# Apprentissage Artificiel

## Chapitre 5 : L'apprentissage Bayésien et son approximation

Ewa Kijak

ESIR/Université de Rennes 1





## Sommaire







## Le principe inductif ERM

**Choisir l'hypothèse minimisant le risque empirique** (*Empirical Risk Minimization*)
Le risque empirique est la perte moyenne mesurée *sur l'échantillon d'apprentissage* $\mathcal{S}$ :

$$R_{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} l(\boldsymbol{u}_i, h(\boldsymbol{x}_i))$$

L'idée est que l'hypothèse qui s'accorde le mieux aux données, si elles sont représentatives, décrit correctement le monde en général.





## Le principe inductif bayésien

**Choisir l'hypothèse la plus probable étant donné** $\mathcal{S}$.

- On suppose qu'il est possible de définir une distribution de probabilité sur les hypothèses.
- La connaissance du domaine préalable à l'apprentissage s'exprime sous la forme d'une distribution de probabilité *a priori* sur les hypothèses.
- L'échantillon d'apprentissage est alors considéré comme une information modifiant la distribution de probabilité sur $\mathcal{H}$.
- On peut choisir l'hypothèse la plus probable *a posteriori* : *Maximum A Posteriori* (*MAP*).





## Sommaire







## Apprentissage bayésien de classes (reconnaissance statistique des formes)

**Notations :**

- classes : $\mathcal{C} = \{\omega_i \mid i = 1, .., C\}$
- ensemble d'apprentissage $\mathcal{S}$ de taille $m$, composé de $m_i$ points $(\boldsymbol{x}_i\ ,\ \omega_i)$ par classe $\omega_i$.
- espace de représentation : $\mathbb{R}^d$.

**Problème à résoudre :**
Attribuer une classe parmi $C$ à un point quelconque $\boldsymbol{x}$ de $\mathbb{R}^d$, à partir de la seule connaissance de l'ensemble d'apprentissage.

## La formule de Bayes

**Formule de Bayes :**

$$P(\omega_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \omega_i)P(\omega_i)}{p(\boldsymbol{x})}$$

- $p(\boldsymbol{x} \mid \omega_i)$ est la **densité de probabilité** de la classe $\omega_i$ au point $\boldsymbol{x}$, aussi appelée **vraisemblance**
- $P(\omega_i)$ est la **probabilité a priori** de la classe $i$
- $P(\omega_i \mid \boldsymbol{x})$ est la **probabilité a posteriori** de : $\boldsymbol{x} \in \omega_i$.

On a :

$$p(\boldsymbol{x}) = \sum_{i=1}^{i=C} P(\omega_i)p(\boldsymbol{x} \mid \omega_i) \qquad \text{et} \qquad \sum_{i=1}^{i=C} P(\omega_i) = 1$$





## La règle de classification bayésienne : Règle du Maximum A Posteriori (MAP)

La **règle de classification bayésienne** ou **règle MAP** $h^*$ attribue au point $\boldsymbol{x}$ la classe $\omega^*$ de plus forte probabilité *a posteriori* d'avoir engendré $\boldsymbol{x}$ :

$$h^* \text{ choisit la classe } \omega^* = \underset{i}{\text{ArgMax}}(P(\omega_i \mid \boldsymbol{x}))$$

On cherche en effet l'hypothèse $h$ la plus probable étant donnée l'observation $\boldsymbol{x}$, c'est-à-dire *a posteriori*.

La règle MAP s'écrit encore :

$$\omega^\star = \underset{i}{\text{ArgMax}}\ p(\boldsymbol{x} \mid \omega_i)P(\omega_i)$$





## Règle du Maximum A Posteriori (MAP)

Cette règle est **optimale** : parmi toutes les règles de classification possibles, elle est celle qui a la plus petite probabilité d'erreur.

$$err(h^*) = \min_h \left[ \int_{\mathbb{R}^d} P^h_{err}(\boldsymbol{x}) d\boldsymbol{x} \right]$$

$P^h_{err}(\boldsymbol{x})$ est la probabilité que $\boldsymbol{x}$ soit mal classé par la règle $h$.
La valeur $err(h^*)$ est appelée **erreur bayésienne de classification**.

