



Introduction à la reconnaissance des formes et au *Machine Learning*

Omar Galarraga, *Ph.D.*

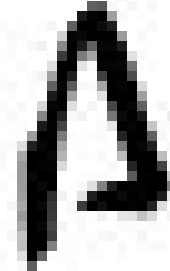
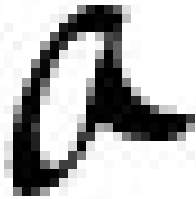
omar.galarraga@ugecam.assurance-maladie.fr

Pattern Recognition & Biometrics

Prof. Sonia García (*coord.*)

2023-2024

Qu'est-ce que la reconnaissance des formes et le *Machine Learning* ?



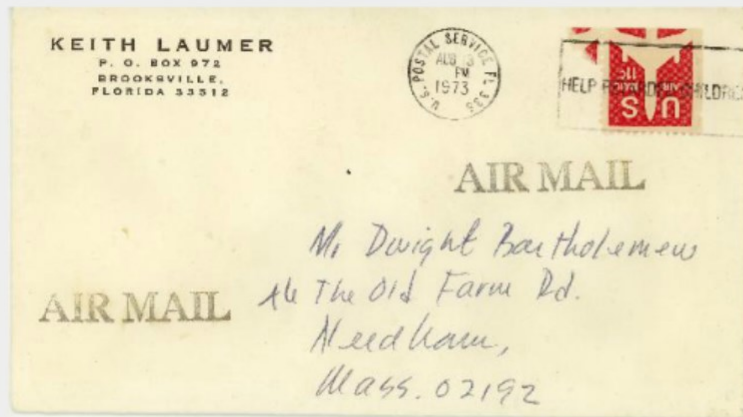
La reconnaissance des formes et le *Machine Learning*

- Trouver des relations significatives entre les données pour la prise de décision
- Apprentissage automatique / apprentissage statistique
- Rendre les machines capables d'apprendre des tâches « humaines »
- Estimation / Filtrage / Prédiction
- Classification / Régression



Applications : reconnaissance des caractères manuscrits

Service postal



Numérisation de documents

Pakistan Peoples Party

To the officials and members of Pakistan Peoples Party I say that I was honoured to lead you. He leader could be as proud of their party. Their dedication, dash and discipline to the mission of Quaid-e-Azam Zulfikar Ali Bhutto for a Federal Democratic and Egalitarian Pakistan as I have been proud of you. I salute your courage and your sense of honour. I salute you for standing by your sister through their military dictatorship.

I fear for the future of Pakistan. Please continue the fight against extremism, dictatorship, poverty and ignorance.

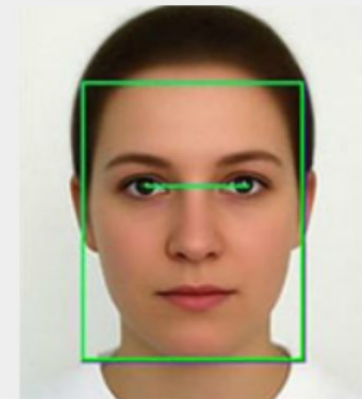
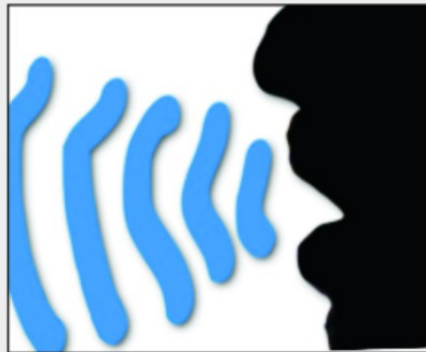
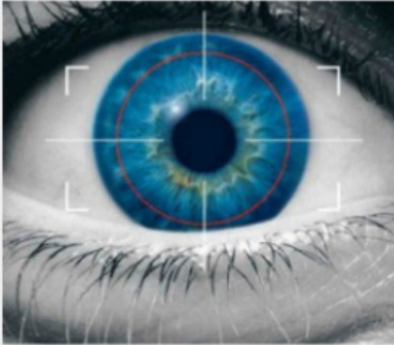
I would like my husband Prof. M. J. Baidan to lead you in this intense period until you and he decide what to do. I say this because he is a man of courage and honour. He spent 11 1/2 years in prison without bonding despite torture. He has the political stature to keep our party united.

I wish all of you sincere in fulfilling the manifesto of your party and in serving the down-trodden, discriminated and oppressed people of Pakistan. Dedicate yourselves to freeing them from poverty and backwardness as you have done in the past.

