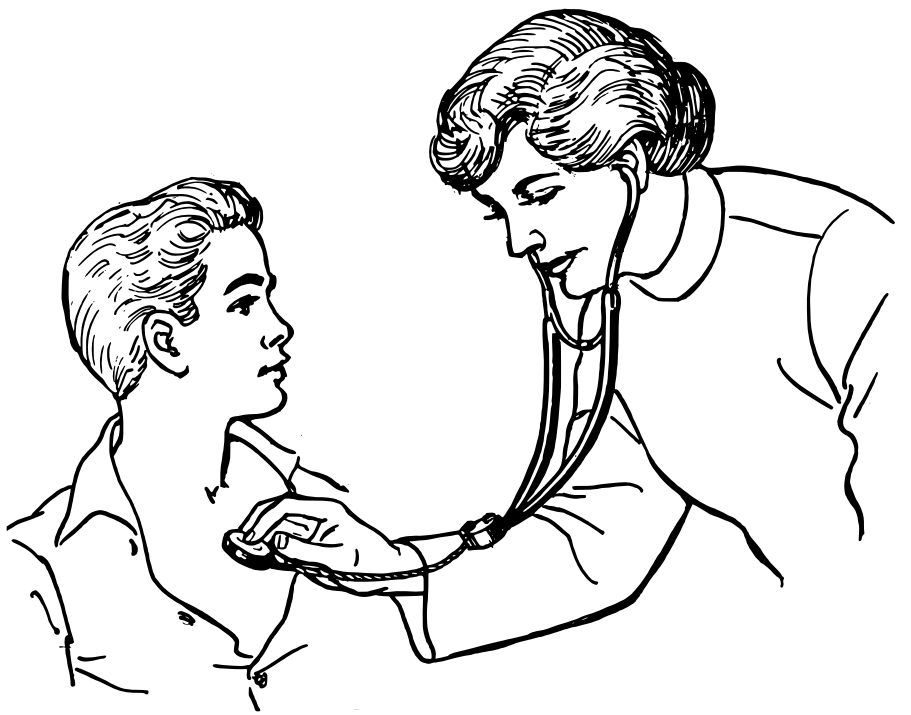
MS BGD
MDI 720 : Modèles linéaires généralisés (GLM)

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

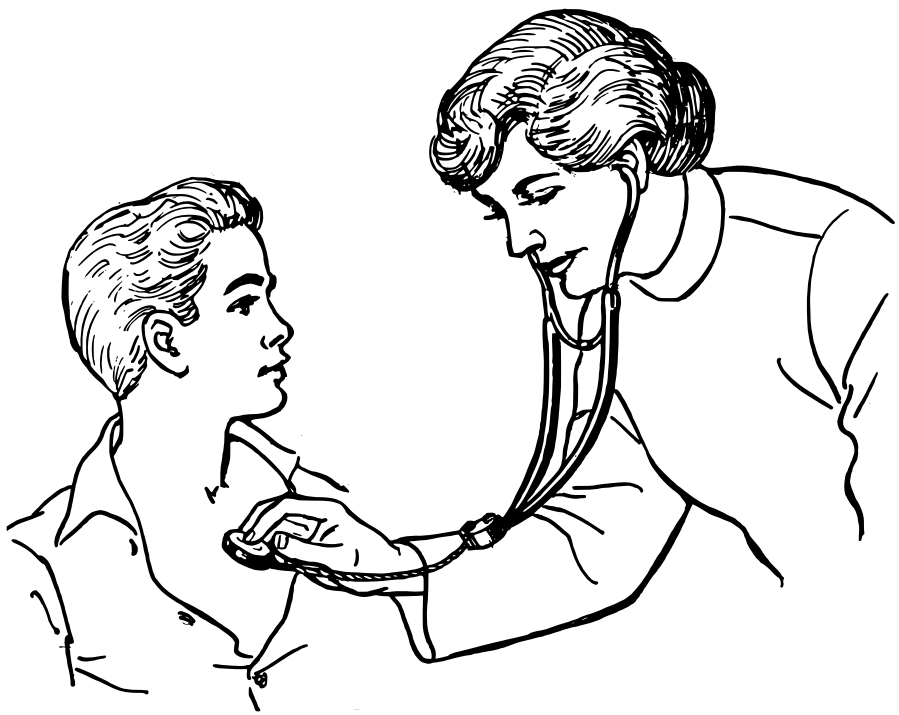
La classification : cadre binaire

Diagnostiquer des patients :



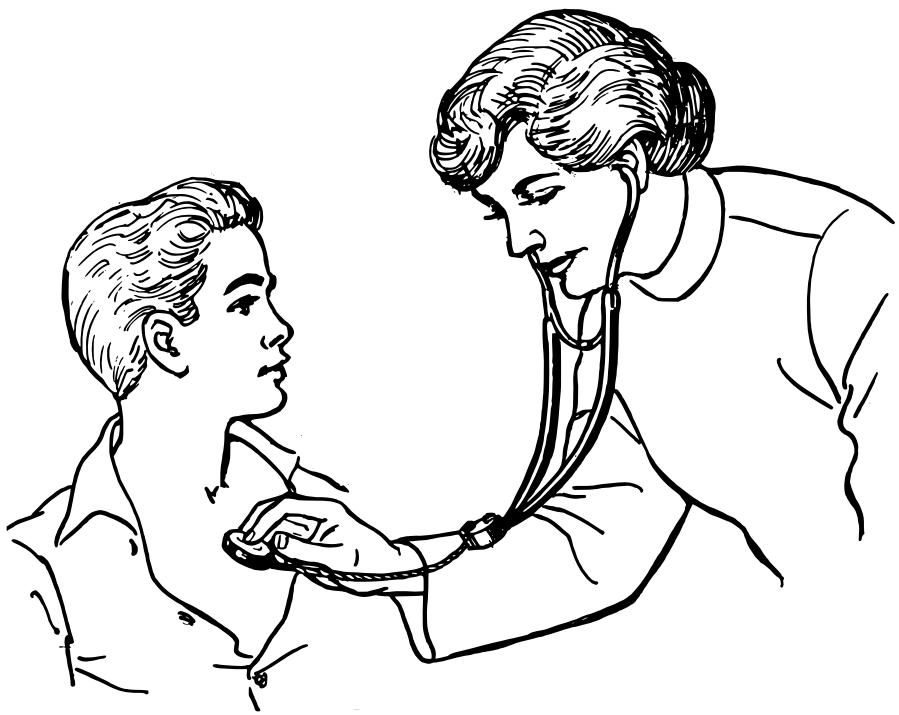
La classification : cadre binaire

Diagnostiquer des patients : malades



La classification : cadre binaire

Diagnostiquer des patients : ~~malades~~ sains



La classification : cadre binaire

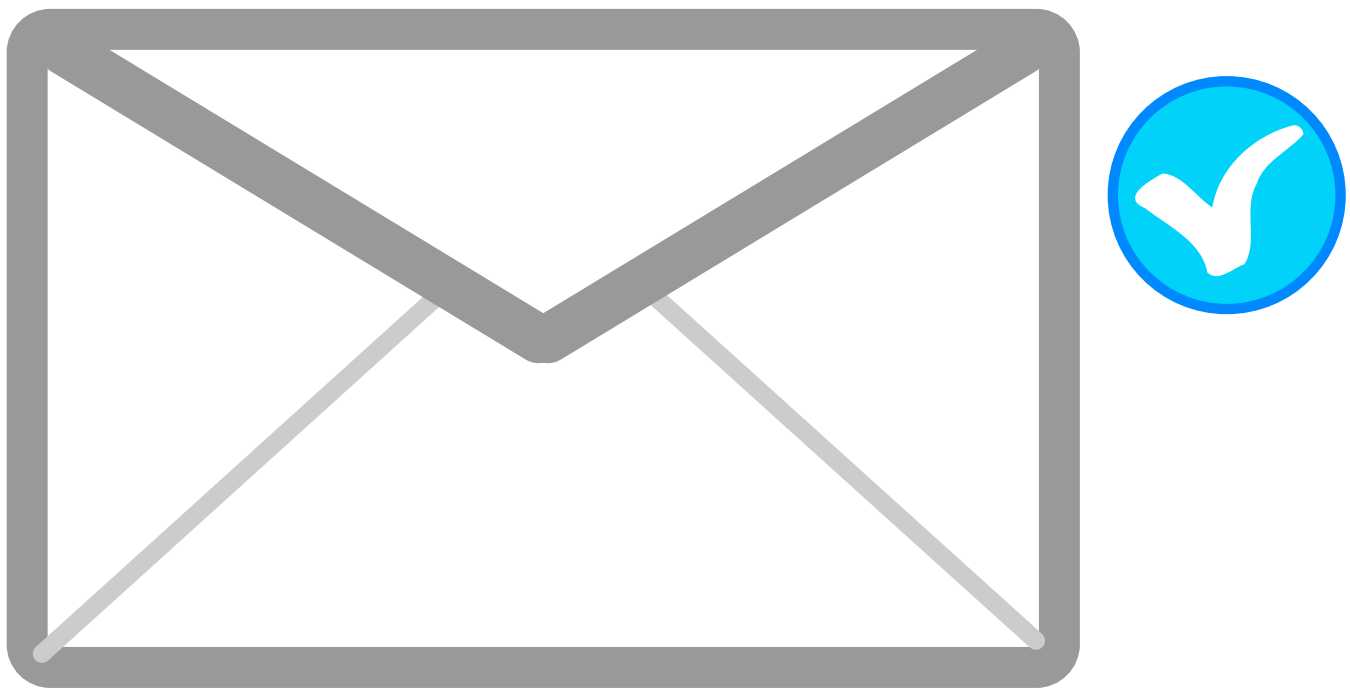
Classer des emails :

La classification : cadre binaire

Classer des emails : pourriels (spams)

La classification : cadre binaire

Classer des emails : ~~pourriels (spams)~~ normaux



La classification : cadre binaire

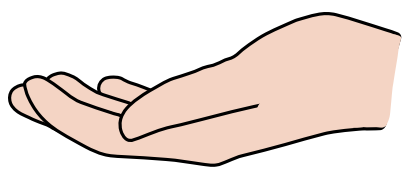
Classer des clients :