Cette règle s'appelle aussi la *règle d'erreur minimale* car elle minimise le nombre d'erreurs de classification.





## Règle du Maximum de Vraisemblance

Si toutes les hypothèses ont la même probabilité *a priori*, alors la règle de *Maximum a Posteriori* devient la règle du *Maximum de Vraisemblance* (Maximum Likelihood ou ML en anglais).

$$\omega^\star = \underset{i}{\text{ArgMax}}\ p(\boldsymbol{x}|\omega_i)$$

Cette règle revient à sélectionner la classe $\omega$ pour laquelle l'observation $\boldsymbol{x}$ est la plus probable, c'est-à-dire l'état du monde qui est le plus à même d'avoir produit l'événement $\boldsymbol{x}$.

Cela traduit l'idée simple que l'observation $\boldsymbol{x}$ n'est pas totalement fortuite et était même fortement probable étant donné l'état du monde $h$ (hypothèse).





## Un cas naïf

Si l'on suppose que les attributs de description $\{a_1, ..., a_d\}$ de l'espace d'entrée $\mathcal{X}$ sont indépendants les uns des autres, alors on peut décomposer $p(\boldsymbol{x}|\omega)$ en $p(a_1 = v_{1\boldsymbol{x}}|\omega) \dots p(a_d = v_{d\boldsymbol{x}}|\omega)$ soit

$$p(\boldsymbol{x}|\omega) = \prod_{i=1}^{d} p(a_i = v_{i\boldsymbol{x}}|\omega)$$

Le classifieur utilisant la règle du Maximum *A Posteriori* basé sur ces hypothèses est appelé *classifieur bayésien naïf*.

Les attributs de description sont en réalité rarement indépendants les uns des autres (par exemple le *poids* et la *taille*). Pourtant le classifieur bayésien naïf donne souvent des résultats proches de ceux obtenus par les meilleures méthodes connues.





## Les surfaces séparatrices

On appelle **surface séparatrice** entre $\omega_i$ et $\omega_j$ le lieu des points où les probabilités *a posteriori* d'appartenir à $\omega_i$ et à $\omega_j$ sont égales.

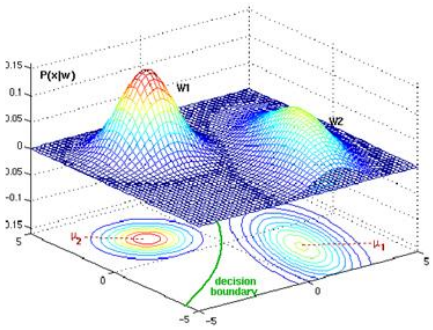
La surface séparatrice entre les classes $\omega_i$ et $\omega_j$ a pour équation :

$$P(\omega_i \mid \boldsymbol{x}) = P(\omega_j \mid \boldsymbol{x})$$

$$\frac{p(\boldsymbol{x} \mid \omega_i)P(\omega_i)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x} \mid \omega_j)P(\omega_j)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x} \mid \omega_i)P(\omega_i) = p(\boldsymbol{x} \mid \omega_j)P(\omega_j)$$

## Les surfaces séparatrices








## Sommaire







## Comment approcher la règle de classification bayésienne ?

Le problème de l'apprentissage d'une règle de classification serait donc résolu si l'on connaissait les $P(\omega_i)$ et les $p(\boldsymbol{x} \mid \omega_i)$.

$P(\omega_i)$ : Les probabilités *a priori* des classes peuvent être soit supposées égales, soit estimées à partir des fréquences d'apparition dans l'ensemble d'apprentissage.

$p(\boldsymbol{x} \mid \omega_i)$ : Pour chaque classe, on se trouve devant un problème d'estimation de densité de probabilité à partir d'un nombre fini d'observations.