Zameer Bhutto
October 16, 2007

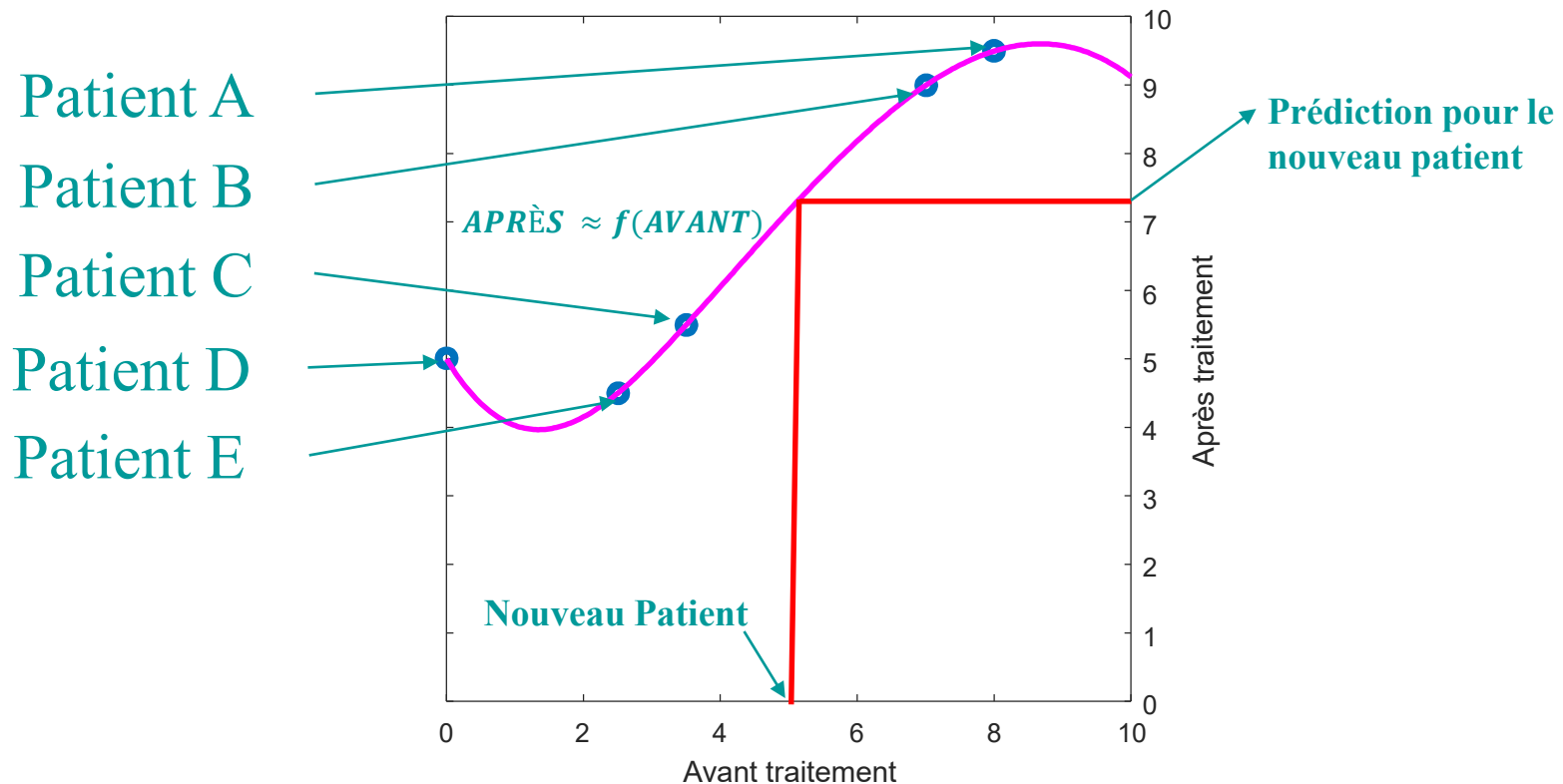


Applications : Biométrie



Apprentissage supervisé

- Classes / valeurs de sortie désirée connues
- Objectif : minimiser l'erreur entre la sortie désirée et la sortie estimée

