La régression logistique

Dr: Hadjadj Abdelhalim

Centre universitaire de Mila
Master I: Apprentissage automatique

Introduction

- **Importance de la classification** : Environ 70 % des problèmes en science des données concernent la classification.
- **Types de classification** :
  - **Classification binaire** : Deux classes possibles (ex. : spam ou non spam).
  - **Classification multinomiale** : Plusieurs classes possibles (ex. : classification des fleurs dans l'ensemble de données IRIS).

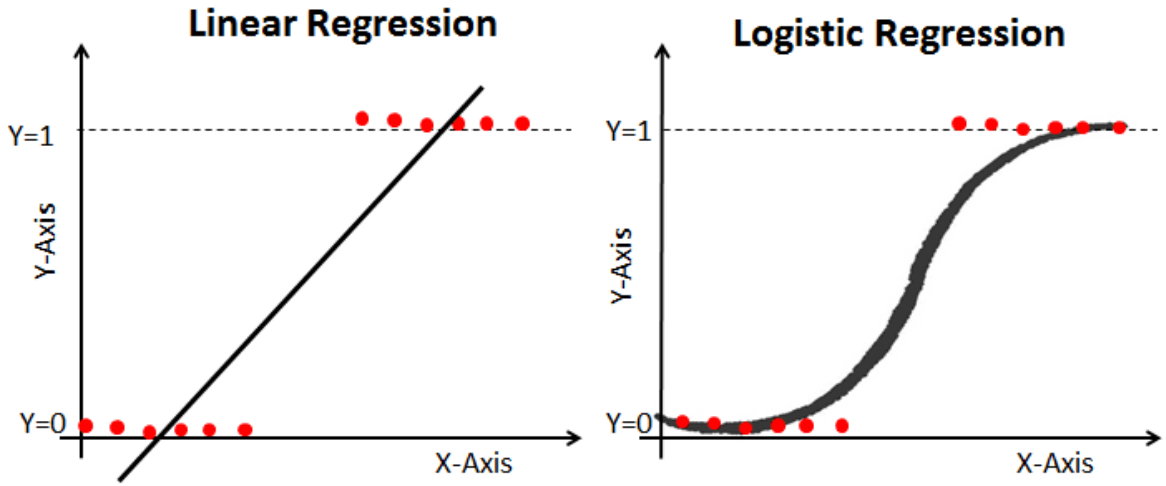
Introduction

❖Définition de la régression logistique

- ❖La régression logistique est une technique utilisée pour résoudre des problèmes de classification binaire(deux classes possibles).
- ❖la régression logistique prédit une probabilité associée à une classe donnée.

❖Lien avec la régression linéaire

- ❖Cas particulier de la **régression linéaire** où la variable cible est **catégorielle**.
- ❖Utilise le **logarithme des chances (logit)** comme variable dépendante.
- ❖Prédit la probabilité d'un événement binaire à l'aide de la **fonction logit**.



Principe de la régression logistique

- Supposons une classification binaire avec C = $\{C_1, C_2\}$.
- Une donnée avec D attributs $\mathbf{x} = (x_1,\ldots,x_D) \in \mathbb{R}D$, on aura:

$$p(C_1|\mathbf{x} = x) = \frac{p(\mathbf{x} = x|C_1)p(C_1)}{p(\mathbf{x} = x|C_1)p(C_1) + p(\mathbf{x} = x|C_2)p(C_2)}$$

- après la division par $p(x/C_1)p(C_1)$ nous avons obtenu :

$$= \frac{1}{1 + \frac{p(\mathbf{x} = x|C_2)p(C_2)}{p(\mathbf{x} = x|C_1)p(C_1)}} \qquad = \frac{1}{1 + \exp\left(-\ln\left(\frac{p(\mathbf{x} = x|C_1)}{p(\mathbf{x} = x|C_2)}\right) - \ln\left(\frac{p(C_1)}{p(C_2)}\right)\right)}$$

Principe de la régression logistique

- **Hypothèse de Bayes Naïf (HBN)**
- Les attributs x_d sont **indépendants** conditionnellement à la classe C_k.
- Pour chaque classe C_k, chaque attribut x_d suit une loi normale
- Chaque attribut x_d a la même variance σ_d dans les 2 classes.

