Semaine IMT Grand-Est

Introduction à l'apprentissage automatique, la science de l'intelligence artificielle

Séance 3

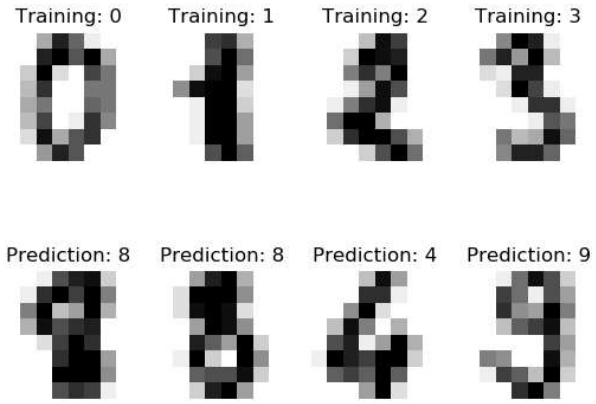
Théorie statistique de la décision et applications

Frédéric Sur

https://members.loria.fr/FSur/enseignement/IMT_GE/

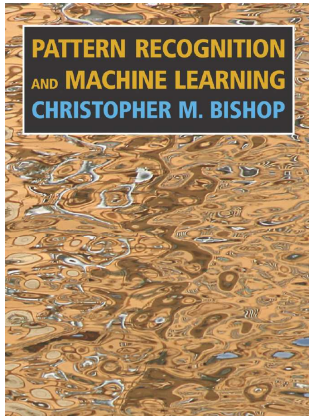
Plan

Classification supervisée



Aujourd'hui : classification supervisée
Cadre : théorie statistique de la décision

Bibliographie



Source principale (& illustrations sauf mention contraire) :
C. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006
→ dans toutes les bonnes médiathèques

Notations

K classes : $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$
classes = partition de l'ensemble des observations possibles
Exemple : $\mathcal{C}_1$ =1, $\mathcal{C}_2$ =2, ..., $\mathcal{C}_{10}$ =0

N observations (base d'apprentissage) : $x_1, \ldots, x_N \in \mathbb{R}^d$
Exemple : $x \in \mathbb{R}^{256}$ = vecteur des niveaux de gris d'une imagette 16×16
on connaît la classe d'appartenance de chaque x_i : y_i $(= \mathcal{C}_1, \mathcal{C}_2, \ldots)$

Question du jour : prédiction de la classe d'une nouvelle observation x ?

Soit f un classifieur : $f(x) = \mathcal{C}_1$ ou $f(x) = \mathcal{C}_2$, etc.
→ une fonction sur l'ensemble des entrées possibles qui prédit la classe d'appartenance

Problème : le classifieur peut faire des erreurs
Exemple : $f(x) = \mathcal{C}_1$ alors que $y = \mathcal{C}_2$

Formalisation probabiliste

Probabilités :

- $p(\mathcal{C}_k)$ *probabilité a priori (prior)*, $\sum_k p(\mathcal{C}_k) = 1$
  → ce qu'on suppose sans connaître d'observations
  Exemple OCR : $p(\text{a}) = 0.07$, $p(\text{b}) = 0.01$, $p(\text{c}) = 0.03\ldots$
- $p(x|\mathcal{C}_k)$ *proba. conditionnelle* (en fait, *densité - vraisemblance*)
  → probabilité pour qu'une obs. tirée dans la classe $\mathcal{C}_k$ vaille x
  Exemple : $p(x|\mathcal{C}_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$
- $p(\mathcal{C}_k|x)$ *probabilité (vraisemblance) a posteriori*
  → probabilité de la classe $\mathcal{C}_k$ étant donnée l'observation x

Théorème de Bayes :

$$\forall\, 1 \leq k \leq K,\ p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x)}$$

et $p(x) = \sum_{k=1}^{K} p(x, \mathcal{C}_k) = \sum_{k=1}^{K} p(x|\mathcal{C}_k)p(\mathcal{C}_k)$