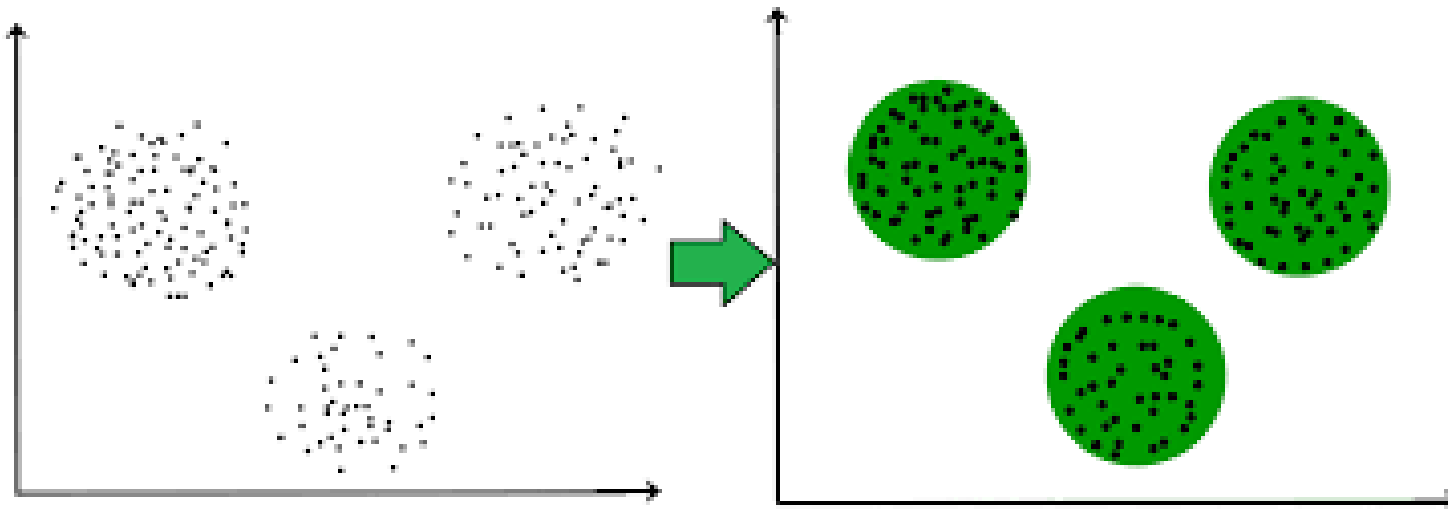


Exemple : reconnaissance du visage



Apprentissage non supervisé

- Données non étiquetées
 - **Classes inconnues**
 - **Valeurs de la sortie désirée inconnues**
- Clustering



Etapes du Machine Learning

Prétraitement

Extraction des
caractéristiques

Classification /
Régression



Autres paradigmes d'apprentissage

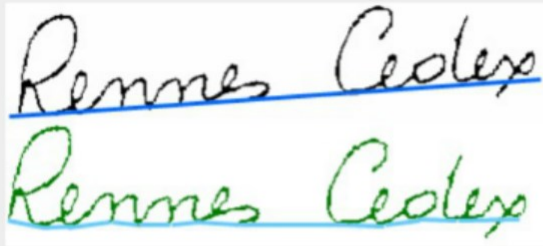
- Apprentissage profond
- Apprentissage par renforcement
- Apprentissage auto-supervisé



