}$$

Fonction de perte à optimiser (minimiser)

$$\mathcal{L}(\boldsymbol{\theta}) = -log(P(\tilde{\mathcal{Y}}|\mathcal{X}, \boldsymbol{\theta})) = -\Big(\sum_{i=1}^{n} \tilde{y}_i log(\pi_i) + (1 - \tilde{y}_i) log(1 - \pi_i)\Big)$$