Erreur de classification

Pour simplifier, $K = 2$ dans la suite

Le classifieur f définit une partition de l'ensemble des observations possibles en régions $\mathcal{R}_i$ telles que :
$\mathcal{R}_i = \{x, f(x) = \mathcal{C}_i\}$

Problème : les partitions $(\mathcal{C}_i)$ et $(\mathcal{R}_i)$ ne coïncident pas
(à cause des erreurs de classification)

Proposition : calcul de la proportion moyenne d'erreur **théorique**

$$\begin{aligned} E_{\text{err}} &= E_{X,Y}\left(1_{f(x)\neq y}\right) = \iint 1_{f(x)\neq y} p(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int 1_{f(x)\neq \mathcal{C}_1} p(x,\mathcal{C}_1)\mathrm{d}x + \int 1_{f(x)\neq \mathcal{C}_2} p(x,\mathcal{C}_2)\mathrm{d}x \\ &= \int_{\mathcal{R}_2} p\left(x,\mathcal{C}_1\right)\mathrm{d}x + \int_{\mathcal{R}_1} p\left(x,\mathcal{C}_2\right)\mathrm{d}x \end{aligned}$$

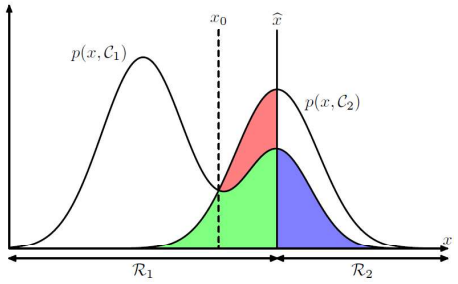
Question : existe-t-il un classifieur f minimisant E_{err} ?

Minimisation de l'erreur moyenne

$$E_{\text{err}} = \int_{\mathcal{R}_2} p\left(x,\mathcal{C}_1\right)\mathrm{d}x + \int_{\mathcal{R}_1} p\left(x,\mathcal{C}_2\right)\mathrm{d}x$$

Illustration : $x \in \mathbb{R}$, $\mathcal{R}_1 = \{x \in \mathbb{R}, x < \hat{x}\}$, $\mathcal{R}_2 = \{x \in \mathbb{R}, x > \hat{x}\}$

Question : comment fixer $\hat{x}$ de manière à minimiser E_{err} ?



E_{err} = rouge + vert + bleu

Or vert + bleu = Cste
lorsque $\hat{x}$ varie

→ minimum atteint pour $\hat{x} = x_0$.

Classifieur de Bayes

(ce raisonnement se généralise à $x \in \mathbb{R}^d$, et $\mathcal{R}_i \neq$ intervalles)

Conséquence : la règle de classification minimisant l'erreur moyenne est

$$f(x) = \text{argmax}_{\mathcal{C}_k}\, p(x, \mathcal{C}_k) = \text{argmax}_{\mathcal{C}_k}\, p(x)\, p(\mathcal{C}_k|x)$$

Classifieur de Bayes – maximum a posteriori (MAP)

$$f(x) \;=\; \text{argmax}_{\mathcal{C}_k}\, p(\mathcal{C}_k|x) \;=\; \text{argmax}_{\mathcal{C}_k}\, p(\mathcal{C}_k)\, p(x|\mathcal{C}_k)$$

ou, dans le cas biclasse :
$f(x) = \mathcal{C}_1$ ssi $\dfrac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} > \dfrac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}$
(cf test du rapport des vraisemblances)

En pratique ? (1)

Problème : comment estimer $p(\mathcal{C}_k)$ et $p(x|\mathcal{C}_k)$ à partir du jeu de données disponible (ensemble d'observations classifiées) ?