Prétraitement

- Préparer, « nettoyer » les données

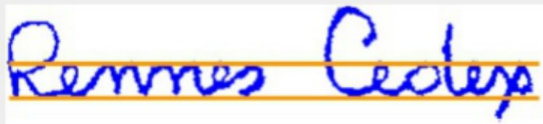
Baseline Correction



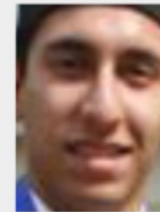
Compression



Normalization

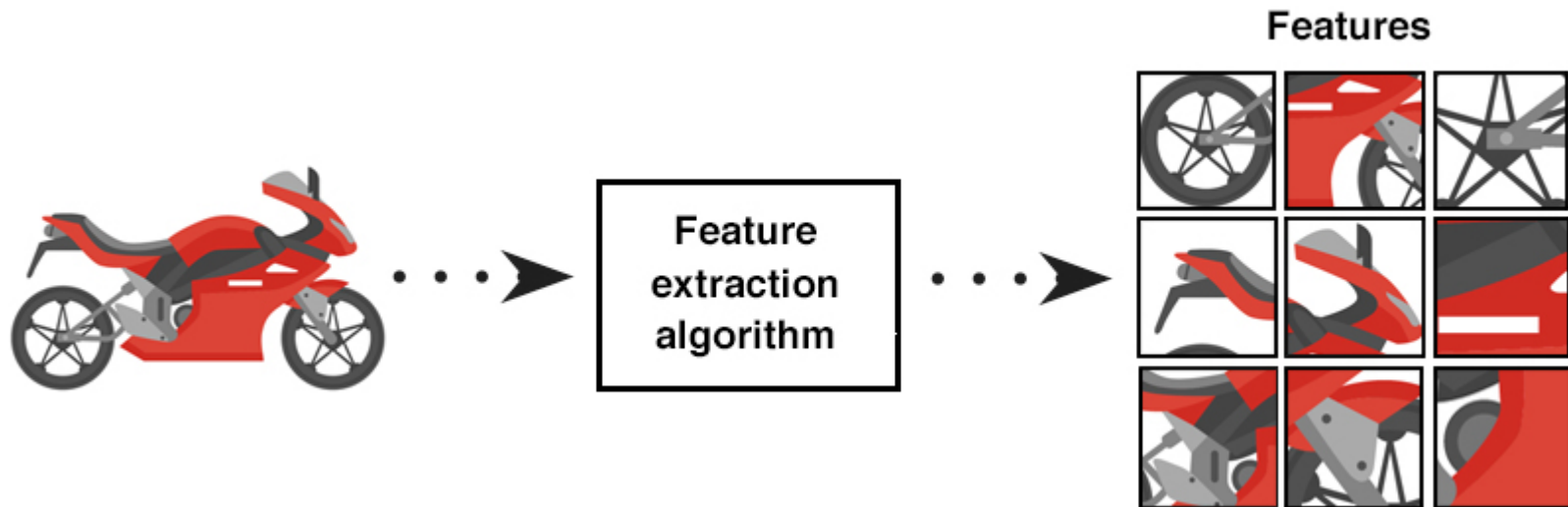


Background removal



Extraction des caractéristiques

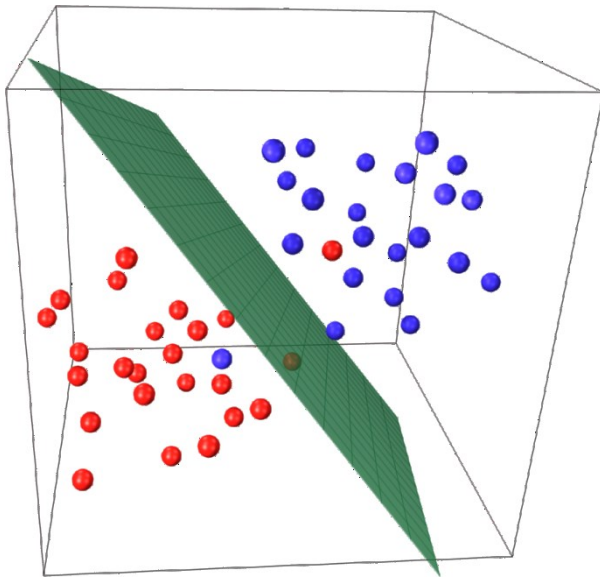
- Variables importantes, prédictives



Classification vs Régression

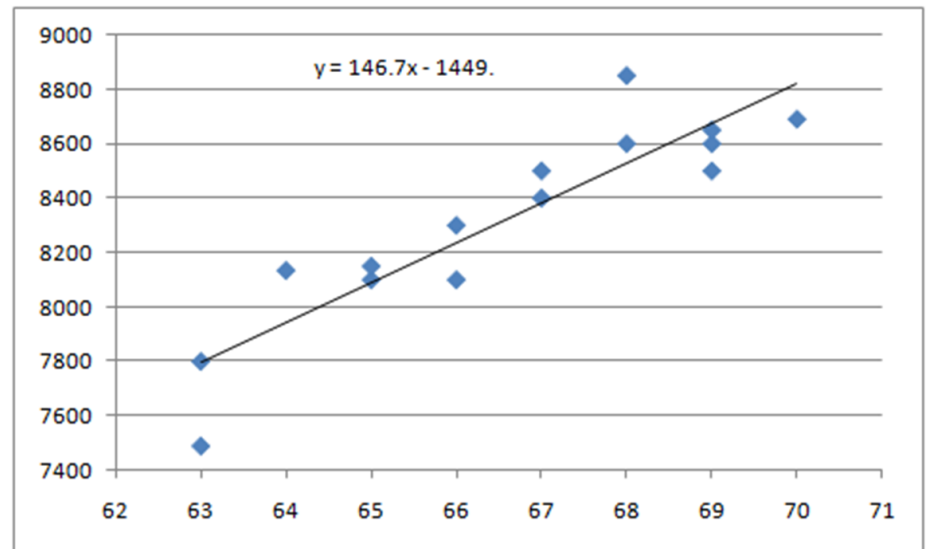
Classification

- Catégories (sortie discrète)



Régression

- Sortie continue



Classification Naïve Bayésienne

- Etant donné $x = (x_1, \dots, x_l)^T$ un vecteur de caractéristiques pouvant appartenir à des classes c_1, \dots, c_K
- $P(c_1|x), \dots, P(c_K|x)$ probabilités « *a posteriori* »
- $\operatorname{argmax}_{i=1, \dots, K} [P(c_i|x)]$ nous donnera la classe la plus probable d'appartenance pour x



Règle de décision : 2 classes

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$

L'observation x sera affecté à la classe c_a avec :

$$a = \begin{cases} 1 \text{ si } P(c_1|x) > P(c_2|x) \\ 2 \text{ si } P(c_1|x) < P(c_2|x) \\ 1 \text{ ou } 2 \text{ si } P(c_1|x) = P(c_2|x) \end{cases}$$


$$P(x|c_1)P(c_1) \leq P(x|c_2)P(c_2)$$



Estimation de la probabilité conditionnelle

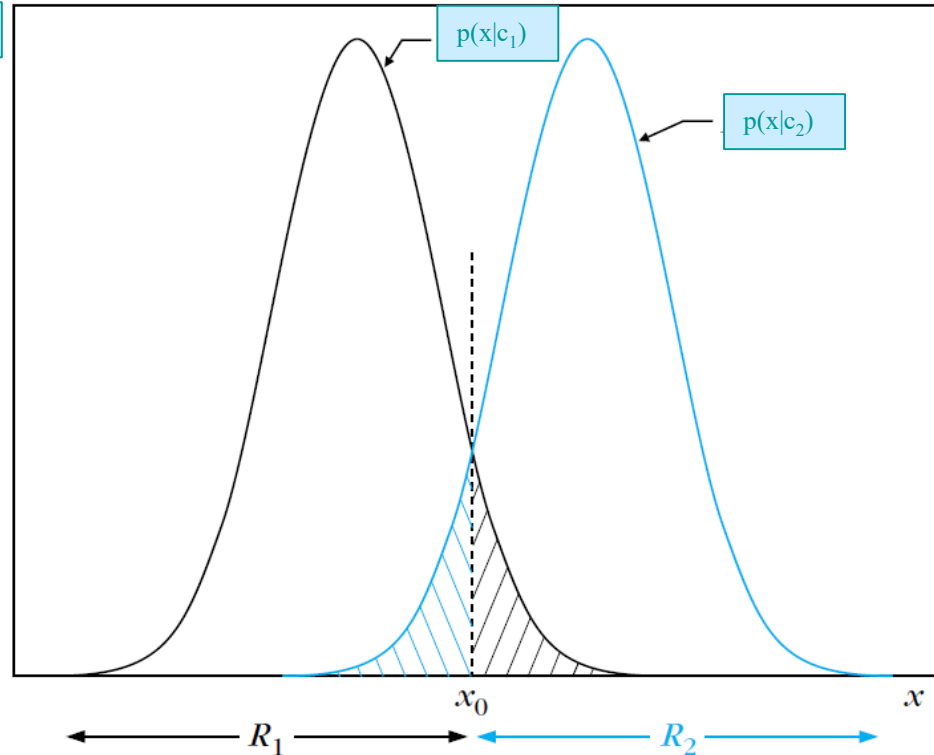
- Si les variables $x_j \forall j = 1, \dots, l$ sont « normales » et indépendants