La classification : cadre binaire

Classer des clients : mauvais payeurs/fraudeurs

La classification : cadre binaire

Classer des clients : ~~mauvais payeurs/fraudeurs~~ bon payeurs

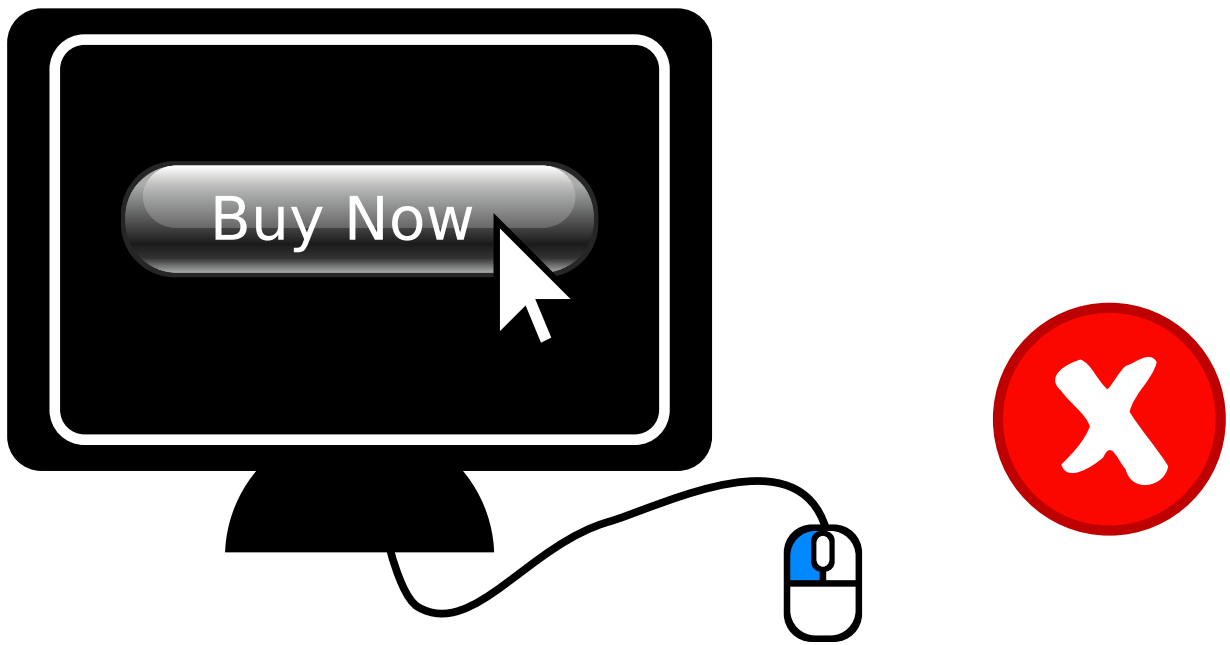


La classification : cadre binaire

Classer les surfeurs :

La classification : cadre binaire

Classer les surfeurs : futurs acheteurs

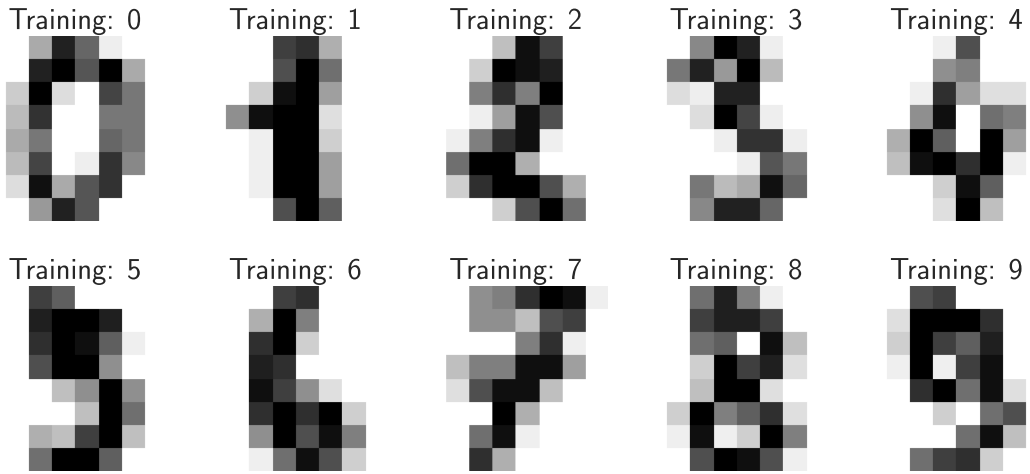


La classification : cadre binaire

Classer les surfeurs : ~~futurs acheteurs~~ ou pas...

La classification : cadre multi-classe

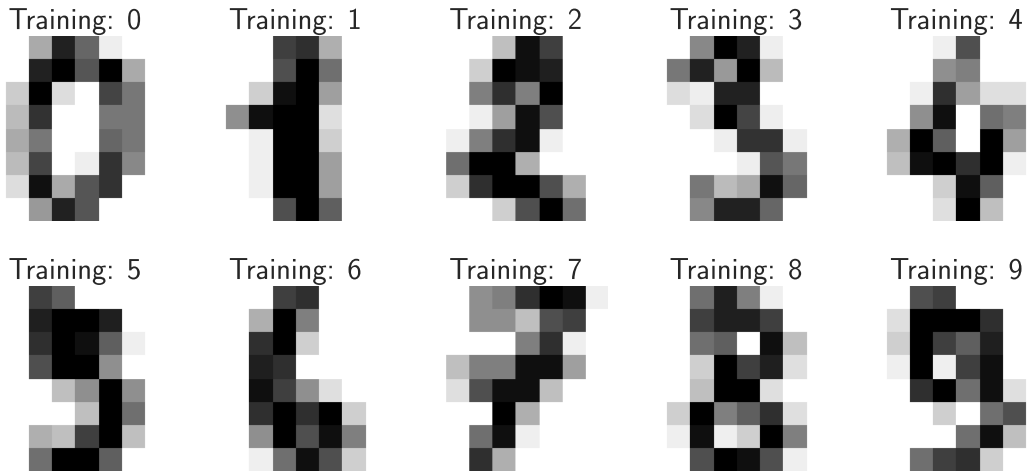
- Classer des chiffres numérisés
  (80's/90's : scans de codes postaux)



- Classer des objets dans des images
  (http://image-net.org/, 2010's)

La classification : cadre multi-classe

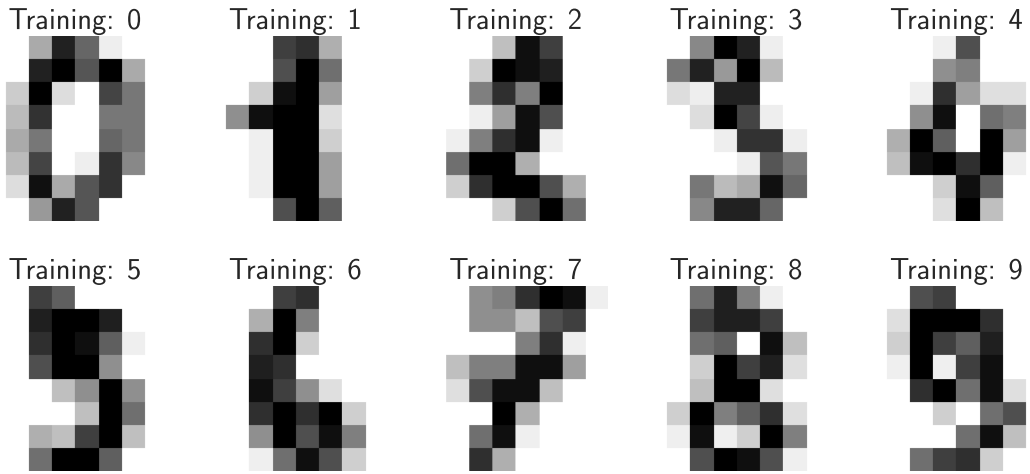
- Classer des chiffres numérisés
  (80's/90's : scans de codes postaux)



- Classer des objets dans des images
  (http://image-net.org/, 2010's)
- Classer des textes par thème (*e.g.*, 20newsgroups)

La classification : cadre multi-classe

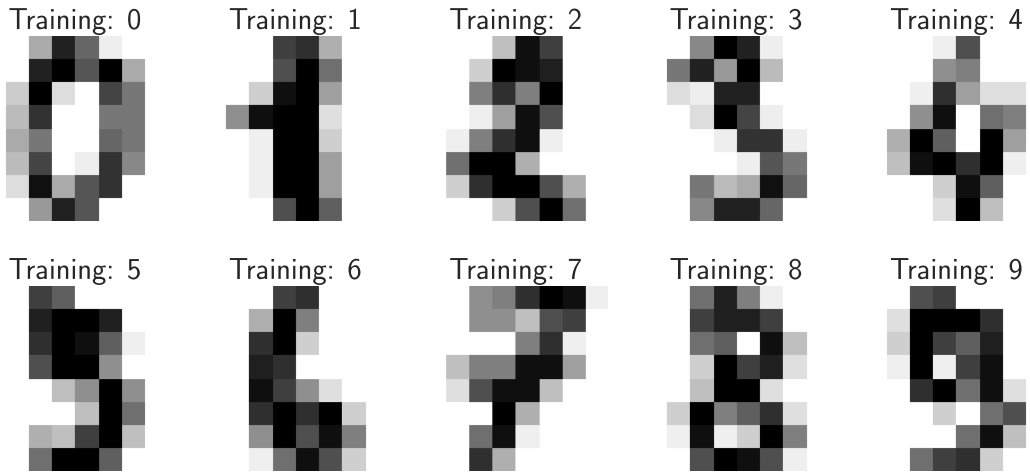
- Classer des chiffres numérisés
  (80's/90's : scans de codes postaux)



- Classer des objets dans des images
  (`http://image-net.org/`, 2010's)
- Classer des textes par thème (*e.g.*, 20newsgroups)
- Classer des espèces animales/végétales (*e.g.*, iris)
- ...

La classification : cadre multi-classe

- Classer des chiffres numérisés
  (80's/90's : scans de codes postaux)



- Classer des objets dans des images
  (`http://image-net.org/`, 2010's)
- Classer des textes par thème (*e.g.*, 20newsgroups)
- Classer des espèces animales/végétales (*e.g.*, iris)
- ...