## L'estimation des probabilités *a priori*

- Soit, en l'absence d'information particulière, on les suppose égales et on prend l'estimateur : $\widehat{P(\omega_i)} = \frac{1}{C}$.
- Soit on suppose l'échantillon d'apprentissage représentatif et on les estime par les fréquences d'apparition de chaque classe dans cet ensemble : $\widehat{P(\omega_i)} = \frac{m_i}{m}$.
- Soit on utilise un estimateur intermédiaire (formule de Laplace)

$$\widehat{P(\omega_i)} = \frac{m_i + M/C}{m + M}$$

où $M$ est un nombre arbitraire. Cette formule est employée quand $m$ est petit, donc quand les estimations $m_i/m$ sont très imprécises. $M$ représente une augmentation virtuelle du nombre d'exemples, pour lesquels on suppose les classes équiprobables.





## Estimation d'une densité de probabilité

Deux techniques :

- les méthodes **paramétriques** : on suppose que les $p(\boldsymbol{x} \mid \omega_i)$ possèdent une certaine forme analytique.
  Si on les suppose **gaussiennes**, il suffit d'estimer la moyenne et la covariance de chaque distribution.
  La probabilité d'appartenance d'un point $\boldsymbol{x}$ à une classe se calcule alors directement à partir des coordonnées de $\boldsymbol{x}$.
- les méthodes **non-paramétriques** : on estime localement les densités $p(\boldsymbol{x} \mid \omega_i)$ au point $\boldsymbol{x}$ en observant l'ensemble d'apprentissage autour de ce point.
  Ces méthodes sont implémentées par la technique des **fenêtres de Parzen** (**noyaux**) ou l'algorithme des **K-plus proches voisins**.





## Sommaire

## Apprentissage au maximum de vraisemblance de classes supposées gaussiennes

Notons $E[x]$ l'**espérance mathématique** de la variable aléatoire $\boldsymbol{x}$.
La **moyenne** d'une densité de probabilité $p$ dans $\mathbb{R}^d$ est un vecteur de dimension $d$ défini par :

$$\mu = E[\mathbf{x}]$$

Sa **matrice de covariance** s'écrit :

$$Q = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T]$$

La $j^{eme}$ composante de $\mu$ vaut : $\mu(j) = E[\mathbf{x}_j] = \int_{\mathbb{R}} \mathbf{x}_j p(\mathbf{x}_j) d\mathbf{x}$
L'élément courant de sa **matrice de covariance** s'écrit :
$Q(j,k) = E[(\mathbf{x}_j - \mu(j))(\mathbf{x}_k - \mu(k))^T]$





## Apprentissage bayésien de classes gaussiennes

Une distribution de probabilité gaussienne est définie par son vecteur moyenne $\boldsymbol{\mu}$ et sa matrice de covariance $Q$.

Pour chaque classe :

$d = 1$ $Q$ se ramène à un scalaire $\sigma^2$ (la variance)

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\mathbf{x} - \boldsymbol{\mu}_i)^2}{\sigma_i^2}\right)$$

$d > 1$

$$p(\mathbf{x} \mid \omega_i) = \frac{\mid Q_i \mid^{-1/2}}{2\pi^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T Q_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right)$$





## Apprentissage bayésien de classes gaussiennes

Une estimation au **maximum de vraisemblance** maximise la probabilité d'observer les données d'apprentissage.
Pour la classe $\omega_i$, on possède $m_i$ points d'apprentissage, notés $\{\mathbf{x}_1, \ldots, \mathbf{x}_j \ldots, \mathbf{x}_{m_i}\}$.
Il est démontré que les estimations au maximum de vraisemblance de la moyenne $\boldsymbol{\mu}_i$ et de la matrice de covariance $Q_i$ se calculent par :

$$\widehat{\boldsymbol{\mu}_i} = \frac{\Sigma_{l=1}^{l=m_i} \mathbf{x}_l}{m_i}$$

$$\widehat{Q_i} = \frac{\Sigma_{l=1}^{l=m_i} (\mathbf{x}_l - \widehat{\boldsymbol{\mu}_i})(\mathbf{x}_l - \widehat{\boldsymbol{\mu}_i})^T}{m_i}$$