$$P(x|c_i) = \prod_{j=1}^l \frac{1}{\sqrt{2\pi\sigma_{j,i}^2}} e^{-\frac{(x - \mu_{j,i})^2}{2\sigma_{j,i}^2}}$$

$$\text{Si } l = 1 \rightarrow \frac{P(c_1)}{2\pi\sigma_1^2} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \leq \frac{P(c_2)}{2\pi\sigma_2^2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$



Risque d'erreur (minimum)

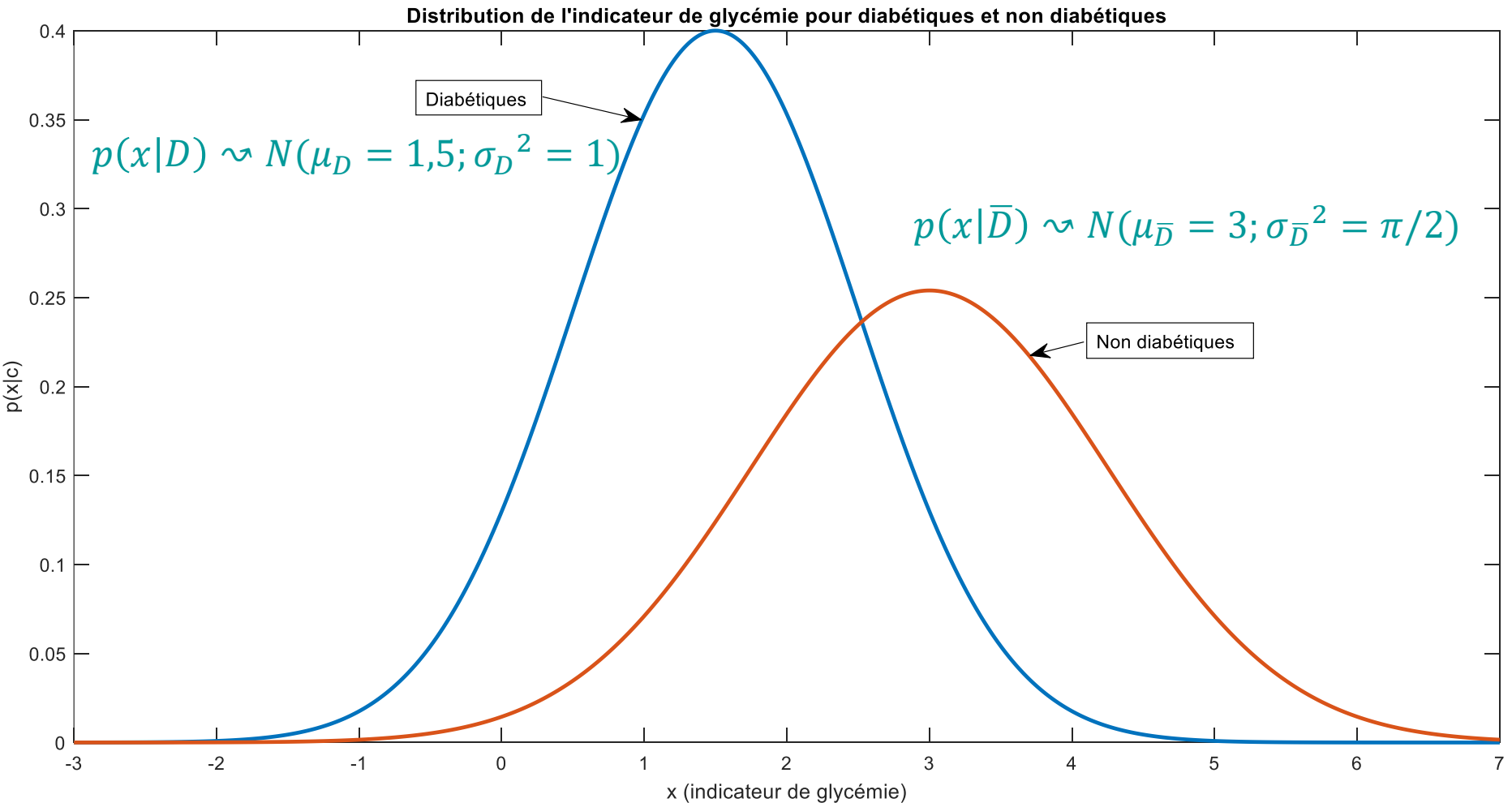


Source : Koutoumbas & Theodoridis (2009)

$$P(\text{Erreur}) = P(c_2) \int_{-\infty}^{x_0} p(x|c_2) dx + P(c_1) \int_{x_0}^{\infty} p(x|c_1) dx$$