Modèle de classification

K représente le nombre de classes ; on suppose que les classes sont indexées par l'ensemble $[\![0, K-1]\!] := \{0, \ldots, K-1\}$

Observations : $\mathbf{y} \in [\![0, K-1]\!]^n$

Variables explicatives : $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times p}$, n observations, p variables

Classifieur : c'est un estimateur $h_{\boldsymbol{\theta}} : \begin{cases} \mathbb{R}^p & \mapsto [\![0, K-1]\!] \\ \mathbf{x} & \to h_{\boldsymbol{\theta}}(\mathbf{x}) \end{cases}$

Objectif : minimiser l'erreur $\boxed{\mathbb{P}(y \neq h_{\boldsymbol{\theta}}(\mathbf{x})) = \mathbb{E}(\mathbb{1}_{y \neq h_{\boldsymbol{\theta}}(\mathbf{x})})}$,
équivalent à maximiser la précision (🇬🇧 : *accuracy*) : $\mathbb{P}(y = h_{\boldsymbol{\theta}}(\mathbf{x}))$

Sommaire

Prédiction linéaire et indicatrices

Tentative naïve : utiliser une prédiction linéaire pour faire de la classification

Résultat : simple, mais ne marche pas (on va le voir quand même pour s'en convaincre)

Traitement par régression linéaire brute

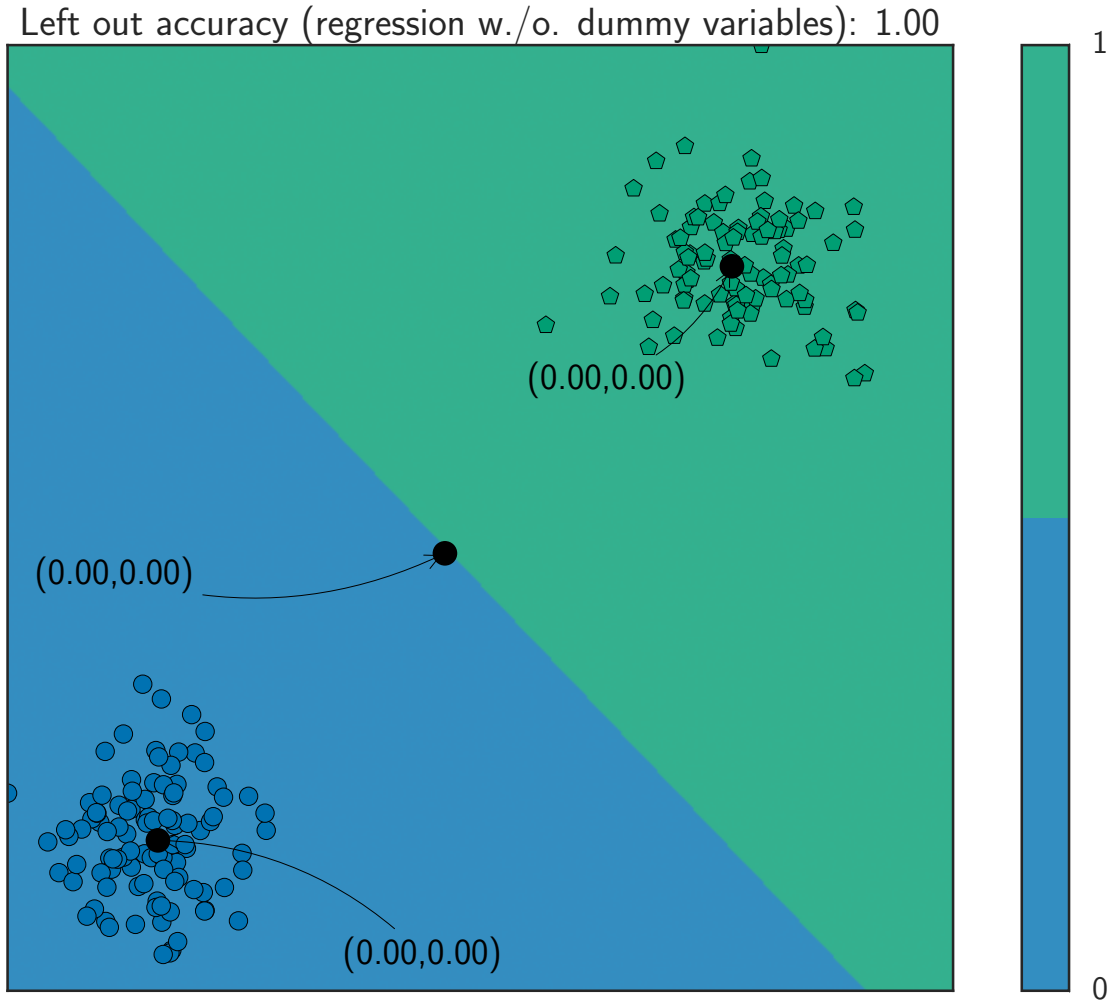
Idée naïve (I) : choisir comme classifieur $h_{\boldsymbol{\theta}}(\mathbf{x}) \approx \langle \mathbf{x}, \boldsymbol{\theta} \rangle$ (*i.e.*, faire le choix $\mathbf{y} \approx X\boldsymbol{\theta}$!), ou plus exactement, la classe la plus proche

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - X\boldsymbol{\theta}\|^2$$

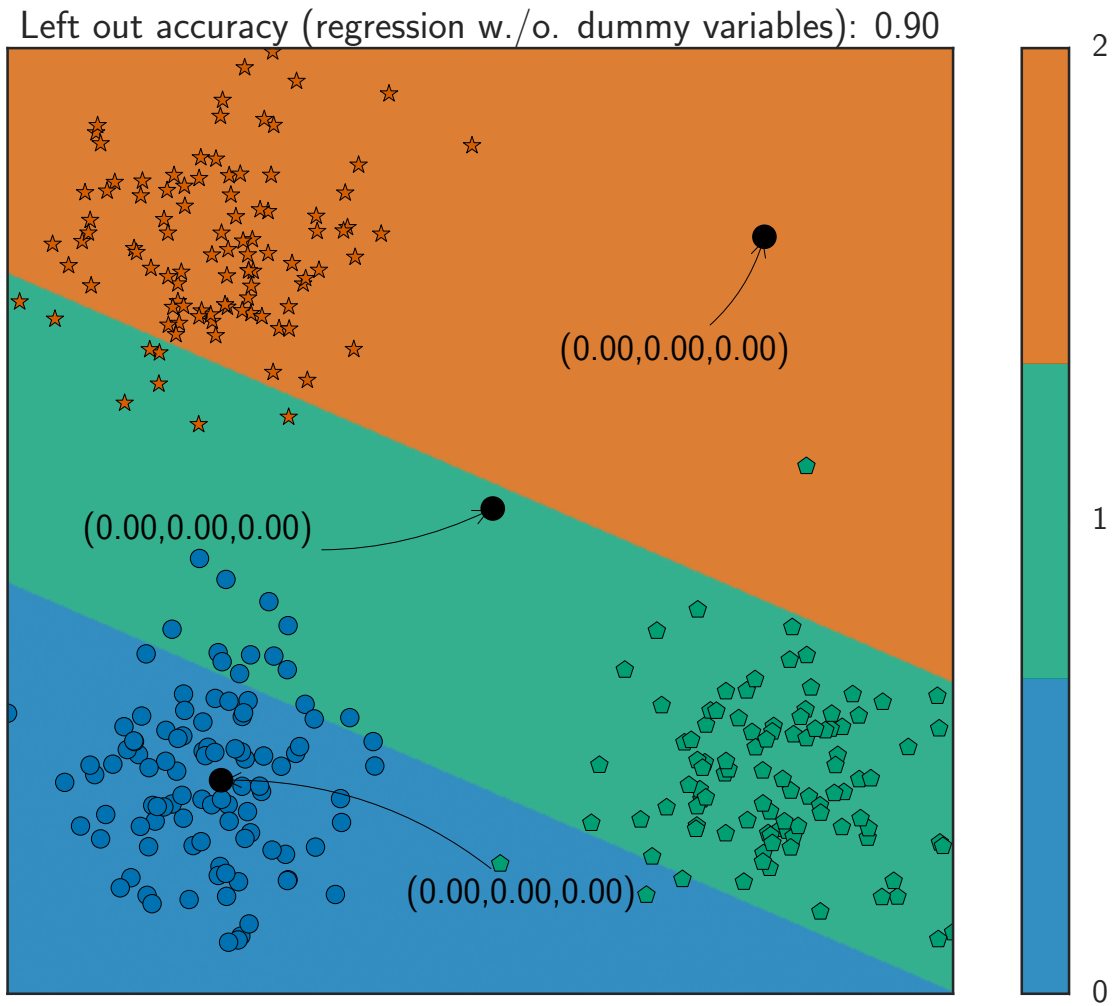
$$h_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = \underset{k \in [\![0, K-1]\!]}{\arg\min} \left| k - \langle \mathbf{x}, \hat{\boldsymbol{\theta}} \rangle \right|$$

