La régression logistique (RL)

Du classificateur de Bayes à la RL

- Supposons une classification binaire avec $\mathscr{C} = \{C_1, C_2\}$.
- Une donnée avec D attributs $\mathbf{x} = (\mathrm{x}_1, \ldots, \mathrm{x}_D) \in \mathbb{R}^D$, on aura:

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

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

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

Du classificateur de Bayes à la RL

- On suppose dans chaque classe C_k:

  ☞l'HBN pour la classification.

  ☞Chaque attribut suit une distribution Gaussienne.

  ☞Chaque attribut x_d a la même variance σ_d dans les 2 classes.

- On a alors:

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

Du classificateur de Bayes à la RL

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

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

- En posant:

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

- On obtient:

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

La régression logistique (*)

- Plus généralement, on suppose **des classes non Gaussiennes**, et on aura des paramètres généraux w_0 et $\mathbf{w}$, de sorte que:

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

- Et puisque : $p(C_1|\mathbf{x} = x) + p(C_2|\mathbf{x} = x) = 1$, on aura:

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

(*) La régression logistique est une appellation erronée

La régression logistique

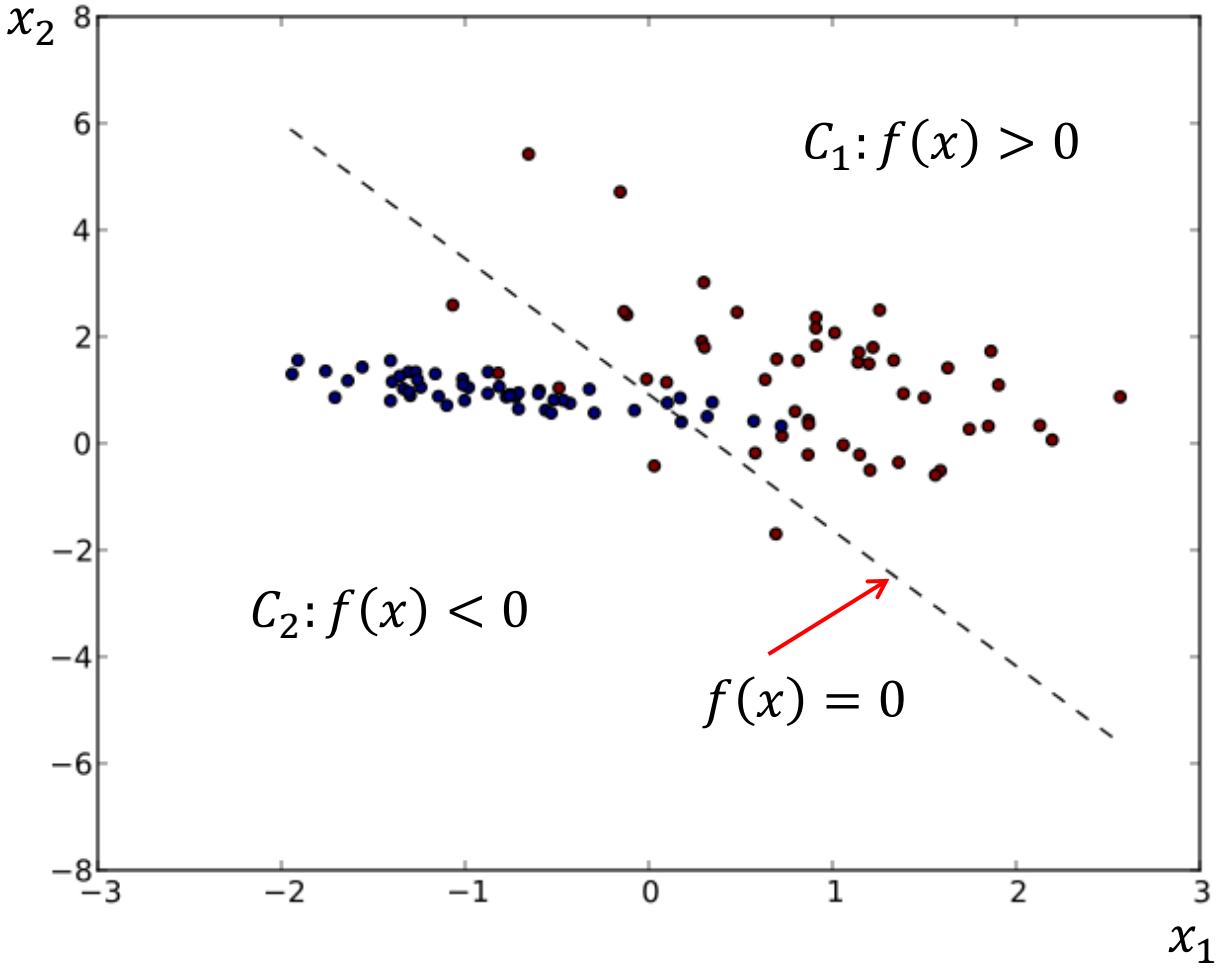
- Notons dans ce cas que si $f(x) = 0$, on a:

$$p(C_1|\mathbf{x} = x) = p(C_2|\mathbf{x} = x) = 0.5$$

- De plus, si $f(x) \neq 0$, on peut mesurer la quantité:

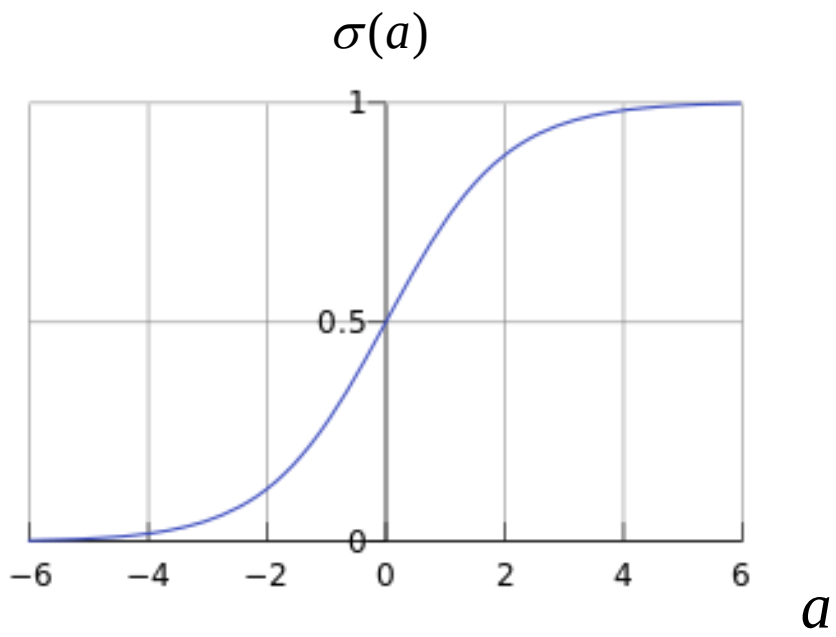
$$\ln\left(\frac{p(C_1|\mathbf{x} = x)}{p(C_2|\mathbf{x} = x)}\right) = f(x) = w_0 + \sum_{d=1}^{D}(w_d x_d) \quad \begin{cases} > 0? \\ \\ < 0? \end{cases}$$

La régression logistique



La fonction sigmoïde

- La fonction sigmoïde est définie par: $$\sigma(a) = \frac{1}{1+\exp(-a)}$$



- Quand a varie de $-\infty$ à $+\infty$, $\sigma(a)$ variera de 0 à 1.

Estimation des paramètres de la RL

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

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

- Comment calculer $\tilde{\mathbf{w}} = (w_0, w_1, \ldots, w_D)$?

☞ **Forme exacte impossible** à cause de la **fonction sigmoïde**.

☞ Utiliser **le maximum de vraisemblance**.

Estimation des paramètres de la RL

- On peut choisir le codage suivant pour la variable cible:

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

- Le **maximum de vraisemblance** vise à maximiser le produit **des probabilités des classes**, comme suit:

$$\tilde{\mathbf{w}}^* = \text{argmax}_{\tilde{\mathbf{w}}}(L(\tilde{\mathbf{w}}))$$

$$L(\tilde{\mathbf{w}}) = p\left(x^{(1)}, y^{(1)}\right) p(x^{(2)}, y^{(2)}) \cdots p(x^{(N)}, y^{(N)})$$

Estimation des paramètres de la RL

- **Le maximum de vraisemblance conditionnel** vise à maximiser le produit **des probabilités conditionnelles** suivant:

$$L(\tilde{\mathbf{w}}) = \prod_{i=1}^{N} \left[\left(p(y^{(i)} = 1 | x^{(i)}) \right)^{y^{(i)}} \left(p(y^{(i)} = 0 | x^{(i)}) \right)^{(1-y^{(i)})} \right]$$

- En posant **le logarithme au produit**, on obtient:

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

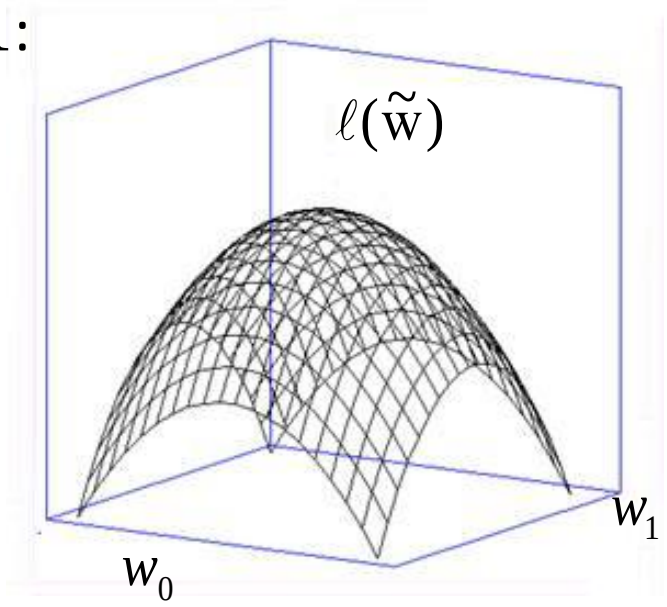
$$= \sum_{i=1}^{N} y^{(i)} \ln\left(\frac{p(y^{(i)} = 1 | x^{(i)})}{p(y^{(i)} = 0 | x^{(i)})} \right) + \ln\left(p(y^{(i)} = 0 | x^{(i)}) \right)$$

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

Estimation des paramètres de la RL

- La fonction $\ell(\tilde{\mathbf{w}})$ est **concave**. Elle peut être **maximisée**.

Exemple: $D = 1$:



- Comment trouver $\tilde{\mathbf{w}}$ qui maximise $\ell(\tilde{\mathbf{w}})$?

Estimation des paramètres de la RL

- Puisqu'on ne peut pas utiliser un forme exacte pour isoler la valeur de $\tilde{\mathbf{w}}$, on peut utiliser la **montée du gradient**.

- Le gradient de la fonction $\ell(\tilde{\mathbf{w}})$ est donnée par:

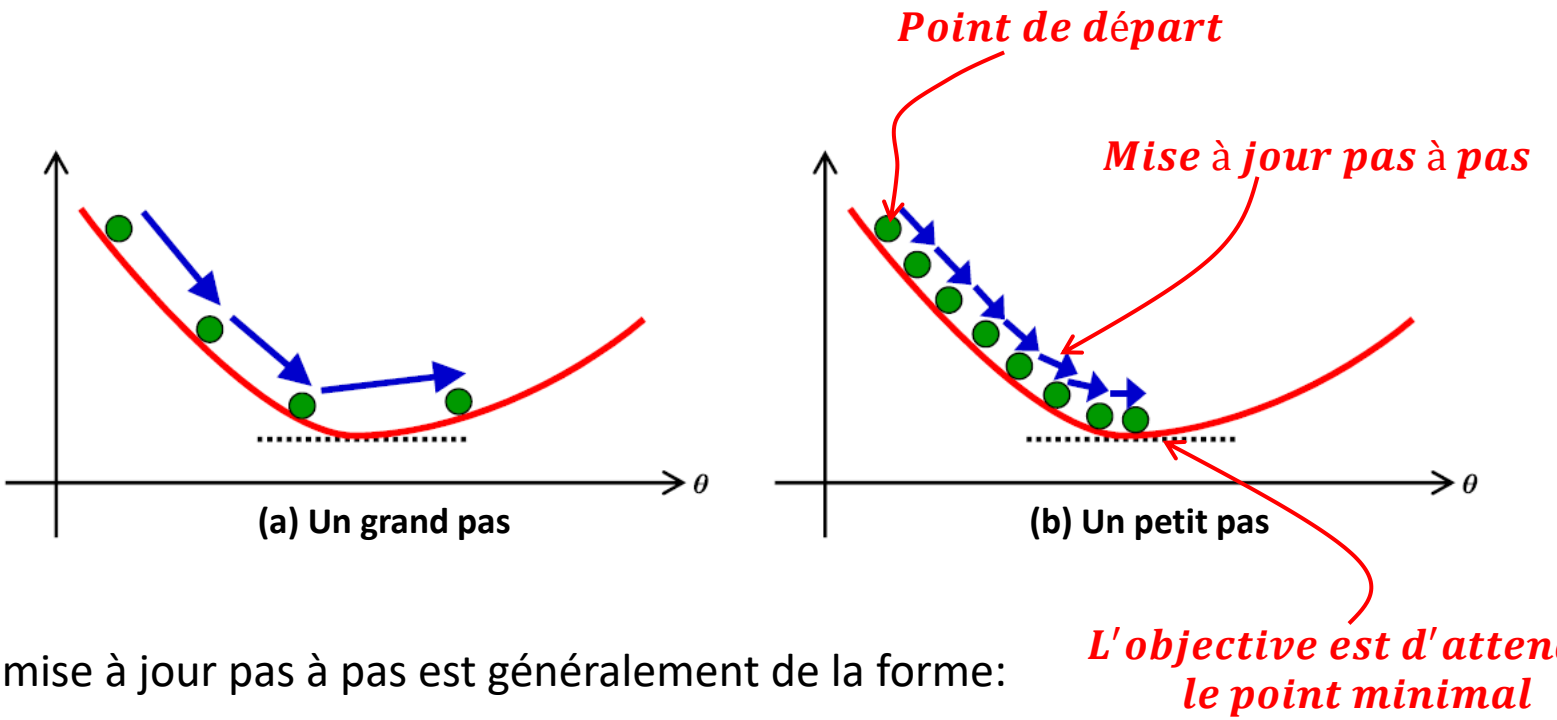
$$\nabla \ell(\tilde{\mathbf{w}}) = \left(\frac{\partial \ell(\tilde{\mathbf{w}})}{\partial w_0}, \frac{\partial \ell(\tilde{\mathbf{w}})}{\partial w_1}, \cdots, \frac{\partial \ell(\tilde{\mathbf{w}})}{\partial w_D} \right)^T$$