## Apprentissage bayésien de classes gaussiennes : surfaces séparatrices

Le lieu des points où les probabilités d'appartenir aux deux classes $\omega_i$ et $\omega_j$ sont égales est par définition :

$$\frac{\mid Q_i \mid^{-1/2}}{2\pi^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T Q_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right)$$
$$= \frac{\mid Q_j \mid^{-1/2}}{2\pi^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)^T Q_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j)\right)$$

Après simplification on obtient une *forme quadratique* :

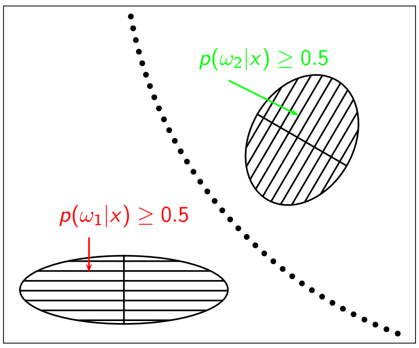
$$\mathbf{x}^T \Phi \mathbf{x} + \mathbf{x}^T \phi + \alpha = 0$$

La matrice $\Phi$, le vecteur $\phi$ et $\alpha$ ne dépendent que de $\boldsymbol{\mu}_i$, $\boldsymbol{\mu}_j$, $Q_i$, $Q_j$.





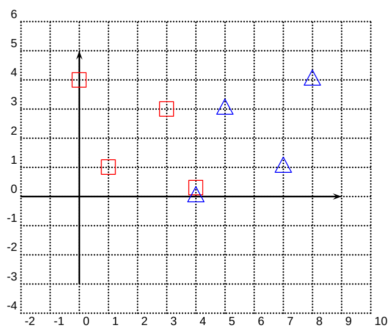
## Apprentissage bayésien de classes gaussiennes : Surface séparatrice de 2 classes dans $\mathbb{R}^2$








## Exercice 1 : un exemple à deux dimensions



## Exercice 1 : un exemple à deux dimensions

Ensemble d'apprentissage :

$$\omega_1 \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 4 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} \begin{pmatrix} 4 \\ 0 \end{pmatrix}$$

$$\omega_2 \quad \begin{pmatrix} 4 \\ 0 \end{pmatrix} \begin{pmatrix} 7 \\ 1 \end{pmatrix} \begin{pmatrix} 8 \\ 4 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

En supposant que les 2 classes sont gaussiennes, quelle est l'équation de la surface séparatrice ?





## Un cas plus compliqué : la modélisation par un mélange de gaussiennes

Mélange de $K$ gaussiennes :

$$p(\boldsymbol{x}|\omega_i) = \sum_{k=1}^{K} \alpha_k \frac{\mid Q_k \mid^{-1/2}}{(2\pi)^{d/2}} exp \left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T Q_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k) \right\} \text{ avec } \sum_{k=1}^{K} \alpha_k = 1$$

On apprend pour chaque classe $\omega_i$ tous les paramètres :

- la moyenne de chaque gaussienne : $\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K\}$
- la covariance de chaque gaussienne : $\{Q_1, ..., Q_K\}$
- les valeurs de mélange : $\{\alpha_1, ..., \alpha_K\}$

par l'algorithme d'optimisation *EM (Expectation-Maximization)*.





## Un cas simplifié : la classification bayésienne naïve

On suppose ici que chaque classe possède une matrice de covariance diagonale. Cette hypothèse revient à dire que les attributs sont statistiquement décorrélés.
Dans cette simplification, la probabilité d'observer $\boldsymbol{x}^T = (x_1, \ldots, x_d)$ pour un point de n'importe quelle classe $\omega_i$ est la probabilité d'observer l'attribut $x_1$ pour cette classe, multipliée par celle d'observer l'attribut $x_2$ pour cette classe, etc. Donc, par hypothèse :

$$\omega^* = \underset{i \in \{1,\ldots,C\}}{\text{ArgMax}} \quad P(\omega_i) \prod_{j=1}^{d} p(x_j \mid \omega_i)$$

Chaque valeur $p(x_j \mid \omega_i)$ s'estime par comptage dans un intervalle (histogramme monodimensionnel).