Numériquement : on utilise un solveur de moindres carrés

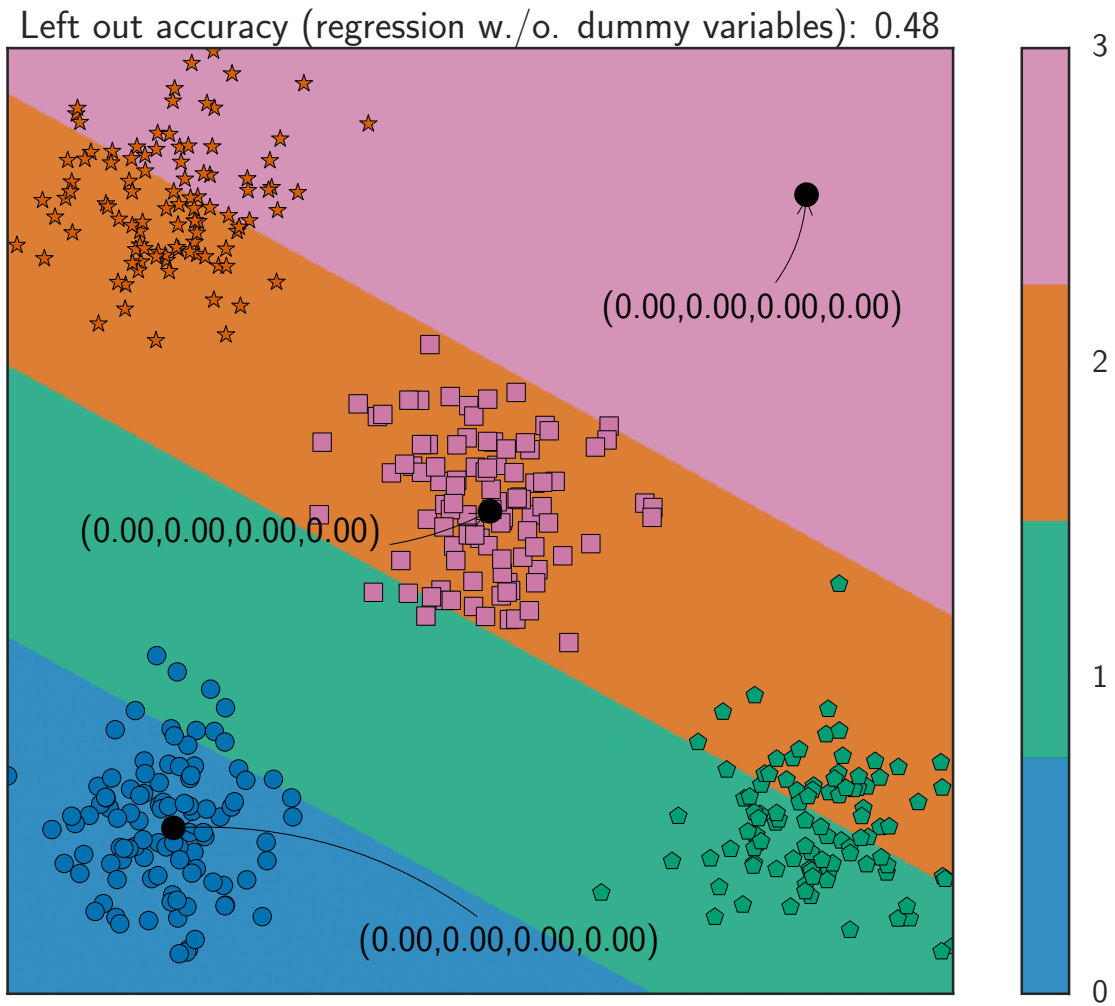
Un exemple avec deux classes



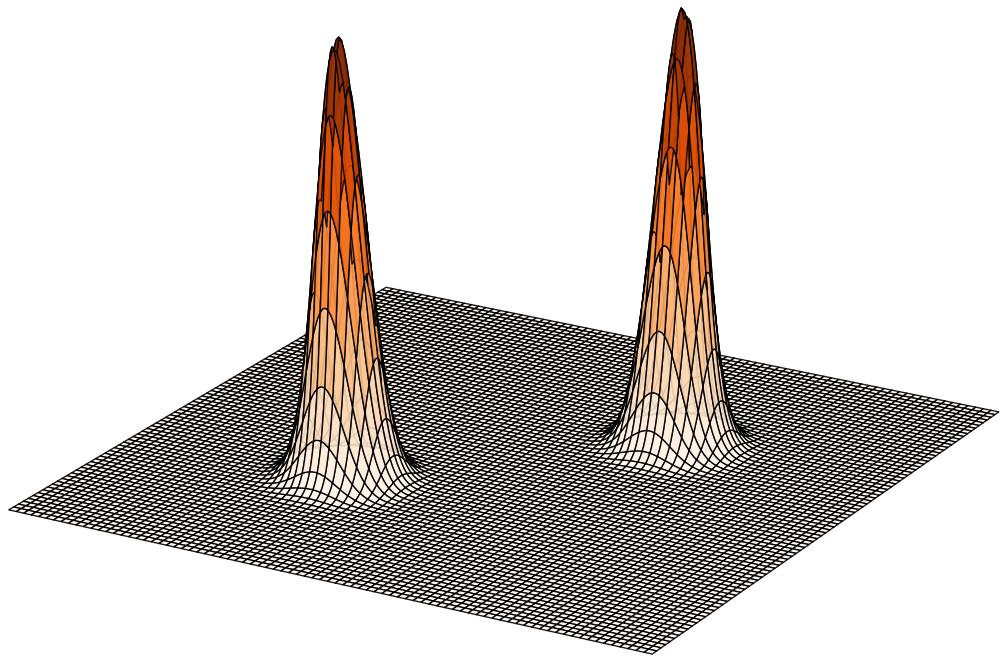
Un exemple avec trois classes



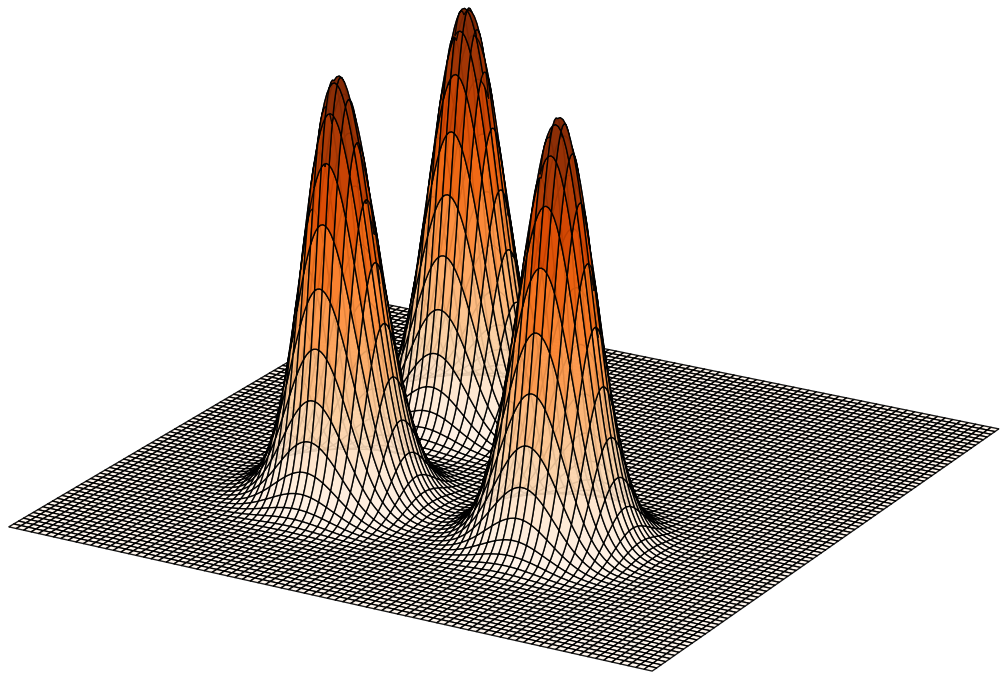
Un exemple avec quatre classes



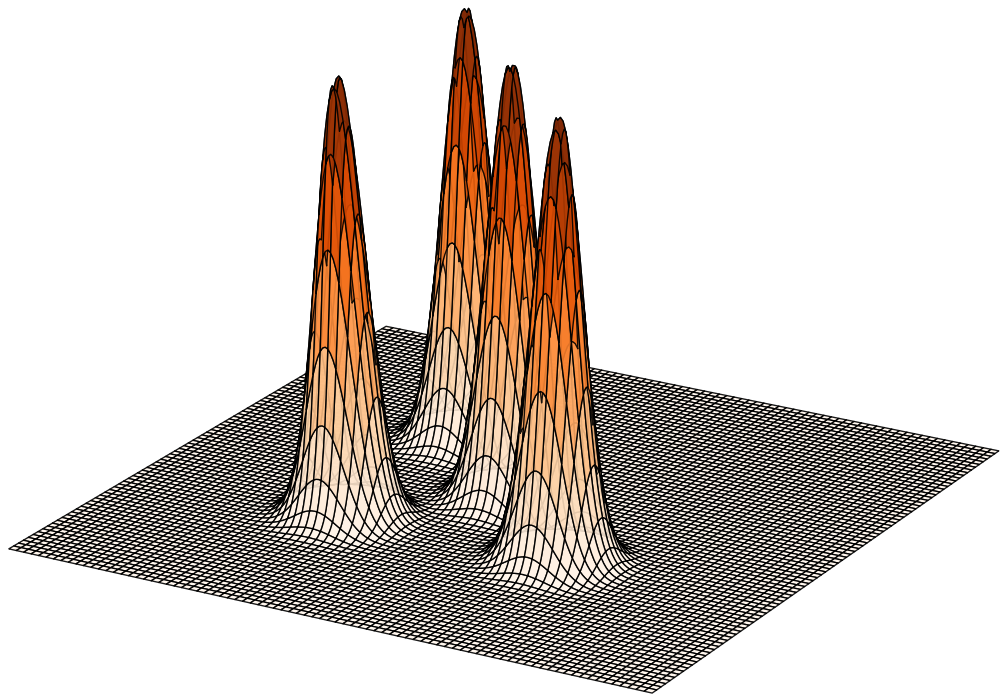
Distribution sous-jacente : deux classes



Distribution sous-jacente : trois classes



Distribution sous-jacente : quatre classes



Sommaire

Prédiction linéaire et indicatrices

Seconde idée naïve : utiliser une méthode de régression pour estimer $\mathbb{P}(y=0|\mathbf{x})$, $\mathbb{P}(y=1|\mathbf{x}), \dots, \mathbb{P}(y=K-1|\mathbf{x})$ et choisir la classe qui donne la plus grande probabilité.

Rem: $\mathbb{P}(y=k|\mathbf{x}) = \mathbb{E}(Z^{(k)}|\mathbf{x})$ et l'on définit

$$Z^{(k)} \in \mathbb{R}^n, \quad Z_i^{(k)} = \mathbb{1}_{y_i=k} = \begin{cases} 1, & \text{si } y_i = k, \\ 0, & \text{sinon.} \end{cases}$$

Régression multi-tâches :

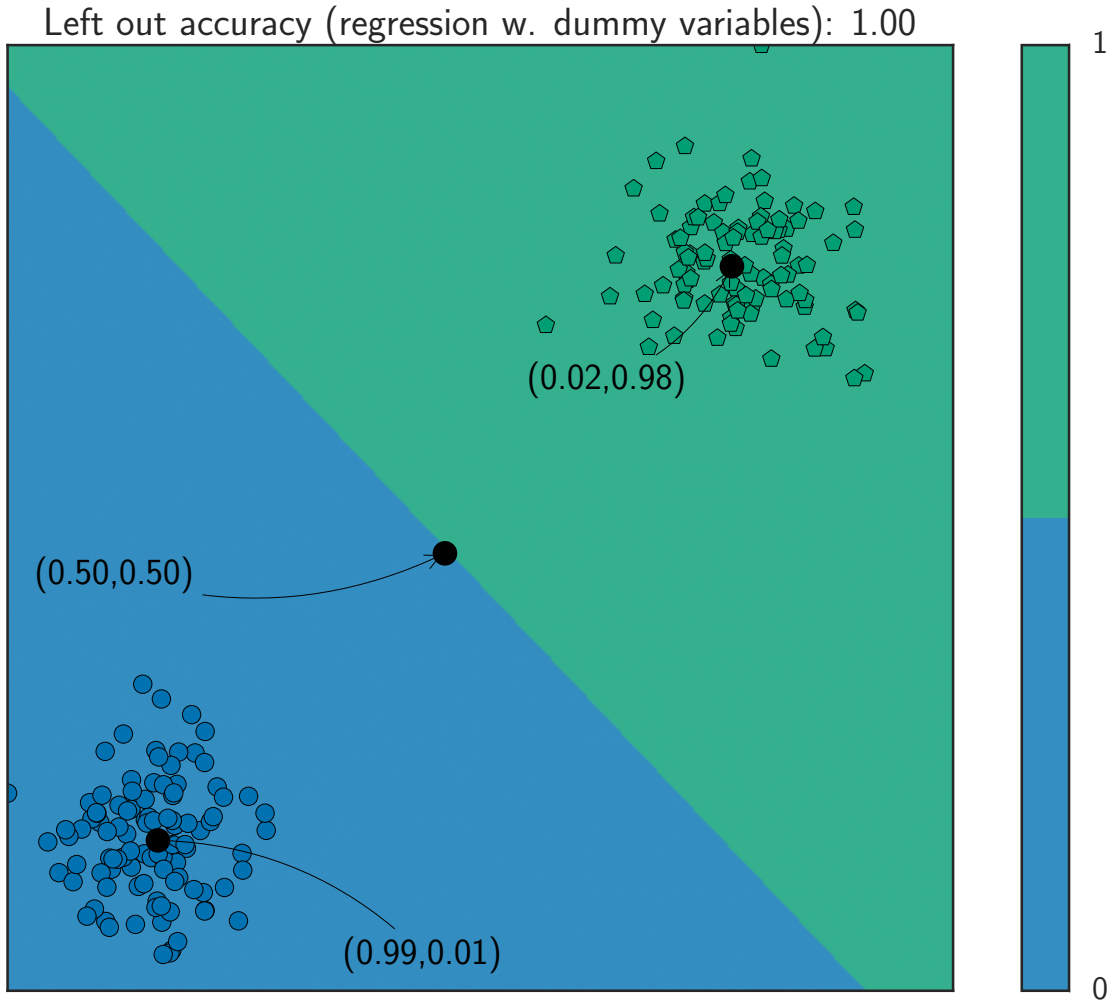
$$\boxed{[\hat{\boldsymbol{\theta}}^{(1)}, \dots, \hat{\boldsymbol{\theta}}^{(K)}] = \underset{\Theta \in \mathbb{R}^{p \times K}}{\arg\min} \|X\Theta - Z\|^2}$$