- On peut démontrer que:

$$\frac{\partial \ell(\tilde{\mathbf{w}})}{\partial w_d} = \sum_{i=1}^{N} {x_d}^{(i)} (y^{(i)} - p(y^{(i)} = 1 | x^{(i)}))$$

Estimation des paramètres de la RL

- Algorithme de descente du gradient



- 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 RL

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 RL

- On peut estimer les paramètres $\tilde{\mathbf{w}}$ par **la montée du gradient** via la mise à jour suivante:

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

α : est un **coefficient d'apprentissage.**

- **Inconvénients:**
  - **Choix de la valeur initiale** de $\tilde{\mathbf{w}}$?
  - **Sur-apprentissage:** Tendance à produire des $\tilde{\mathbf{w}}$ avec une grande amplitude (**frontière rigide**).

Régularisation des paramètres de la RL

- Utiliser **une régularisation** pour la fonction à maximiser:

$$\ell(\tilde{\mathbf{w}}) = \sum_{i=1}^{N} y^{(i)}\left(w_0 + \mathbf{w}^T x^{(i)}\right) - \ln\left(1 + \exp\left(w_0 + \mathbf{w}^T x^{(i)}\right)\right) - \frac{\lambda}{2}\tilde{\mathbf{w}}^T\tilde{\mathbf{w}}$$

- Les entrées du vecteurs $\nabla\ell(\tilde{\mathbf{w}})$ seront alors calculés par:

$$\frac{\partial\ell(\tilde{\mathbf{w}})}{\partial w_d} = \sum_{i=1}^{N} {x_d}^{(i)}\left(y^{(i)} - p\left(y^{(i)} = 1\middle|x^{(i)}\right)\right) - \lambda\tilde{\mathbf{w}}$$

Classificateur de Bayes (CB) versus la RL

- **Classificateur de Bayes (CB):**

  ☞ Suppose une forme des fonctions $p(\mathbf{x}|y)$ et $p(\mathbf{x})$.

  ☞ Estime les paramètres de $p(\mathbf{x}|y)$ et $p(\mathbf{x})$à partir de$\mathcal{D}$.

  ☞ Utilise la **règle de Bayes** pour calculer $p(y|\mathbf{x})$.

- **La régression logistique (RL):**

  ☞ Suppose une forme de la fonction $p(y|\mathbf{x})$.

  ☞ Estime les paramètres de $p(y|\mathbf{x})$ à partir des $\mathcal{D}$.

Le CB versus la RL

- On suppose le cas simple de la classification binaire $y \in \{1,0\}$ et $\mathbf{x} \in \mathbb{R}^D$ (c.-à-d., $\mathrm{x}_d \rightsquigarrow \mathcal{N}(\mu_d, \sigma_d)$)

  ☞ Le nombre de paramètres du CB est: $4D + 1$.