Exemple : détection du diabète



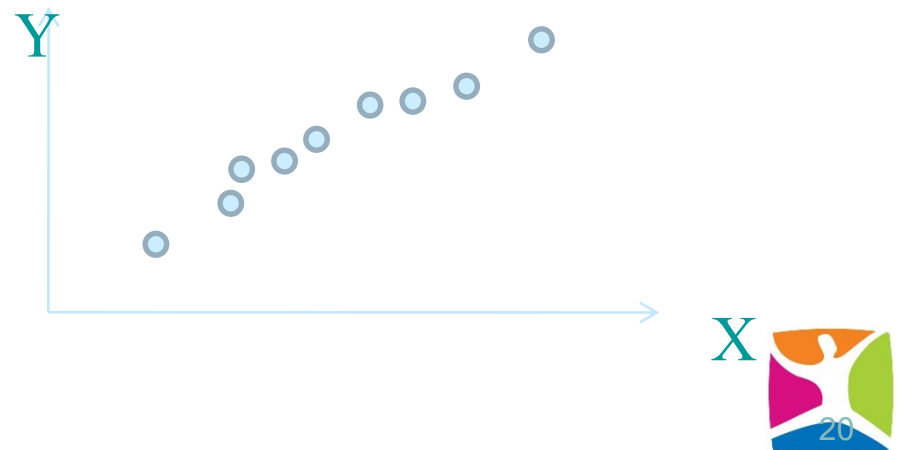
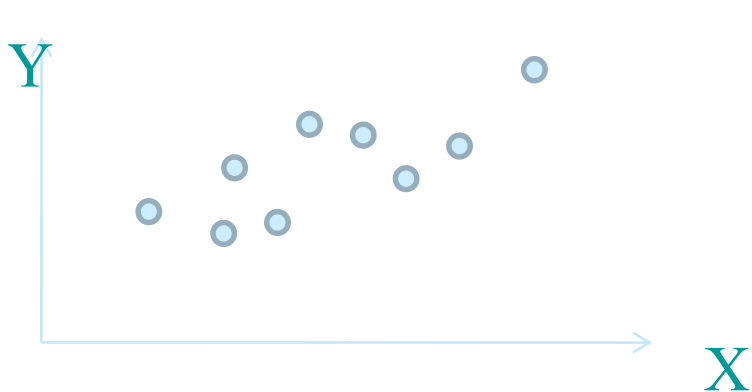
Régression

- Modéliser la relation entre deux (ou plus) variables (continues)
- Estimation, filtrage, prédiction...
- Types :
 - ▶ **Linéaire**
 - ▶ **Multilinéaire**
 - ▶ **Non linéaire**
 - ▶ **Etc...**



Régression linéaire simple

- Cas de deux variables quantitatives
 - X : facteur ou variable indépendante fixée par l'expérimentateur
 - Y : réponse ou variable dépendante
- Dispersion