$p(\mathcal{C}_k)$:

- information connue a priori
  exemple : OCR
- ou fréquence estimée à partir de la base d'observations
  exemple : $p(\mathcal{C}_k) = \dfrac{\#\{x_i \in \mathcal{C}_k,\ 1 \leq i \leq N\}}{N}$
  où # désigne le cardinal d'un ensemble
- ou, dans le cas où les $p(\mathcal{C}_k)$ sont égaux :
  le classifieur de Bayes se simplifie en $f(x) = \text{argmax}_k p(x|\mathcal{C}_k)$
  → règle du maximum de vraisemblance (ML)

En pratique ? (2)

$p(x|\mathcal{C}_k)$: toute une partie de l'apprentissage non-supervisé concerne l'*estimation de densités de probabilité*

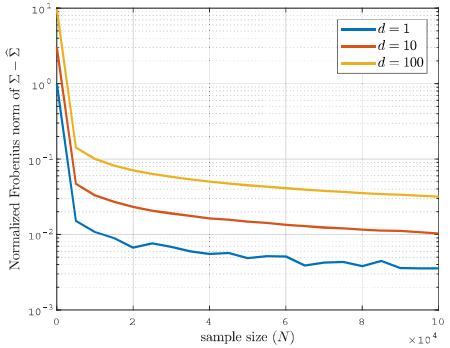
si $x \in \mathbb{R}^d$ avec $d \ll$ grand $\gg$: attention, *curse of dimensionality !*

Expérience :
estimation matrice de covariance

N observations x_n i.i.d. $\mathcal{N}_d(0, \Sigma)$

$$\widehat{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} x_n x_n^T$$

$||\Sigma - \widehat{\Sigma}||^2_{\text{NFn}} = \frac{1}{d} \sum_{i=1}^{d} \lambda_i^2$
où λ_i v.p. de $\Sigma - \widehat{\Sigma}$



→ en pratique, on a intérêt à réduire la dimension d...

Classifieur naïf de Bayes

Une manière de battre la malédiction de la dimensionnalité...

Si $x = (x^1, x^2, \ldots, x^d) \in \mathbb{R}^d$, on suppose les composantes (conditionnellement) statistiquement indépendantes

Donc : $p(x|\mathcal{C}_k) = \prod_{i=1}^{d} p(x^i|\mathcal{C}_k)$

Gros avantage : plutôt qu'estimer la distribution $p(x|\mathcal{C}_k)$ sur $\mathbb{R}^d$, on estime les d distributions $p(x^i|\mathcal{C}_k)$ sur $\mathbb{R}$.

Classifieur naïf de Bayes

$$f(x) \;=\; \text{argmax}_{\mathcal{C}_k}\, p(\mathcal{C}_k) \prod_{i=1}^{d} p(x^i|\mathcal{C}_k)$$

Exemple : classifieur naïf Gaussien
→ on suppose les distributions $p(x^i|\mathcal{C}_k)$ gaussiennes
→ deux paramètres : μ_k, σ_k

Plan

Estimateur de distribution aux plus proches voisins

On va chercher à estimer $\phi(x_0) = p(x_0|\mathcal{C}_k)\ldots$

Pour estimer $\phi(x_0)$, on considère une boule B_{x_0} centrée en x_0, contenant P échantillons de $\mathcal{C}$ parmi $M = \#\mathcal{C}$:

- $\int_B \phi(x)\mathrm{d}x \simeq \frac{P}{M}$ (par la loi des grands nombres)
- et $\int_B \phi(x)\mathrm{d}x \simeq \phi(x_0) \times \text{Volume}(B_{x_0})$ (si ϕ constant sur B_{x_0})

Donc : $\frac{P}{M} \simeq \int_B \phi(x)\mathrm{d}x \simeq \phi(x_0) \times \text{Volume}(B_{x_0})$

D'où l'estimateur des P plus proches voisins :

$$\phi(x) = \frac{P}{M\, V_P(x)}$$

où $V_P(x)$: volume d'une boule contenant les P p.p.v.

→ hypothèse ϕ constant sur B, donc il faudrait une boule « pas trop grosse » (malédiction dimensionnalité ?)

→ en fait, utilisé pour la classification supervisée...