  ☞ Le nombre de paramètres de la RL est: $D + 1$.

- Quand **l'HBN est satisfaite**, le CB et la RL ont en générale la même performance.

- Quand **l'HBN n'est pas satisfaite**, le CB est biaisé et obtient moins de performance que la RL.

Le CB versus la RL

- Le CB est **biaisé** car il est basé sur l'HBN.

- On peut démontrer que si $\mathcal{D}$ contient N données:

  ☞ La complexité algorithmique du CB est: $O(\log(N))$.

  ☞ La complexité algorithmique de la RL est: $O(N)$.

- En d'autres mots: le CB converge plus rapidement que la RL, pour sa **solution** (**restreinte**)

Étude d'une application: analyse de spams

"Wow! I've got one from someone I know!"

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.

TrustedBank™

Dear valued customer of TrustedBank,

We have recieved notice that you have recently attempted to withdraw the following amount from your checking account while in another country: $135.25.

If this information is not correct, someone unknown may have access to your account. As a safety measure, please visit our website via the link below to verify your personal information:

http://www.trustedbank.com/general/custverifyinfo.asp

Once you have done this, our fraud department will work to resolve this discrepency. We are happy you have chosen us to do business with.

Thank you,
TrustedBank

From: LuxR-Clones [mailto: [redacted]
Sent: Wednesday, December 14, 2011 5:39 AM
To: [redacted]
Subject: The best Christmas ever

mso-padding-alt:0in 5.4pt 0in 5.4pt; mso-para-margin:0in; mso-para-margin-bottom:.0001pt; mso-pagination:widow-orphan; font-size:10.0pt; font-family:"Times New Roman";}

Luxury Replicas

MERRY CHRISTMAS

Dec.19th is our Christmas Order Cut-Off Date
All orders will receive 15% to 20% discounts!

CAN YOU SEE THE DIFFERENCE?

Genuine Rolex
Seadweller edition

Our Replica Rolex
Seadweller edition

THE ONLY DIFFERENCE IS PRICE

MERRY CHRISTMAS!
Notice : Orders must be in by December 19th!

Christmas orders need to be placed prior to December 19th, this is our cut-off date.
There will be instant rebates of 15-20% on all multiple purchases.
All orders are shipped immediately to help ensure pre-Christmas arrival.

WE HAVE THE WORLD'S BEST WATCH => EQUIVALENTS!

WHAT MAKES THEM SO =>EXACT?

ALL WATCHES ARE MADE WITH:

Identical Materials as the originals
Identical Metals as the originals
Identical Labeling as the originals
Identical Functionality as the originals

These watches are so exact that your friends, loved-ones, co-workers or your local jeweler will never know the difference.

2012 WATCH-MODELS ARE AVAILABLE:
http://www.pointpurpose.com

From: Tiger Autumn via LinkedIn <member@linkedin.com>
Subject: Join my network on LinkedIn
Date: September 16, 2011 12:42:12 AM CDT
To: Lance Spitzner <lance@spitzner.net>
Reply-To: Tiger Autumn <tiger.autumn@chinamail.com>

LinkedIn

Tiger Autumn has indicated you are a Friend

I'd like to talk about cyber warfare with you.

I am from ShangHai,China.

- Tiger Autumn

Accept View invitation from Tiger Autumn

DID YOU KNOW **LinkedIn can help you find the right service providers using recommendations from your trusted network?**

Using LinkedIn Services, you can take the risky guesswork out of selecting service providers by reading the recommendations of credible, trustworthy members of your network.

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 $\mathcal{D}$ contenant des **emails étiquetés**.

- Les filtres peuvent faire des erreurs. Idéalement, les filtres doivent **s'adapter et s'améliorer** avec le temps.

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

- Pour importer les données sur Matlab:

  donnees_spam = load('spambase.txt');

- 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.

Exemple avec Matlab

Exercice:

- Étudier la régression logistique sur Matlab.
  https://www.mathworks.com/help/stats/glmfit.html

- 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)

Références

1. M. S. Allili. Techniques d'apprentissage automatique (Cours de 2e cycle). Université du Québec en Outaouais (UQO), Québec, Canada. Hivers 2015.
2. S. Rogers et M Girolami. A first Course in machine learning, CRC press, 2012.
3. C. Bishop. Pattern Recognition and Machine learning. Springer 2006.
4. R. Duda, P. Storck et D. Hart. Pattern Classification. Prentice Hall, 2002.