$$p(x_d = x_d | C_k) = \frac{1}{\sigma_d \sqrt{2\pi}} \exp\left(-\frac{(x_d - \mu_{kd})^2}{2{\sigma_d}^2}\right)$$

Principe de la régression logistique

- On aura alors: $p(C_1|\mathbf{x} = x) =$

$$\frac{1}{1 + \exp\left(-\ln\left(\frac{p(C_1)}{p(C_2)}\right) - \sum_{d=1}^{D}\left(\frac{\mu_{1d} - \mu_{2d}}{{\sigma_d}^2} x_d + \frac{{\mu_{2d}}^2 - {\mu_{1d}}^2}{2{\sigma_d}^2}\right)\right)}$$

$$= \frac{1}{1 + \exp\left(-\ln\left(\frac{p(\mathbf{x} = x|C_1)}{p(\mathbf{x} = x|C_2)}\right) - \ln\left(\frac{p(C_1)}{p(C_2)}\right)\right)}$$

$$p(\mathrm{x}_d = x_d|C_k) = \frac{1}{\sigma_d\sqrt{2\pi}} \exp\left(-\frac{(x_d - \mu_{kd})^2}{2{\sigma_d}^2}\right)$$

- En posant:

$$\left(\frac{\mu_{1d} - \mu_{2d}}{{\sigma_d}^2}\right) = w_d \quad \text{et} \quad \sum_{d=1}^{D}\left(\frac{{\mu_{2d}}^2 - {\mu_{1d}}^2}{2{\sigma_d}^2}\right) + \ln\left(\frac{p(C_1)}{p(C_2)}\right) = w_0$$

$$= \frac{1}{1 + \exp(-f(x))} \quad \text{où} \quad f(x) = w_0 + \sum_{d=1}^{D}(w_d x_d)$$

Principe de la régression logistique

- Plus généralement, lorsque les classes ne suivent pas une distribution gaussienne, on adopte une approche plus flexible en introduisant des paramètres généraux w_0 et w , de sorte que la probabilité a posteriori de la classe C_1 donnée une observation x soit exprimée comme :

$$p(C_1|\mathbf{x} = x) = \frac{1}{1 + \exp(-f(x))}$$

- on en déduit que :