Le classifieur (choisit la probabilité estimée la plus grande) :

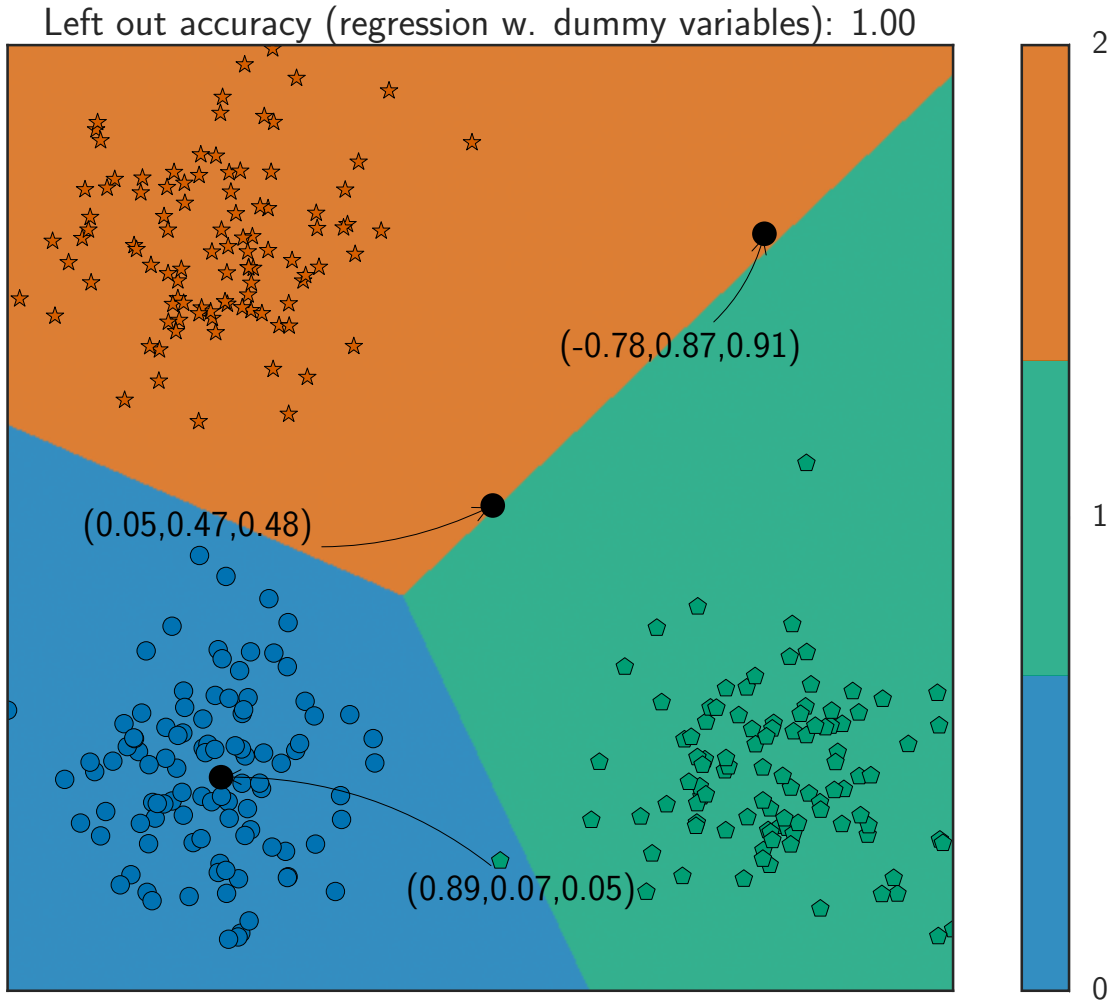
$$h_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = \underset{k \in [\![0, K-1]\!]}{\arg\max} \langle \mathbf{x}, \hat{\boldsymbol{\theta}}^{(k)} \rangle$$

Rem: on peut utiliser un solveur de moindres carrés multi-tâches

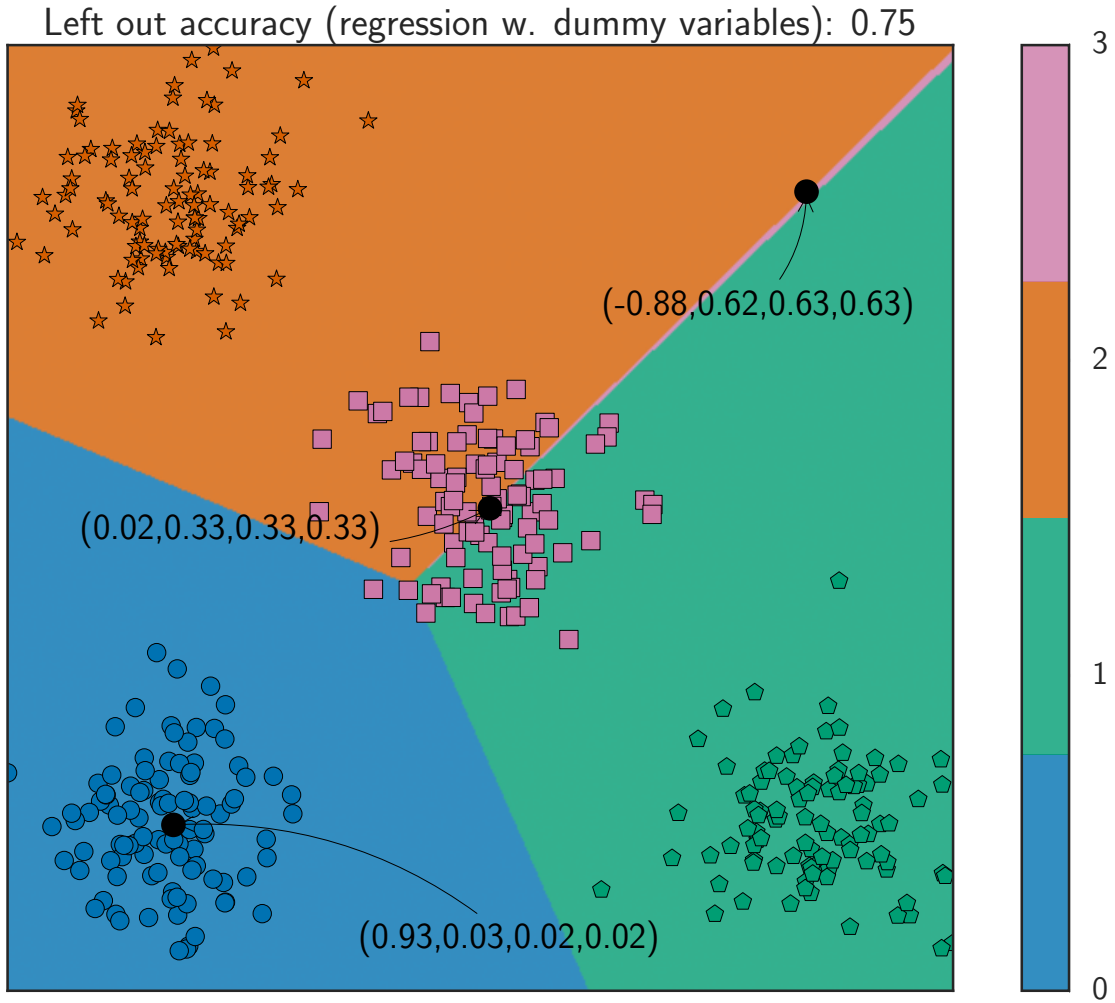
Un exemple avec deux classes



Un exemple avec trois classes



Un exemple avec quatre classes



Avantages / Inconvénients : prédiction linéaire et indicatrice

Avantages

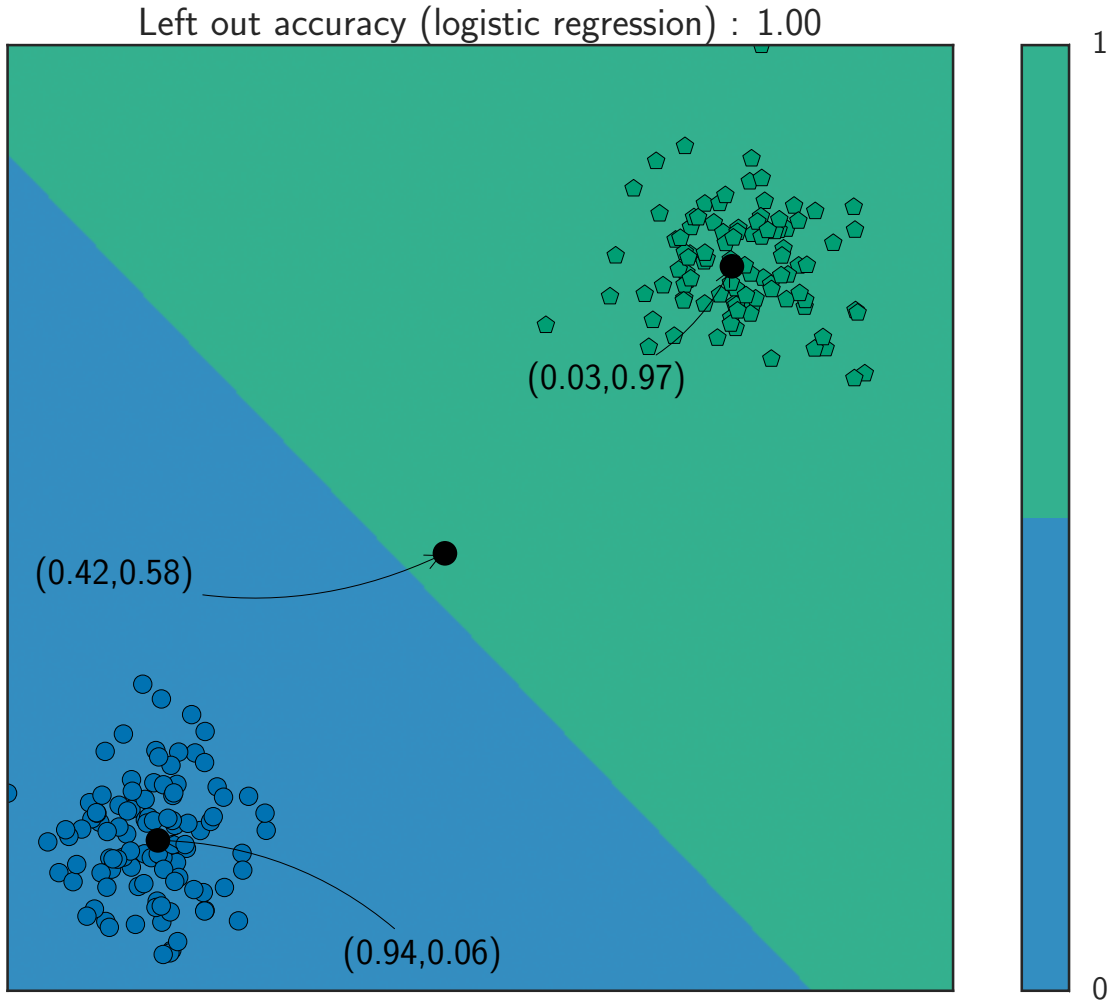
- Simplicité : peu d'hypothèse sur le modèle
- Implémentable : facile avec un solveur de moindres carrés