Classification aux P plus proches voisins

Base de données : N observations $x_1, \ldots, x_N$ et classes associées parmi $\mathcal{C}_1, \ldots, \mathcal{C}_K$.

→ N_1 observations dans $\mathcal{C}_1, \ldots$, N_K dans $\mathcal{C}_K$, t.q. $\sum_k N_k = N$

Problème : étant donnée une nouvelle observation x, comment prédire sa classe ?

Parmi les P p.p.v. de x : P_1 dans $\mathcal{C}_1, \ldots$, P_K dans $\mathcal{C}_K$ $(\sum_k P_k = P)$

Par estimation aux P-p.p.v. : $p(x|\mathcal{C}_k) = \frac{P_k}{N_k V_P(x)}$

De plus : $p(\mathcal{C}_k) = \frac{N_k}{N}$

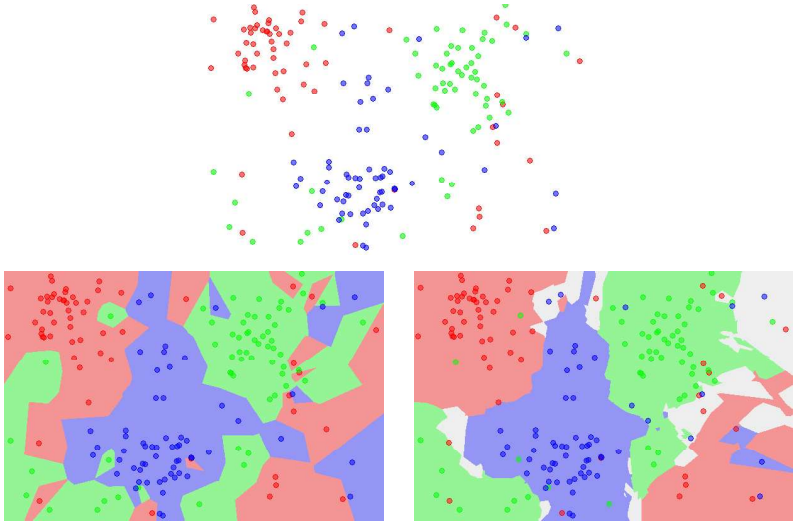
Classifieur MAP :

$\text{argmax}_k p(\mathcal{C}_k)p(x|\mathcal{C}_k) = \text{argmax}_k P_k/(NV_P(x)) = \text{argmax}_k P_k$

Définition : le *classifieur aux P-p.p.v.* affecte une nouvelle observation à la classe majoritaire parmi les P observations les plus proches

→ implémente le classifieur de Bayes sous hypothèses (très) simplificatrices

Exemple



1-p.p.v. 5-p.p.v.

La régression logistique

Dans le cas bi-classe :

$$p(\mathcal{C}_1|x) = \frac{p(\mathcal{C}_1)p(x|\mathcal{C}_1)}{p(\mathcal{C}_1)p(x|\mathcal{C}_1) + p(\mathcal{C}_2)p(x|\mathcal{C}_2)} = \frac{1}{1 + \frac{p(\mathcal{C}_2)p(x|\mathcal{C}_2)}{p(\mathcal{C}_1)p(x|\mathcal{C}_1)}}$$

Avec $f(x) = \log\left(\frac{p(x|\mathcal{C}_2)}{p(x|\mathcal{C}_1)}\right) + \log\left(\frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}\right)$: $p(\mathcal{C}_1|x) = \frac{1}{1+e^{-f(x)}}$

Définition : fonction logistique (ou sigmoïde) : $\sigma(t) = \frac{1}{1+e^{-t}}$

Hypothèse simplificatrice : $f(x) = \beta_0 + \beta_1 \cdot x$

→ c'est aussi une manière de contrer la malédiction de la dimensionnalité, en réduisant le nombre de paramètres à estimer