$$p(C_2|\mathbf{x}) = \frac{\exp(-f(\mathbf{x}))}{1 + \exp(-f(\mathbf{x}))}$$

Principe de la régression logistique

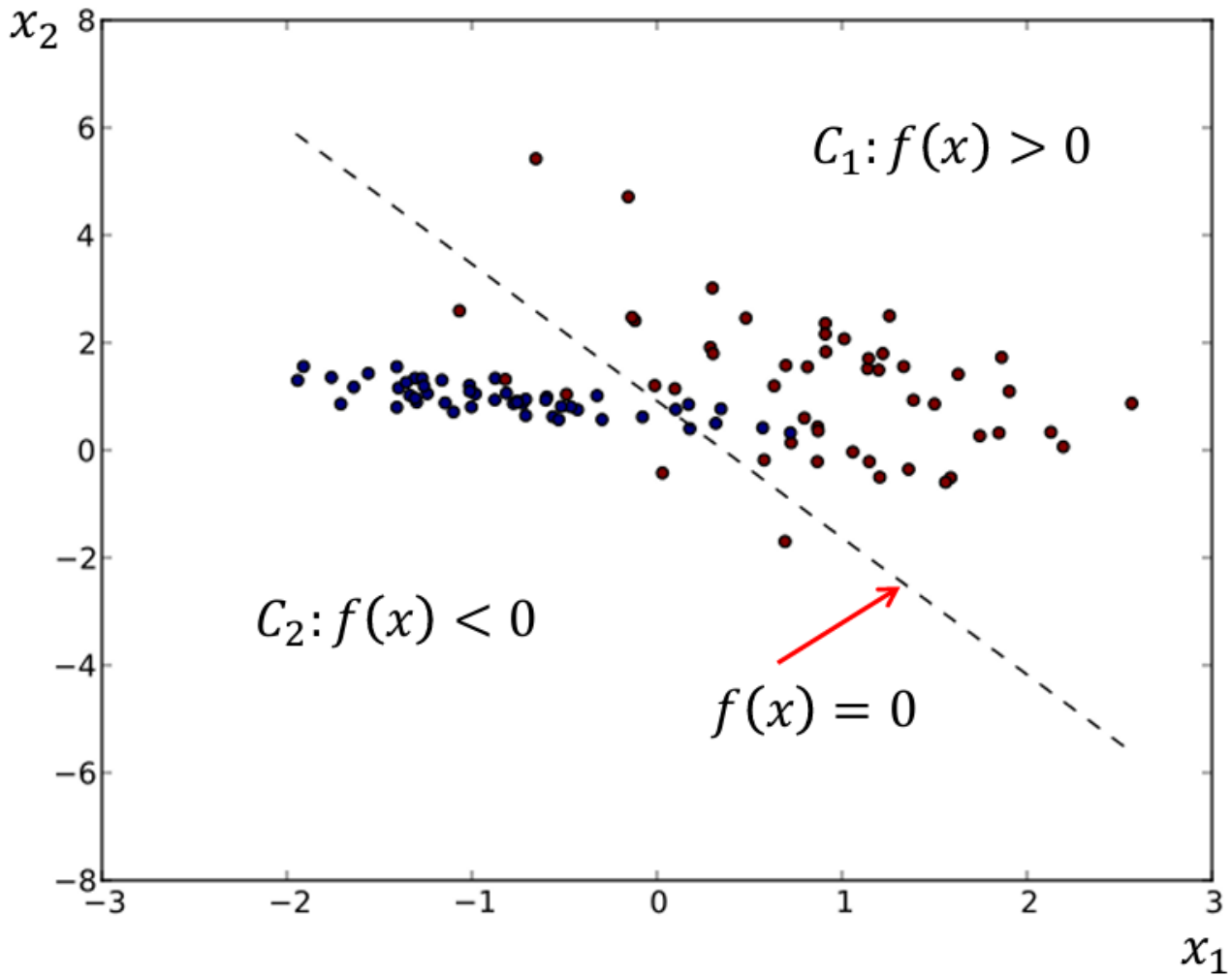
- Dans ce cas, on note que :
- Si f(x)=0, alors les probabilités a posteriori des deux classes sont égales :

$$p(C_1/x)=p(C_2/x)=0.5$$

- De plus, lorsque f(x)≠0, on peut mesurer le rapport des probabilités a posteriori sous forme logarithmique :

$$\ln\left(\frac{p(C_1|\mathbf{x}=x)}{p(C_2|\mathbf{x}=x)}\right)=f(x)=w_0+\sum_{d=1}^{D}(w_d x_d) \quad \begin{cases} > 0? \\ \\ < 0? \end{cases}$$

la régression logistique (frontière de décision)



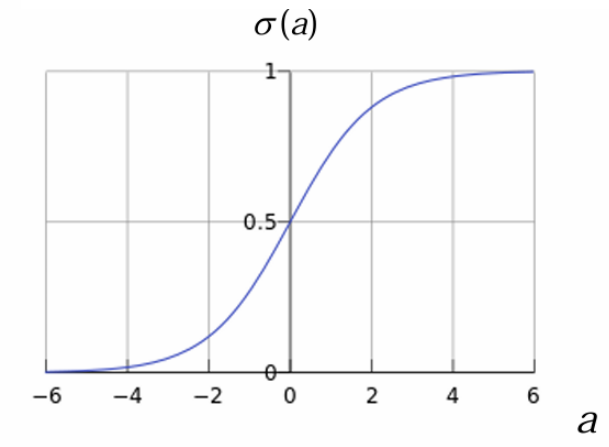
$$p(C_1|\mathbf{x} = x) = \frac{1}{1 + \exp(-f(x))}$$

$$p(C_2|\mathbf{x}) = \frac{\exp(-f(\mathbf{x}))}{1 + \exp(-f(\mathbf{x}))}$$

Fonction sigmoïde

- **Transforme une valeur réelle** en une probabilité entre **0 et 1**.
- Courbe en **forme de "S"** :
  - À l'infini positif → **sortie = 1**.
  - À l'infini négatif → **sortie = 0**.
- Seuil à **0,5** :
  - Si **sortie ≥ 0,5** → classé comme **1 (OUI)**.
  - Si **sortie < 0,5** → classé comme **0 (NON)**.