## Sommaire

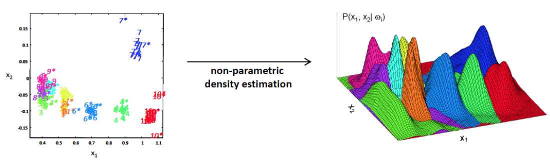






## Apprentissage bayésien non paramétrique

- Soit un point $\boldsymbol{x}$ dont on cherche la classe.
- On va estimer en $\boldsymbol{x}$ les densités de probabilités de chaque classe $\omega_i$, puis appliquer la règle de classification bayésienne.
- Pour chaque classe $\omega_i$, on a le même problème : on possède $m_i$ points d'apprentissage de $\mathbb{R}^d$ obtenus par tirages indépendants selon une densité $p(\boldsymbol{x} \mid \omega_i)$.
- Comment estimer $p(\boldsymbol{x} \mid \omega_i)$ au point $\boldsymbol{x}$ à partir de ces $m_i$ points de l'ensemble d'apprentissage ?








## Explication 1

remarque : pour simplifier les notations, l'indice $i$ est supprimé des transparents suivants. On notera $\omega$ une classe donnée $\omega_i$ qui contient $m$ points dans l'ensemble d'apprentissage (au lieu de $m_i$).

- Soit une région $\mathcal{R} \in \mathbb{R}^d$ de volume $V$.
- Soit un point $\boldsymbol{x}$ tiré aléatoirement selon une distribution de probabilité de densité $p$ (inconnue)
- Soit $P = P(\boldsymbol{x} \in \mathcal{R})$ la probabilité que ce point $\boldsymbol{x}$ tombe dans la région $\mathcal{R}$
- Soit $B$, la variable de Bernoulli définie par :

$$B = \begin{cases} 1 & \text{si } \boldsymbol{x} \in \mathcal{R} \quad \text{avec une proba P} \\ 0 & \text{sinon} \end{cases}$$

## Explication 2

D'une part,

- On tire indépendamment $m$ points selon $p$ : $\{x_1, x_2, \cdots, x_m\}$
- Alors $K = \sum_{i=1}^{m} B_i =$ "Nombre de fois où $\boldsymbol{x}$ tombe dans la région $\mathcal{R}$" suit un loi binomiale $\mathcal{B}(m, P)$ :

$$P(K = k) = \begin{pmatrix} m \\ k \end{pmatrix} P^k (1 - P)^{m-k}$$

- On tire de cette distribution que l'espérance de $K$ vaut $mP$ et donc $\mathbb{E}(\frac{K}{m}) = P$

$\Rightarrow$ $\frac{K}{m}$ est un estimateur de $P = P(\boldsymbol{x} \in \mathcal{R})$ : $\hat{P} = \frac{K}{m}$ (1)





## Explication 3

D'autre part,

- Soit une région $\mathcal{R} \in \mathbb{R}^d$ de volume $V$.
- Soit un point $\boldsymbol{x}$ tiré aléatoirement selon une distribution de probabilité de densité $p$ (inconnue)
- Alors $P = P(\boldsymbol{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\boldsymbol{u}) du$
- En prenant $\mathcal{R}$ assez petit pour que $p$ y soit constante, on a :
$$P = \int_{\mathcal{R}} p(\boldsymbol{u}) du \simeq p(\boldsymbol{x})\ V \quad (2)$$
- De (1) et (2) on déduit : $P = \frac{K}{m} = p(\boldsymbol{x})\ V$

$$\Rightarrow \quad p(\boldsymbol{x}) = \frac{K/m}{V}$$





## La technique

Soit $m$ le nombre de points de l'échantillon d'apprentissage de la classe $\omega$.