Inconvénients

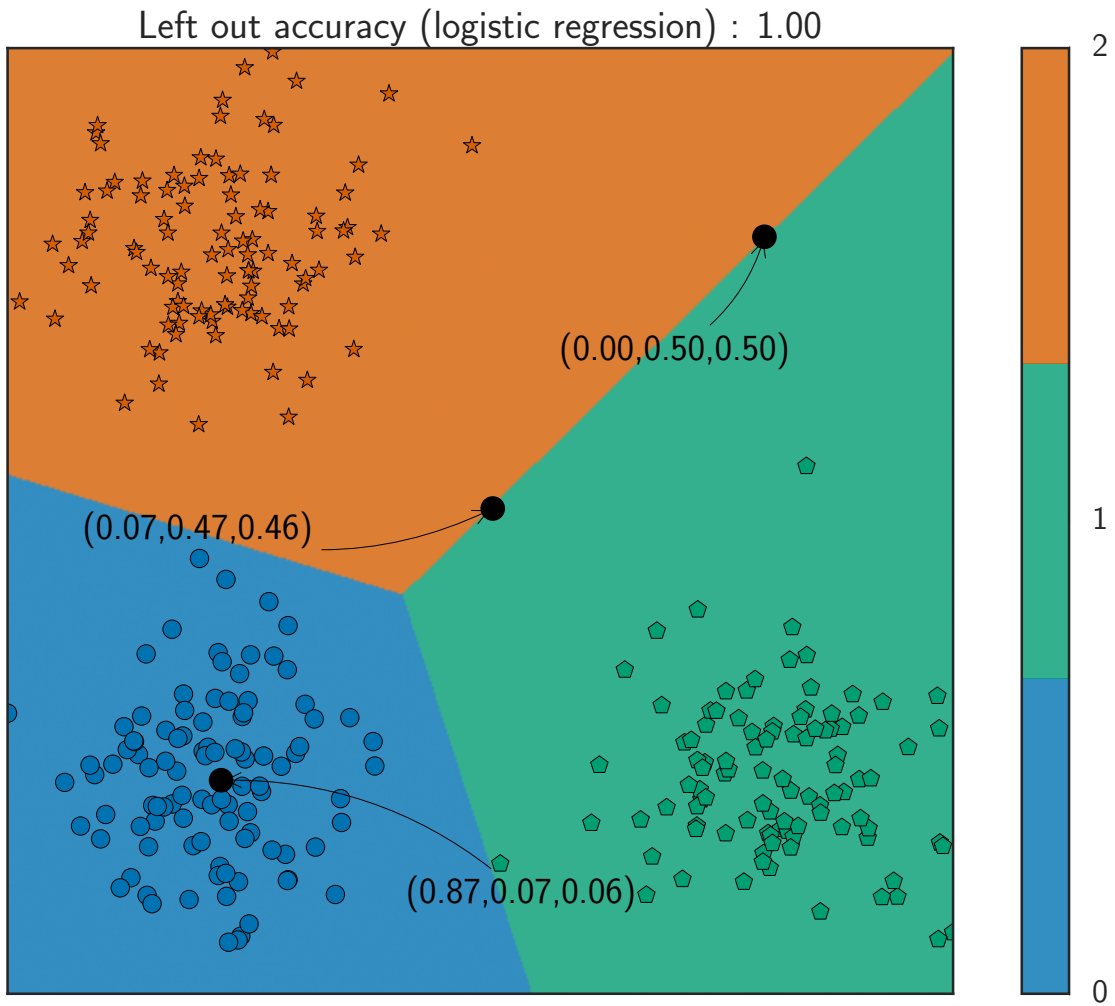
- les estimations $\langle \mathbf{x}, \hat{\boldsymbol{\theta}}^{(k)} \rangle$ de $\mathbb{P}(y = k|\mathbf{x})$ peuvent être négatives
- effet masque
- ne pas utiliser (sauf peut-être en binaire)

Sommaire

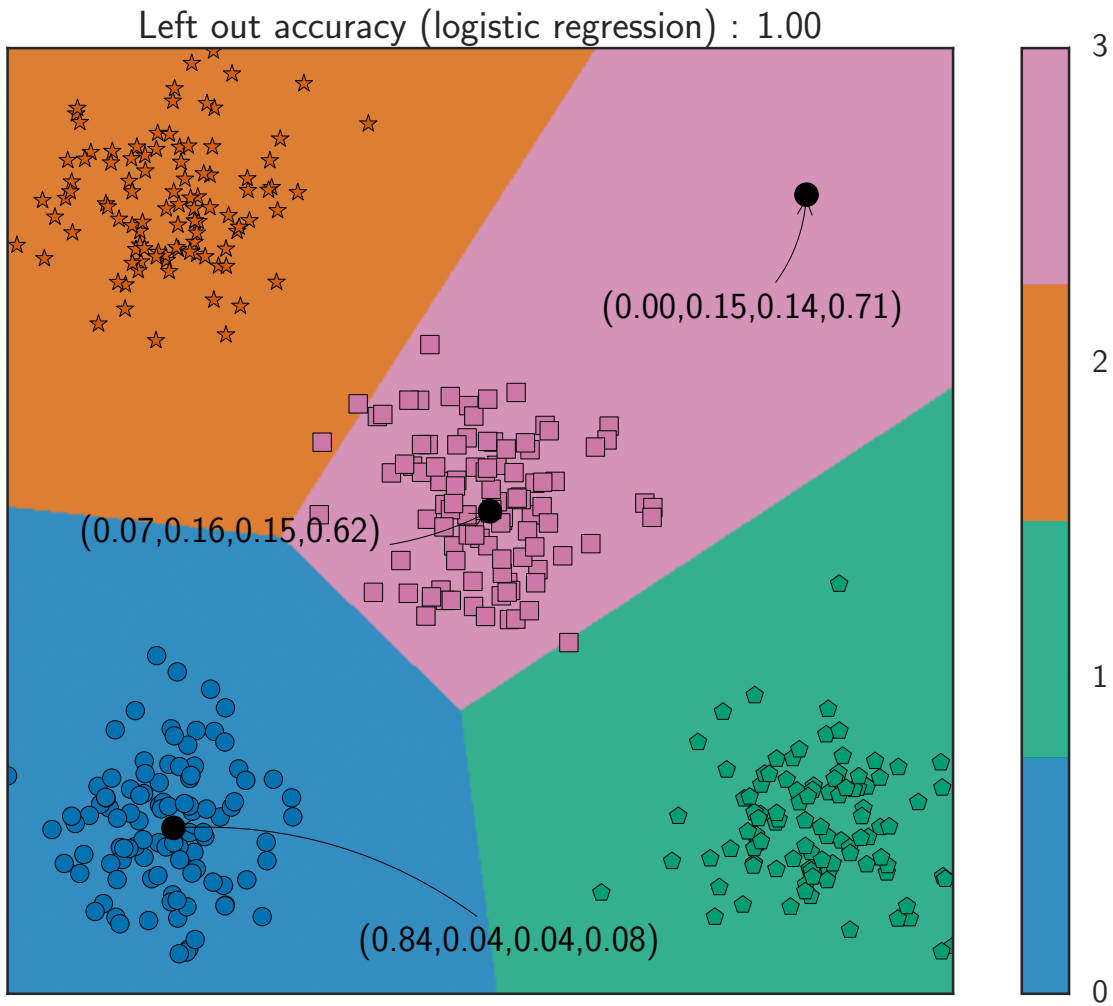
Un exemple avec deux classes



Un exemple avec trois classes



Un exemple avec quatre classes



L'approche gaussienne

Supposons que les densités des observations des classes 0 et 1 sont gaussiennes isotropes (de même variance) :

$$\varphi_{\mu_0,\sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^{n/2}} \exp\left(-\frac{\|\mathbf{x}-\mu_0\|^2}{2\sigma^2}\right)$$

$$\varphi_{\mu_1,\sigma}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sigma^{n/2}} \exp\left(-\frac{\|\mathbf{x}-\mu_1\|^2}{2\sigma^2}\right)$$

La règle de Bayes donne alors : $\mathbb{P}(y=k|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|y=k)\mathbb{P}(y=k)}{\mathbb{P}(\mathbf{x})}$, puis

$$\log\left(\frac{\mathbb{P}(y=0|\mathbf{x})}{\mathbb{P}(y=1|\mathbf{x})}\right) = \theta_0 + \langle \boldsymbol{\theta}, \mathbf{x}\rangle$$

Exo: donner θ_0 et $\boldsymbol{\theta}$ en fonction de μ_1, μ_0, σ et $\pi_0 = \mathbb{P}(y=0)$

Régression logistique : cas binaire

Ainsi, il est raisonnable de modéliser le log-ratio (des probabilités conditionnelles par classe) linéairement

$$\log\left(\frac{\mathbb{P}(y=0|\mathbf{x})}{\mathbb{P}(y=1|\mathbf{x})}\right) = \boldsymbol{\theta}^\top \mathbf{x}, \quad \text{avec } \boldsymbol{\theta} \in \mathbb{R}^p$$

Rem: constantes incorporées si besoin dans les variables explicatives

Sous cette hypothèse la règle de classification est simplement :

$$\begin{cases} \text{si } \langle \boldsymbol{\theta}, \mathbf{x} \rangle \leqslant 0, \text{on étiquette } 1 \text{ au point } \mathbf{x} \\ \text{si } \langle \boldsymbol{\theta}, \mathbf{x} \rangle > 0, \text{on étiquette } 0 \text{ au point } \mathbf{x} \end{cases}$$