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



- Ex. : **Si la sortie est 0,75**, alors **75 % de chances** qu'un patient ait un cancer.
- Quand a varie de $-\infty$ à $+\infty$, $\sigma(a)$ variera de 0 à 1

Exemple

$e = 2.718$

x	z	e^{-z}	$1 + e^{-(z)}$	g(z)	y
3	0	e^0 1	2	0.5	1
2	2	$\frac{1}{e^2}$ 0.13536335	1.13536335	0.8808	1
1	4	$\frac{1}{e^4}$ 0.018323237	1.018323237	0.9820	1
5	-4	e^4 54.57551085	55.57551085	0.0179	0
4	-2	e^2 7.387524	8.387524	0.1192	0
6	-6	e^6 403.1778962	404.1778962	0.00247	0

f(x)=-2x+6

z=f(x)

$g(z)= \frac{1}{1+e-^z}$

Estimation des Paramètres de la Régression Logistique

- Dans le cadre d'un problème de classification binaire avec deux classes C_1 et C_2, on modélise la probabilité conditionnelle d'appartenance à la classe C1 en utilisant la fonction sigmoïde :

$$p(C_1|\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))}$$

- où f(x) est un modèle linéaire donné par :

$$f(\mathbf{x}) = w_0 + \sum_{d=1}^{D} w_d x_d = w_0 + \mathbf{w}^T \mathbf{x}$$

La probabilité d'appartenance à la classe C2 est alors :

$$p(C_2|\mathbf{x}) = 1 - p(C_1|\mathbf{x}) = \frac{\exp(-f(\mathbf{x}))}{1 + \exp(-f(\mathbf{x}))}$$

Estimation des Paramètres de la Régression Logistique

- Contrairement à la régression linéaire, la régression logistique utilise la fonction sigmoïde, qui rend l'estimation directe de **w** compliquée. Il est impossible d'obtenir une solution analytique (formule explicite) car l'équation obtenue lors de la dérivation ne peut pas être résolue directement. Comment calculer $\tilde{\mathbf{w}} = (w_0, w_1, \ldots, w_D)$?

- **Solution :**
  On utilise une approche basée sur le **maximum de vraisemblance** pour estimer **w**.

$$y = \begin{cases} 1, & \text{si } x \in C_1 \\ 0, & \text{si } x \in C_2 \end{cases}$$

Estimation des Paramètres de la Régression Logistique

- L'objectif est de maximiser la vraisemblance, c'est-à-dire le produit des probabilités des observations

$$L(\mathbf{w}) = \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)})$$

où :

$$p(y^{(i)}|\mathbf{x}^{(i)}) = p(C_1|\mathbf{x}^{(i)})^{y^{(i)}} \cdot p(C_2|\mathbf{x}^{(i)})^{(1-y^{(i)})}$$

En remplaçant par les expressions des probabilités :

$$p(y^{(i)}|\mathbf{x}^{(i)}) = \left(\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right)^{y^{(i)}} \cdot \left(\frac{\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right)^{(1-y^{(i)})}$$

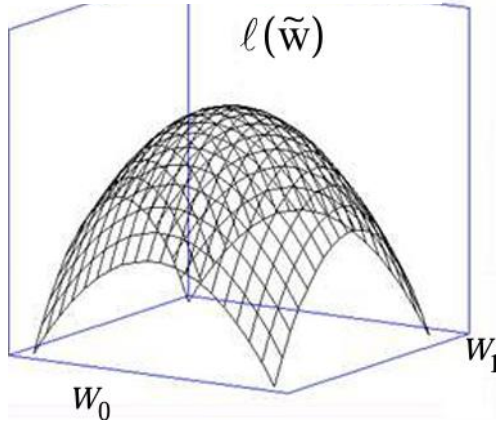
Estimation des Paramètres de la Régression Logistique

Plutôt que de maximiser $L(\mathbf{w})$, on maximise sa version logarithmique, qui est plus simple à manipuler :

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} \ln p(y^{(i)} = 1|\mathbf{x}^{(i)}) + (1 - y^{(i)}) \ln p(y^{(i)} = 0|\mathbf{x}^{(i)}) \right]$$

En substituant les expressions des probabilités :