On définit autour de $\boldsymbol{x}$ une région $\mathcal{R}_m$ de volume $V_m$ et on compte le nombre $k_m$ de points de l'échantillon d'apprentissage de la classe $\omega$ qui sont inclus dans cette région.

$\rightarrow$ Estimateur de $p(\boldsymbol{x} \mid \omega)$ pour un échantillon de taille $m$ :

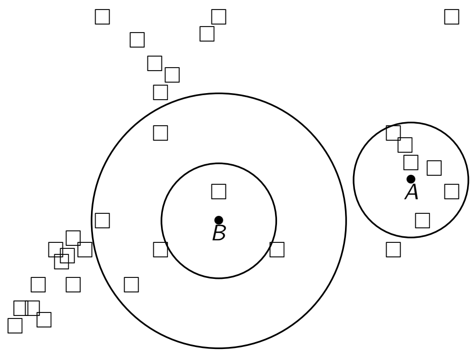
$$\widehat{p_m}(\boldsymbol{x} \mid \omega) = \frac{k_m/m}{V_m}$$

L'estimateur $\widehat{p_m}(\boldsymbol{x} \mid \omega)$ converge vers $p(\boldsymbol{x} \mid \omega)$ quand $m$ augmente si :

- $\lim V_m = 0$
- $\lim k_m = \infty$
- $\lim(k_m/m) = 0$








Les points $\square$ sont des tirages indépendants selon une certaine distribution dans le plan $\mathbb{R}^2$, dont la densité est plus forte au point $A$ qu'au point $B$. En effet, pour le même volume autour du point $A$ et du point $B$, $k_m$ vaut respectivement 6 et 1 . Pour avoir $k_m = 6$ autour du point $B$, il faut augmenter le volume.





## Apprentissage bayésien non paramétrique

La densité $p(\boldsymbol{x} \mid \omega)$ est estimée par la proportion d'exemples de la classe $\omega$ au voisinage de $\boldsymbol{x}$. Il y a deux solutions :

- **Fenêtres de Parzen** : subdivision de l'espace en boules de rayon $\rho$ (fixé) centré en $\boldsymbol{x}$. Soit $N(x)$, le nombre de points de la classe $\omega$ contenus dans la boule :
$$\widehat{p_m}(\boldsymbol{x} \mid \omega) \propto \frac{N(x)}{\rho}$$
- **K-plus proches voisins**, ou K-ppv : former des boules de rayon $\rho$ variable mais contenant exactement $K$ (fixé) points de l'ensemble d'apprentissage (les K-ppv du centre $\boldsymbol{x}$ de la boule) :
$$\widehat{p_m}(\boldsymbol{x} \mid \omega) \propto \frac{K}{\rho_K(x)}$$

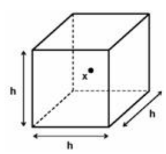




## Fenêtres de Parzen : le cas élémentaire

Considérons une région $\mathcal{R}$ qui est un hypercube de côté $h$ centré sur le point $\boldsymbol{x} \in \mathbb{R}^d$ :

- $V_m = h^d$
- Soit la fonction $\phi$ représentant un cube unité centré sur l'origine :




$$\phi(\boldsymbol{u}) = \begin{cases} 1 & \text{si } |u_j| < 1/2 \quad j = 1, \cdots, d \\ 0 & \text{sinon} \end{cases}$$

Parmi les $m$ points de la classe $\omega$, le nombre total de points $\boldsymbol{x}_i$ tombant à l'intérieur de cet hypercube est :

$$k_m = \sum_{i=1}^{m} \phi\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)$$

## Fenêtres de Parzen : le cas élémentaire

Alors :

$$\widehat{p_m}(\boldsymbol{x} \mid \omega) = \frac{1}{m}\frac{1}{V_m}\sum_{i=1}^{m}\phi\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right)$$

La fonction $\phi$ est un exemple de *fonction noyau* (*noyau uniforme*).