→ si on sait estimer β_0 et β_1, le classifieur de Bayes devient :
classifieur de la régression logistique :
x dans $\mathcal{C}_1$ ssi $p(\mathcal{C}_1|x) > 1/2 \iff f(x) > 0$
(séparation des deux classes par un hyperplan)

Régression logistique et estimation des paramètres

Hypothèse simplificatrice : $f(x) = \beta_0 + \beta_1 \cdot x$

Remarque : cas de classes gaussiennes de même variance
si $p(x|\mathcal{C}_k) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}$,
$f(x) = -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$
→ d'où $f(x) = \beta_0 + \beta_1 \cdot x$ avec :
$\beta_0 = \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$
$\beta_1 = \Sigma^{-1}(\mu_1 - \mu_2)$

Estimation de (β_0, β_1) :
maximisation de la log-vraisemblance conditionnelle

$$\ell_{(x_i,y_i)_{1\leq i\leq N}}(\beta_0,\beta_1) = \log\prod_{i=1}^{n}\Big((p(\mathcal{C}_1|x_i))^{y_i}\left(1-p(\mathcal{C}_1|x_i)\right)^{1-y_i}\Big)$$
$$= \sum_{i=1}^{n}\Big(-(1-y_i)(\beta_0+\beta_1 x_i) - \log(1+\exp(\beta_0+\beta_1 x_i))\Big)$$

avec **ici** : $y_i = 1 \leftrightarrow x_i \in \mathcal{C}_1$, $y_i = 0 \leftrightarrow x_i \in \mathcal{C}_2$, $p(\mathcal{C}_2|x_i) = 1 - p(\mathcal{C}_1|x_i)$

→ fonction **concave**, donc maximum unique

Illustration

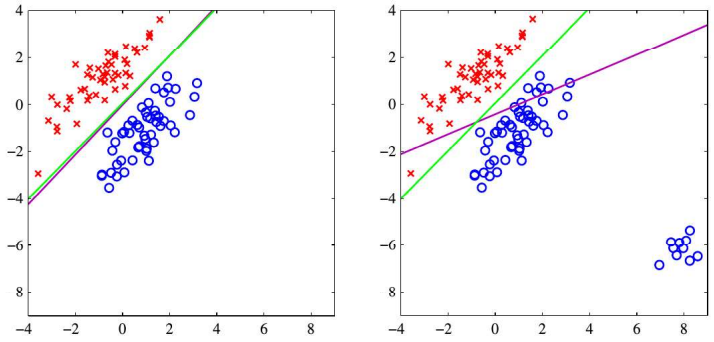


Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

Illustration : C. Bishop, *Pattern Recognition and Machine Learning*, Springer 2006

→ on reviendra sur cette propriété en séance 4

Plan

Fonction discriminante

Notion de **fonction discriminante** :
f_k tel que $f(x) = \text{argmax}_k f_k(x)$

- MAP (classifieur de Bayes, **théoriquement** optimal) :

$$f(x) = \text{argmax}_k p(\mathcal{C}_k|x) = \text{argmax}_k p(\mathcal{C}_k)p(x|\mathcal{C}_k)$$

  Problème : on ne connaît pas les distributions de probabilité
  $\rightarrow$ on ajoute des hypothèses : classifieur naïf de Bayes, régression logistique, k-p.p.v.. . .

- Cours suivants : autres hypothèses, autres fonctions discriminantes

Conclusion – Résumé

Théorie statistique de la décision, et mise en œuvre :

- le classifieur bayésien (MAP) minimise l'erreur moyenne de classification (classifieur **idéal**)
- simplification si *prior* uniformes : classification au maximum de vraisemblance (ML)
- simplification si composantes indépendantes pour $x \in \mathbb{R}^d$: classifieur naïf de Bayes
- simplification si $p(\mathcal{C}_1|x) = \sigma(\beta_0 + \beta_1 \cdot x)$: régression logistique
- simplification si $p(x|\mathcal{C}_1)$ ne varie pas trop et suffisamment d'observations : classification aux K plus proches voisins