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)}) - \ln(1 + \exp(\mathbf{w}^T \mathbf{x}^{(i)})) \right]$$



Estimation des Paramètres de la Régression Logistique

Descente de gradient :

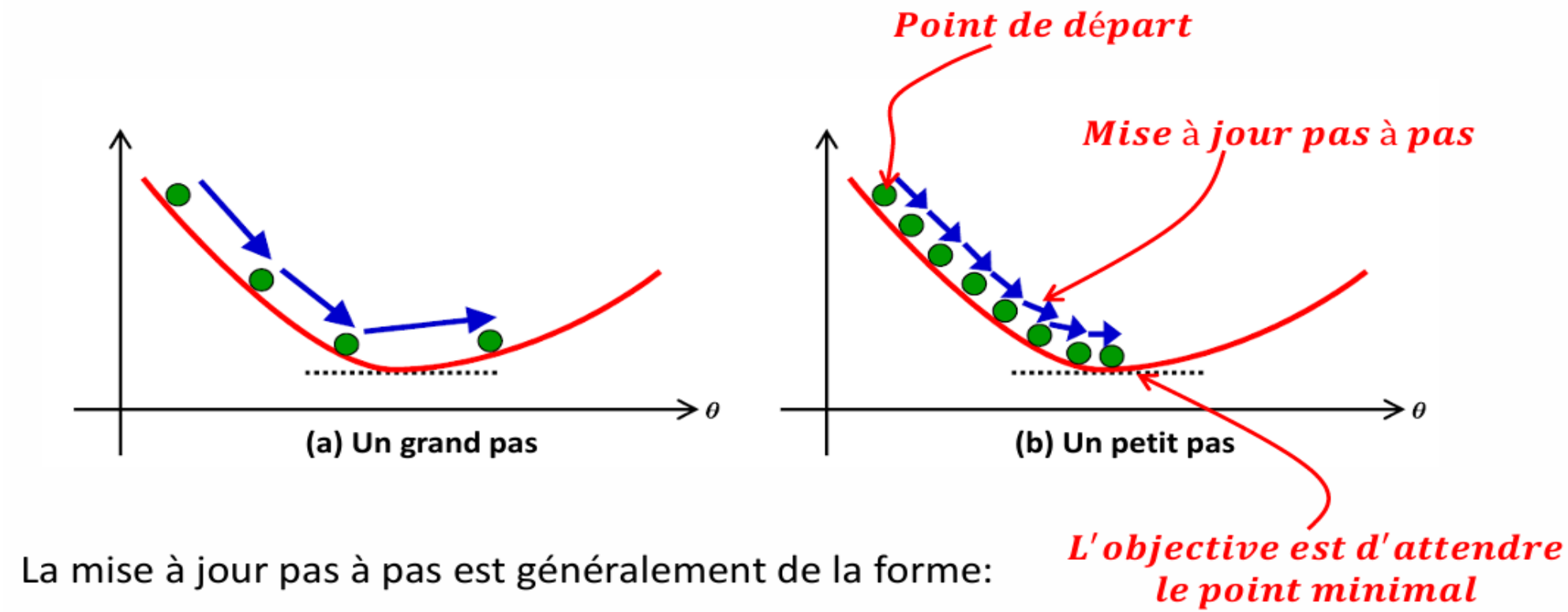
- On met à jour $\mathbf{w}$ selon la règle :

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \nabla \ell(\mathbf{w})$$

- Où η est un taux d'apprentissage et $\nabla \ell(\mathbf{w})$ est le gradient :

$$\nabla \ell(\mathbf{w}) = \sum_{i=1}^{N} \left(y^{(i)} - p(y^{(i)} = 1 | \mathbf{x}^{(i)}) \right) \mathbf{x}^{(i)}$$

Estimation des Paramètres de la Régression Logistique



La mise à jour pas à pas est généralement de la forme:

$$\boldsymbol{w}(t+1) = \boldsymbol{w}\,(t) + \alpha\nabla(\boldsymbol{w})$$

Un **pas trop grand** risque de dépasser la solution optimale, tandis qu'un **pas trop petite** entraine une convergence très lente.