Rem: en binaire il peut être plus simple de modéliser les classes par des -1 (au lieu des 1) et des $+1$ (au lieu de 0), le classifieur étant alors $h_{\boldsymbol{\theta}}(\mathbf{x}) = \text{sign}(\langle \boldsymbol{\theta}, \mathbf{x} \rangle)$, et non $h_{\boldsymbol{\theta}}(\mathbf{x}) = -(1 + \text{sign}(\langle \boldsymbol{\theta}, \mathbf{x} \rangle))/2$

Régression logistique et probabilités

On peut alors estimer les probabilités conditionnelles facilement :

$$\mathbb{P}(y=0|\mathbf{x}) = \frac{\exp\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)}{1+\exp\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)}$$
$$\mathbb{P}(y=1|\mathbf{x}) = \frac{1}{1+\exp\left(\langle \boldsymbol{\theta}, \mathbf{x}\rangle\right)}$$

Ainsi connaissant une estimation de $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ on pourra proposer comme estimation des probabilités :

$$\widehat{\mathbb{P}}(y=0|\mathbf{x}) = \frac{\exp\left(\langle \hat{\boldsymbol{\theta}}, \mathbf{x}\rangle\right)}{1+\exp\left(\langle \hat{\boldsymbol{\theta}}, \mathbf{x}\rangle\right)}$$
$$\widehat{\mathbb{P}}(y=1|\mathbf{x}) = \frac{1}{1+\exp\left(\langle \hat{\boldsymbol{\theta}}, \mathbf{x}\rangle\right)}$$

Régression logistique et vraisemblance

Maximisation de la (log-)vraisemblance ($\ell(\boldsymbol{\theta})$)

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log(\mathbb{P}(y = y_i | \mathbf{x} = x_i, \boldsymbol{\theta})) \\ &= \sum_{i=1}^{n} \sum_{k=0}^{1} \mathbb{1}_{\{y_i = k\}} \log(\mathbb{P}(y = k | \mathbf{x} = x_i, \boldsymbol{\theta}))\end{aligned}$$

On résout : $\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^p} \ell(\boldsymbol{\theta}) \quad \left(= \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} -\ell(\boldsymbol{\theta}) \right)$

Exo: montrer qu'avec $u_i = (2\mathbb{1}_{\{y_i=0\}} - 1)$

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\mathbb{1}_{\{y_i=0\}} \langle \boldsymbol{\theta}, x_i \rangle - \log[1 + \exp(\langle \boldsymbol{\theta}, x_i \rangle)] \right) \\ &= \sum_{i=1}^{n} \left(\log[1 + \exp(-u_i \langle \boldsymbol{\theta}, x_i \rangle)] \right)\end{aligned}$$

Interprétation des coefficients

Pour un vecteur $\mathbf{x} \in \mathbb{R}^p$, et une variable $j \in [\![1, p]\!]$, augmenter $\mathbf{x}_j$ d'une unité (en gardant toutes les autres composantes fixes)

- augmente le log-ratio de la classe 0 d'un facteur additif $\hat{\theta}_j$
- augmente la probabilité de la classe 0 d'un facteur multiplicatif $\exp(\hat{\theta}_j)$

Rem: en pratique il n'est pas toujours possible de ne faire varier qu'une seule variable

Régression logistique et méthode Newton

- Hessienne calculable : on peut appliquer la méthode de Newton
- Approche descente par coordonnées envisageable aussi (notamment si l'on régularise)

Pour les détails techniques, calculs de la Hessienne etc. Hastie *et al.* (2009, page 120)

http://www-stat.stanford.edu/~tibs/ElemStatLearn/

Sommaire

Détour : de deux à plusieurs classes

On peut passer du cadre binaire au multi-classe pour toute méthode, *e.g.*, il suffit de tester :

- "un contre tous" (en : **One-vs.-rest/all**) : créer un classifieur par classe, et produire un score (par exemple une probabilité). Prédire la classe de score maximum
  Coût : K classifieurs
- "un contre un" (en : **One-vs.-one**) : calculer un classifieur pour toutes les $K(K-1)/2$ paires. Prédire la classe qui gagne le plus de "duels"
  Coût : $K(K-1)/2$ classifieurs

Rem: dans `sklearn` la descente par coordonnée avec LIBLINEAR
http://www.csie.ntu.edu.tw/~cjlin/liblinear/
utilise"one-vs-rest"

Détour : de deux à plusieurs classes

On peut passer du cadre binaire au multi-classe pour toute méthode, *e.g.*, il suffit de tester :

- "un contre tous" (en : **One-vs.-rest/all**) : créer un classifieur par classe, et produire un score (par exemple une probabilité). Prédire la classe de score maximum
  Coût : K classifieurs
- "un contre un" (en : **One-vs.-one**) : calculer un classifieur pour toutes les $K(K-1)/2$ paires. Prédire la classe qui gagne le ~~plus de "duels"~~ championnat
  Coût : $K(K-1)/2$ classifieurs

Rem: dans `sklearn` la descente par coordonnée avec LIBLINEAR
http://www.csie.ntu.edu.tw/~cjlin/liblinear/
utilise"one-vs-rest"

Régression logistique (I)

De nouveau, on modélise les probabilités conditionnelles des classes, ou plutôt leur log-ratio, par des quantités linéaires :

$$\log\left(\frac{\mathbb{P}(y=k\,|\mathbf{x})}{\mathbb{P}(y=K-1|\mathbf{x})}\right) = \langle \boldsymbol{\theta}_k, \mathbf{x}\rangle, \text{ avec } \forall k \in [\![0, K-2]\!], \boldsymbol{\theta}_k \in \mathbb{R}^p$$

Paramètre globale : $\Theta = [\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{K-2}, \underbrace{\boldsymbol{\theta}_{K-1}}_{=0}] \in \mathbb{R}^{p\times K}$

Rem: constante incorporée si besoin dans les variables explicatives

Sous cette hypothèse les séparatrices inter-classes sont **linéaires** : la bascule entre deux classes a lieu le long d'hyperplans. On teste par exemple k et k' ainsi

$$\begin{cases} \text{si } \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}'_k, \mathbf{x}\rangle \geqslant 0, \text{ on préfère } k \text{ à } k' \text{ au point } \mathbf{x} \\ \text{si } \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}'_k, \mathbf{x}\rangle < 0, \text{ on préfère } k' \text{ à } k \text{ au point } \mathbf{x} \end{cases}$$

Régression logistique (II)

On peut alors estimer les probabilités conditionnelles facilement :

$$\text{Pour } k = 0, \dots, K-2 : \mathbb{P}(y = k|\mathbf{x}) = \frac{\exp\left(\langle \boldsymbol{\theta}_k, \mathbf{x} \rangle\right)}{1 + \sum_{l=0}^{K-2} \exp\left(\langle \boldsymbol{\theta}_l, \mathbf{x} \rangle\right)}$$

$$\text{Pour } k = K-1 : \mathbb{P}(y = K-1|\mathbf{x}) = \frac{1}{1 + \sum_{l=0}^{K-2} \exp\left(\langle \boldsymbol{\theta}_l, \mathbf{x} \rangle\right)}$$

Règle de classification : choisir la classe qui a la plus probable

$$h_\Theta(\mathbf{x}) = \underset{k \in [\![0, K-1]\!]}{\arg\max}\, \widehat{\mathbb{P}}(y = k|\mathbf{x})$$

Rem: numériquement le problème devient plus dur (à écrire et à traiter) qu'en binaire, *cf.* Hastie *et al.* (2009)

Régression logistique et vraisemblance

Formulation possible : on note $\Theta = [\boldsymbol{\theta}_0, \dots, \boldsymbol{\theta}_{K-1}] \in \mathbb{R}^{p \times K}$

$$\text{Pour } k \in [\![0, K-1]\!] : \mathbb{P}(y = k|\mathbf{x}) = \frac{\exp\left(\langle \boldsymbol{\theta}_k, \mathbf{x} \rangle\right)}{\sum_{l=0}^{K-1} \exp\left(\langle \boldsymbol{\theta}_l, \mathbf{x} \rangle\right)}$$

Rem: on retrouve le précédent modèle en prenant $\boldsymbol{\theta}_{K-1} = 0$

Maximisation de la (log-)vraisemblance ($\ell(\boldsymbol{\theta})$)

$$\begin{aligned}\ell(\Theta) &= \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbb{1}_{\{y_i = k\}} \log(\mathbb{P}(y = k|\mathbf{x} = x_i, \Theta)) \\ &= \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbb{1}_{\{y_i = k\}} \langle \boldsymbol{\theta}_k, x_i \rangle - \log\left(\sum_{k=0}^{K-1} \exp(\langle \boldsymbol{\theta}_k, x_i \rangle)\right)\end{aligned}$$

On résout : $\hat{\Theta} \in \underset{\Theta \in \mathbb{R}^{p \times K}}{\arg\max}\, \ell(\Theta) \quad \left(= \underset{\Theta \in \mathbb{R}^{p \times K}}{\arg\min} -\ell(\boldsymbol{\theta}) \right)$

Numériques / astuces

- Régularisation avec une pénalisation ℓ_1, ℓ_2^2 (on travaille avec l'opposée de la log-vraisemblance pour avoir un problème convexe) :

$$\underset{\Theta \in \mathbb{R}^{p \times K}}{\arg\min} (-\ell(\Theta) + \text{pen}(\Theta))$$

$$\text{avec } -\ell(\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=0}^{K-1} \exp(\langle \boldsymbol{\theta}_k, x_i \rangle) \right) - \sum_{k=0}^{K-1} \mathbb{1}_{\{y_i = k\}} \langle \boldsymbol{\theta}_k, x_i \rangle$$

- Gestion atypique de la constante pour les matrices sparses
- Algorithmes (LBFGS, SAG, Prox-Newton, etc.)