Modèle linéaire

$$y = ax + b + \varepsilon$$

$$\hat{y} = ax + b$$

On veut minimiser l'erreur

$$J(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$$



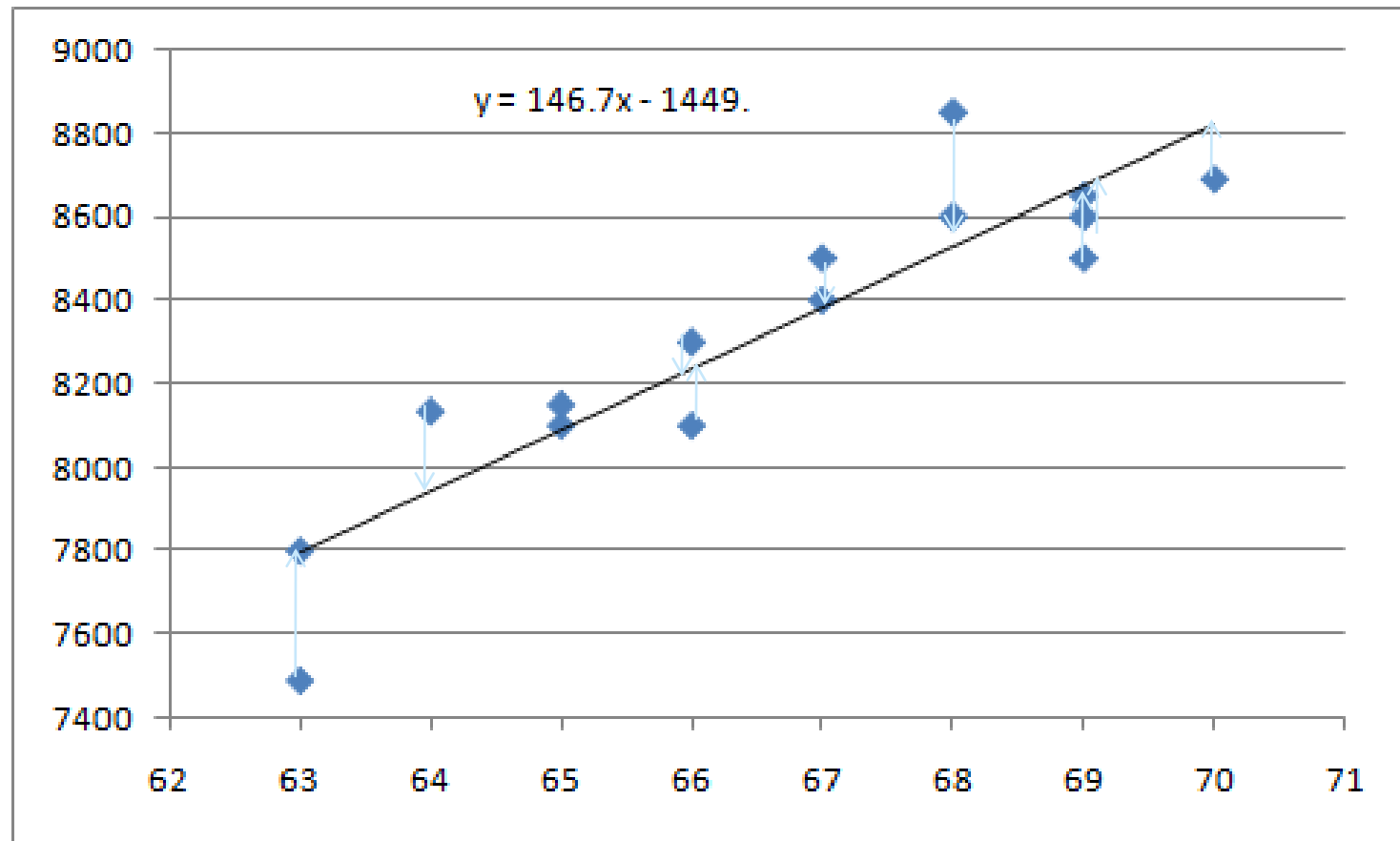
Solution des moindres carrés

$$a = \frac{\text{Cov}(x,y)}{s_x^2} = r_{xy} \frac{s_y}{s_x}$$

$$b = \bar{y} - a\bar{x}$$

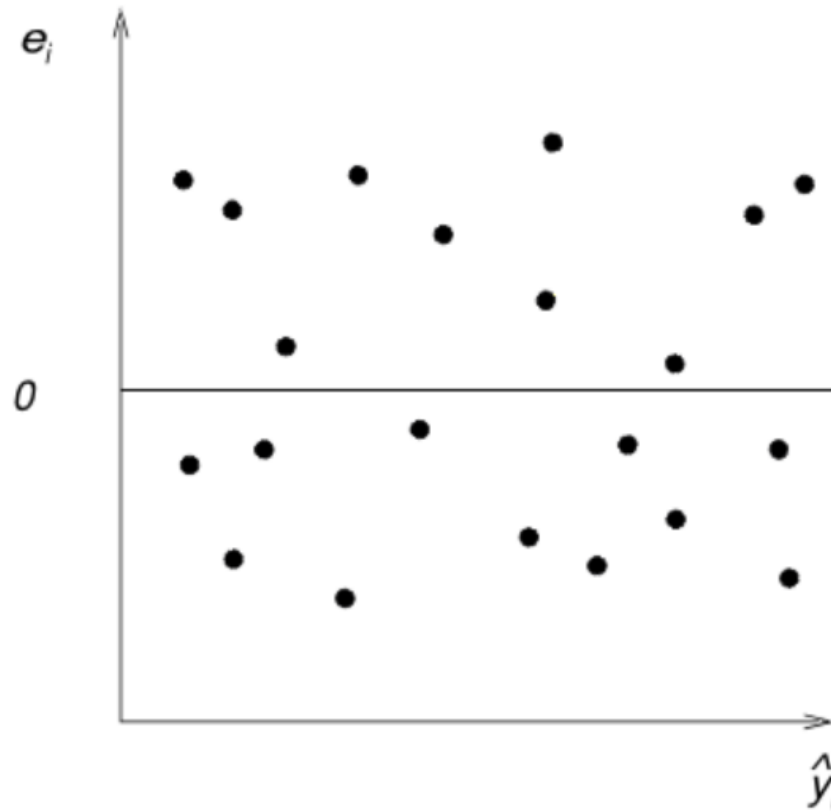


Résidus



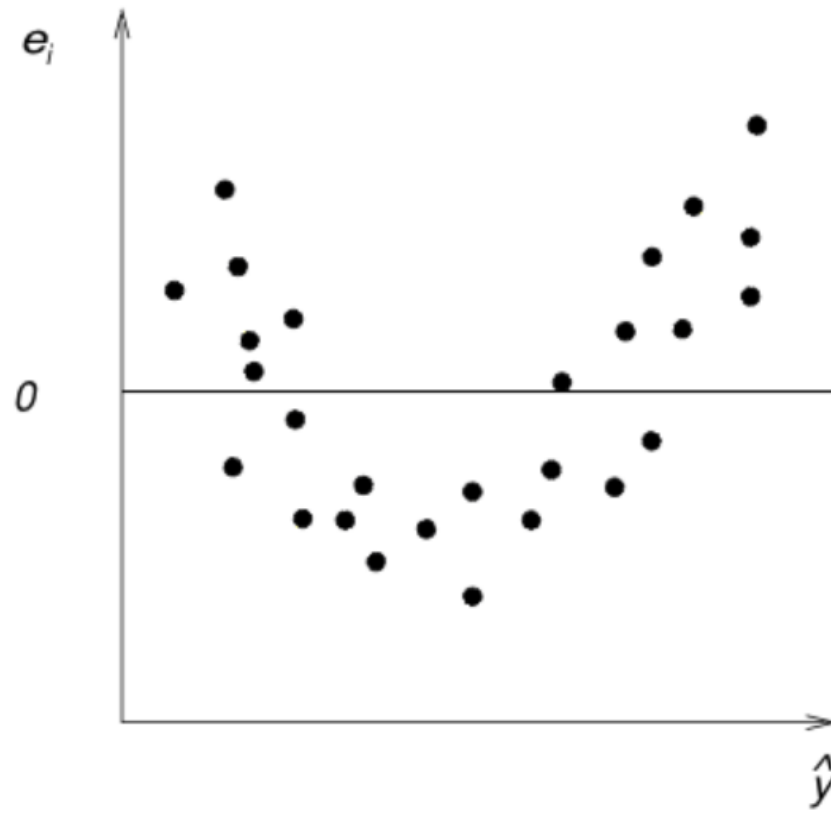
Graphique des résidus

Convenable :



Analyse des résidus

Le modèle linéaire n'est pas approprié :



Régression multilinéaire

$$Y = AX + \epsilon$$

$$A = YX(XX^T)^{-1}$$

$$Y = \begin{pmatrix} y_{1,1} & \cdots & y_{1,n} \\ \vdots & \ddots & \vdots \\ y_{N_{out},1} & \cdots & y_{N_{out},n} \end{pmatrix}$$

$$X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{N_{in},1} & \cdots & x_{N_{in},n} \\ 1 & \cdots & 1 \end{pmatrix}$$

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,N_{in}+1} \\ \vdots & \ddots & \vdots \\ a_{N_{out},1} & \cdots & a_{N_{out},N_{in}+1} \end{pmatrix}$$