Estimation des Paramètres de la Régression Logistique

1. **Initialisation aléatoire de *w***

2. **Calcul de la fonction sigmoid**

3. **Mise à jour de la dérivée**

4. Estimation des paramètres **w**

α : coefficient d'apprentissage

For $d = 0..D$
$w_d \leftarrow rand(-0.01, 0.01)$
Repeat
  For $d = 0..D$
    $\Delta w_d \leftarrow 0$
  For $i = 1..N$
    $\mathbf{x} \leftarrow 0$
    For $d = 0..D$
      $s \leftarrow s + w_d x_d^{(i)}$
    $r \leftarrow sigmoid(s)$
    For $d = 0..D$
      $\Delta w_d \leftarrow \Delta w_d + x_d^{(i)}(y^{(i)} - r)$
  For $d = 0..D$
    $w_d \leftarrow w_d + \alpha \Delta w_d$
Until convergence

Estimation des Paramètres de la Régression Logistique

- Les paramètres w peuvent être estimés par la méthode de la descente du gradient à l'aide de la mise à jour suivante :

$$\tilde{\mathbf{w}}(t+1) = \tilde{\mathbf{w}}(t) + \alpha \nabla \ell(\tilde{\mathbf{w}})$$

- où α est un coefficient d'apprentissage.
- Inconvénients :
  – **Choix de la valeur initiale de** w: La performance de l'algorithme peut être influencée par cette initialisation.
  – **Sur-apprentissage** : Risque d'obtenir des valeurs de w de grande amplitude, entraînant **une frontière de décision trop rigide**

Régularisation des paramètres de la Régression Logistique (RL)

- Pour éviter le **sur-apprentissage** en contrôlant la complexité du modèle. Dans le cas de la régression logistique, cela consiste à ajouter un terme de pénalisation aux paramètres du modèle pour empêcher les coefficients w de devenir trop grands.

- La fonction de log-vraisemblance standard pour la régression logistique est donnée par :

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} \left(w_0 + \mathbf{w}^T x^{(i)} \right) - \ln \left(1 + \exp(w_0 + \mathbf{w}^T x^{(i)}) \right) \right]$$

- Pour régulariser cette fonction, on lui ajoute un terme de pénalisation

$$\ell_{\text{reg}}(\mathbf{w}) = \sum_{i=1}^{N} \left[y^{(i)} (w_0 + \mathbf{w}^T x^{(i)}) - \ln(1 + \exp(w_0 + \mathbf{w}^T x^{(i)})) \right] - \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$$

Étude d’une application: analyse de spams

- **Qu’est-ce qu’un email spam?**
- Un email spam est un email non sollicité et non pertinent, envoyé en grands lots vers des boites emails d’utilisateurs.
- Le but du spammer sont divers: les publicités pour les sites produits/Web, messages en chaines, assurer un gain rapide d'argent, l’usurpation d’identités, etc.

Étude d'une application: analyse de spams

Filtres anti-spams

- Les filtres de spams basés sur l'analyse de texte vérifient l'existence/absence de certains mots ou de symboles.
- Dans un email, la présence de mots, tels que: héritage, loterie, dollars, etc., et de symboles tels que: '$', '¥', '€', '!', etc., augmentent la probabilité d'un spam.
- Ces probabilités sont estimées à partir d'un ensemble d'apprentissage D contenant des emails étiquetés.
- Les filtres peuvent faire des erreurs. Idéalement, les filtres doivent s'adapter et s'améliorer avec le temps.

Étude d'une application: analyse de spams

- Exemple avec Matlab
- L'ensemble d'entrainement est créé par Mark Hopkins et al. de Hewlett-Packard Labs. https://archive.ics.uci.edu/ml/datasets/Spambase
- L'ensemble contient 4601 emails. Chaque email possède 57 valeurs attributs reflétant les propriétés de l'email. Parmi cet attributs, on a:
  - 48 sont des fréquences de certains mots.
  - 06 sont des fréquences de certains caractères.
  - 03 comptent la longueur de chaines non interrompues.

Étude d'une application: analyse de spams

Exercice:

- Étudier la régression logistique
- Utiliser le classificateur de Bayes et la régression logistique pour classer les données de spams.
- Utiliser une validation croisée pour calculer l'erreur de classification (ex. moyenne de 10 validation en retenant à chaque fois 10% de données pour la validation)

FIN

De chapitre05