Avantages / Inconvénients : régression logistique

Avantages

- connu pour avoir des probabilités estimées bonnes
- séparations inter-classes linéaires

Inconvénients

- classification binaire plus facile
- problème d'optimisation plus complexe (temps de calcul)
- parfois géré par la technique du "un contre tous" et non par le cas logistique multinomial (surtout si K est petit)

Sommaire

Formulation GLM

Ce sont des modèles de la forme :

$$\mu = \mathbb{E}(y|\mathbf{x}) = f^{-1}(\langle \mathbf{x}, \boldsymbol{\theta} \rangle)$$

c'est-à-dire qu'on modélise :

$$\mathbf{y} \approx f^{-1}(X\boldsymbol{\theta})$$

On appelle **fonction de lien** la fonction f (on $g = f^{-1}$)

Quelques modèles classiques

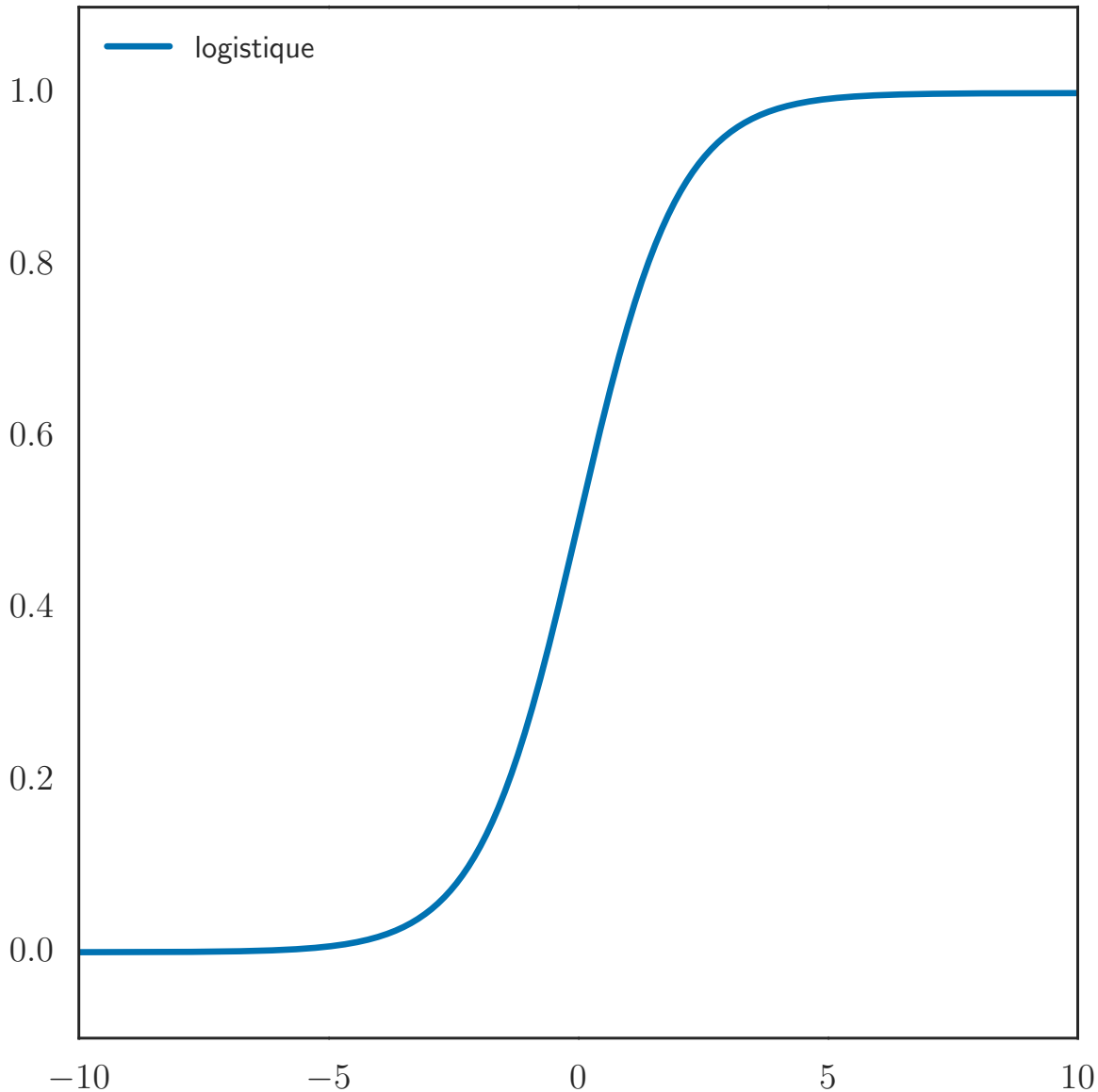
Distribution	Fonction de lien f	Espérance
Gaussienne	Id	$X\boldsymbol{\theta}$
Poisson	$\log$	$\exp(X\boldsymbol{\theta})$
Binomial	logit	$\mathrm{logistic}(X\boldsymbol{\theta})$
$\ldots$	$\ldots$	$\ldots$

où les fonctions logit et $\mathrm{logistic}$ sont définies par

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

$$\mathrm{logistic}(t) = \mathrm{logit}^{-1}(t) = \frac{1}{1+\exp(-t)} = \frac{\exp(t)}{\exp(t)+1}$$

Visualisation : fonction logistique



Sommaire

Points non abordés : extensions possibles

- Robustesse : attache aux données ℓ_1, régression quantile, moyennes tronquées, etc.
- Méthodes gloutones ( : *greedy*)
- boosting/bagging
- Point de vue bayesien
- Arbres / forêts (classification surtout)
- K-plus proches voisins
- SVM (classification)
- Réseaux de neurones (classification surtout)

Références I

- T. Hastie, R. Tibshirani, and J. Friedman.
  The elements of statistical learning.
  Springer Series in Statistics. Springer, New York, second edition, 2009.
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Sommaire

Le cas $y = \{-1, 1\}$

- Lien avec la probabilité $\mathbb{P}(y = 1|x)$
- Remarque :

$$\begin{aligned}\mathbb{E}[y|\mathbf{x}] &= -1 \times \mathbb{P}(y = -1|x) + 1 \times \mathbb{P}(y = 1|x) \\ &= 2\mathbb{P}(y = 1|x) - 1\end{aligned}$$

- Modèle généralisé $y \in \{-1, 1\}$,

$$\widehat{\mathbb{P}(y = 1|x)} = \text{logistic}(\mathbf{x}^\top \boldsymbol{\theta})$$

$$g(\mathbf{x}^\top \boldsymbol{\theta}) = 2\,\text{logistic}(\mathbf{x}^\top \boldsymbol{\theta}) - 1$$

- Classifieur associé :

$$h_{\boldsymbol{\theta}}(x) = \begin{cases} +1 & \text{if } \text{logistic}(\mathbf{x}^\top \boldsymbol{\theta}) > \frac{1}{2} \Leftrightarrow g(\mathbf{x}^\top \boldsymbol{\theta}) > 0 \\ -1 & \text{sinon} \end{cases}$$

- Vrai risque de notre classifieur (notant $\mathbb{P}_1 = \mathbb{P}(y=1)$ et $\mathbb{P}_{-1} = \mathbb{P}(y=-1)$)

$$\mathbb{P}\left(g(\mathbf{x}^\top \boldsymbol{\theta}) \leqslant 0 \middle| y=1\right)\mathbb{P}_1 + \mathbb{P}\left(g(\mathbf{x}^\top \boldsymbol{\theta}) > 0 \middle| y=-1\right)\mathbb{P}_{-1}$$

- Risque empirique de notre classifieur :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i \neq h_{\boldsymbol{\theta}}(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i g(x_i^\top \boldsymbol{\theta})<0}$$

- Minimisation du risque empirique :

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i g(x_i^\top \boldsymbol{\theta})<0}$$

- Problème d'optimisation difficile (non convexe)

Kullback-Leibler

- Divergence de Kullback-Leibler :
$$\mathrm{KL}(\mathcal{B}(\mathbb{P}(y=1|X)), \mathcal{B}(\tilde{g}(X^\top \boldsymbol{\theta}))$$

$$= \mathbb{E}_X \Big[\mathbb{P}(y=1|X) \log \frac{\mathbb{P}(y=1|X)}{\tilde{g}(X^\top \boldsymbol{\theta})} + (1-\mathbb{P}(y=1|X)) \log \frac{1-\mathbb{P}(y=1|X)}{1-\tilde{g}(X^\top \boldsymbol{\theta})} \Big]$$

$$= \mathbb{E}_X \big[-\mathbb{P}(y=1|X) \log(\tilde{g}(X^\top \boldsymbol{\theta})) - (1-\mathbb{P}(y=1|X)) \log(1-\tilde{g}(X^\top \boldsymbol{\theta})) \big] + C_{X,y}$$

- contrepartie empirique :
$$-\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{y_i=1} \log(\tilde{g}(x_i^\top \boldsymbol{\theta})) + \mathbb{1}_{y_i=-1} \log(1-\tilde{g}(x_i^t \boldsymbol{\theta})) \right)$$

Minimisation du risque empirique

- Minimisation possible si $\tilde{g}$ est lisse...

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{y_i=1}\log(\tilde{g}(x_i^\top\boldsymbol{\theta}))+\mathbb{1}_{y_i=-1}\log(1-\tilde{g}(x_i^\top\boldsymbol{\theta}))\right)$$

- Choix classiques $\tilde{g}$:

$$\tilde{g}(t)=\frac{e^t}{1+e^t} \qquad \text{logit or logistic}$$

$$\tilde{g}(t)=F_{\mathcal{N}}(t) \qquad \text{probit}$$

$$\tilde{g}(t)=1-e^{-e^t} \qquad \text{log-log}$$

Régression logistique

- Modèle : $\tilde{g}(t) = \frac{e^t}{1+e^t}$
- La loi de Bernoulli $\mathcal{B}(\tilde{g}(t))$ satisfait

$$\frac{\mathbb{P}(y=1)}{\mathbb{P}(y=-1)} = e^t \Leftrightarrow \log \frac{\mathbb{P}(y=1)}{\mathbb{P}(y=-1)} = t$$

- Opposée de la $\log$-vraisemblance :

$$
\begin{aligned}
& -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{y_i=1}\log(\tilde{g}(x_i^\top\boldsymbol{\theta}))+\mathbb{1}_{y_i=-1}\log(1-\tilde{g}(x_i^t\boldsymbol{\theta}))\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{y_i=1}\log\frac{e^{x_i^t\boldsymbol{\theta}}}{1+e^{x_i^\top\boldsymbol{\theta}}}+\mathbb{1}_{y_i=-1}\log\frac{1}{1+e^{x_i^\top\boldsymbol{\theta}}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\left(1+e^{-y_i(x_i^\top\boldsymbol{\theta})}\right)
\end{aligned}
$$

- Fonction convexe et dérivable de $\boldsymbol{\theta}$
- Optimisation facile

Classifieur associé

- $$h_{\boldsymbol{\theta}}(x) = \begin{cases} 1 & \text{if } \frac{e^{\mathbf{x}^\top \boldsymbol{\theta}}}{1+e^{\mathbf{x}^\top \boldsymbol{\theta}}} > 1/2 \Leftrightarrow \mathbf{x}^\top \boldsymbol{\theta} > 0 \\ -1 & \text{otherwise} \end{cases}$$
- Lien entre le coût de prédiction empirique et la vraisemblance :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i \neq h_{\boldsymbol{\theta}}(x_i)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i (x_i^\top \boldsymbol{\theta}) < 0} \leqslant \frac{1}{n \log 2} \sum_{i=1}^{n} \log\left(1 + e^{-y_i (x_i^\top \boldsymbol{\theta})}\right)$$

- Preuve :

$$\mathbb{1}_{y_i (x_i^\top \boldsymbol{\theta}) < 0} \leqslant \frac{\log\left(1 + e^{-y_i (x_i^\top \boldsymbol{\theta})}\right)}{\log 2}$$

- Convexification du risque