Régularisation

- Pour limiter l'impact des « *outliers* »

$$J(A) = \sum_{i=1}^n (y_i - Ax_i)^2 + \lambda \sum_{j=1}^{N_{out}} \sum_{k=1}^{N_{in}} (A_{j,k})^2$$



Evaluation de la performance du modèle

Classification

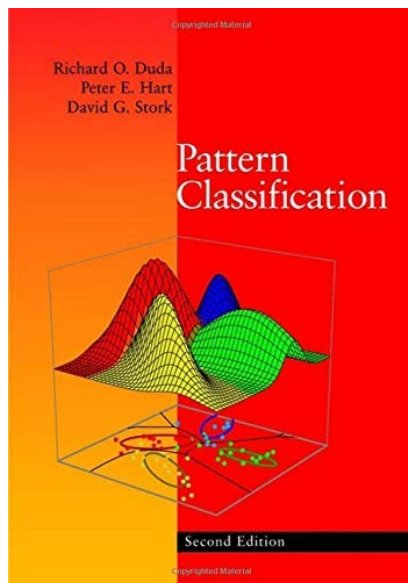
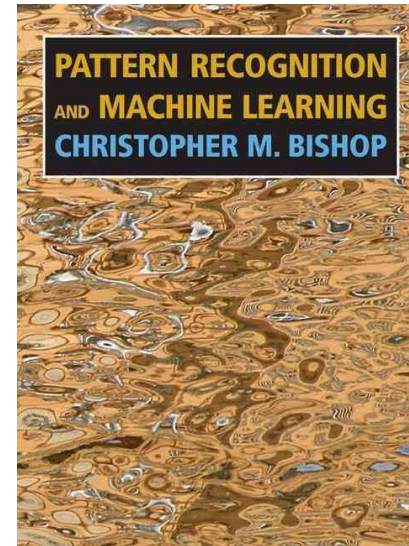
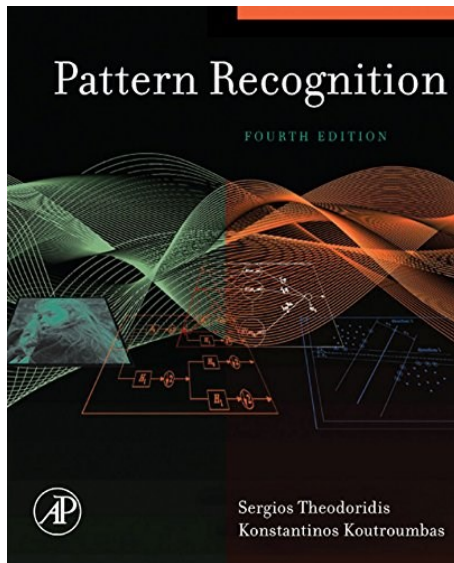
- Matrice de confusion
- Prévalence (Accuracy)
- Sensibilité
- Spécificité
- Entropie croisée (binaire)

Régression

- Erreur quadratique moyenne (MSE)
- Root-MSE (RMSE)
- Ecart moyen (MAE)
- Coefficient de détermination (R^2)



Lectures pour approfondir



MOOCS

- Machine Learning in Python with sci-kit learn. FUN-MOOC. INRIA.
- Machine Learning. COURSERA. Deep.AI Andrew Ng
- Deep Learning for Coders. FAST.AI
- From Deep Learning to Stable Diffusion. FAST.AI
- Deep Learning Specialization. Deep.AI (COURSERA). *[Plusieurs cours payants]*.