Remarque : On doit avoir $\phi(\boldsymbol{u}) \geq 0$ et $\int_{\mathbb{R}^d} \phi(\boldsymbol{u})\, d\boldsymbol{u} = 1$ pour garantir que l'estimation $\widehat{p_m}(\boldsymbol{x} \mid \omega)$ est une densité de probabilité, ie :

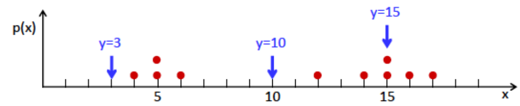
- $\widehat{p_m}(\boldsymbol{x} \mid \omega) \geq 0$
- $\int_{\mathbb{R}^d} \widehat{p_m}(\boldsymbol{x} \mid \omega)\, d\boldsymbol{x} = 1$





## Exercice

- Etant donné l'ensemble $X = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$, estimez la densité $p(y)$ pour $y = 3, 10, 15$ en utilisant les fenêtres de Parzen avec $h = 4$.
- Représentation graphique de l'ensemble $X$








## Exercice





## Fenêtres de Parzen : les noyaux

La technique se décrit plus généralement par le calcul :

$$\widehat{p_m}(\boldsymbol{x} \mid \omega) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{V_m}\kappa(\boldsymbol{x}, \boldsymbol{x}_i)$$

- $\kappa(\boldsymbol{x}, \boldsymbol{x}_i)$ est centrée en $\boldsymbol{x}_i$ et décroît quand $\boldsymbol{x}$ s'éloigne de $\boldsymbol{x}_i$.
- Elle a une intégrale finie : le volume $V_m$
- Elle est symétrique et positive

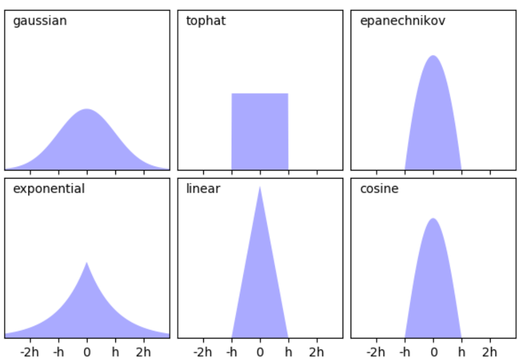
Par exemple $\kappa$ peut être un rectangle de largeur $h$ variable, ou une gaussienne de variance $h$ variable : $\kappa(\boldsymbol{x}, \boldsymbol{x}_i) = \frac{1}{(h\sqrt{2\pi})^d}\exp(-\frac{1}{2}(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h})^2)$





## Parzen windows : estimating with kernels

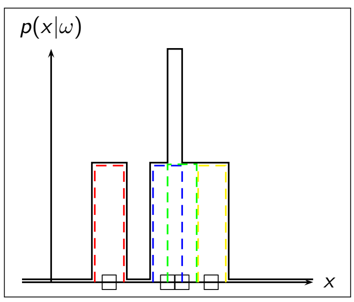
Exemples de noyaux proposés par scikit-learn :





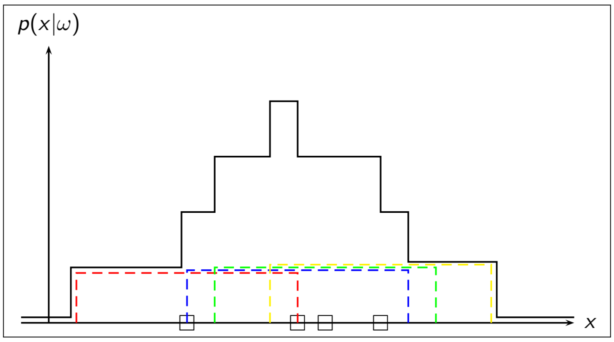



## Fenêtres de Parzen (noyaux rectangles)




Estimation de densité par la méthode des fenêtres de Parzen. Il y a quatre points d'apprentissage, dans un espace à une dimension. La densité (en trait plein) est calculée comme la somme des fenêtres centrées sur chaque point. Ici, cette fenêtre est étroite ($h$ est petit) : la densité résultante est peu lisse.

## Fenêtres de Parzen (noyaux rectangles)

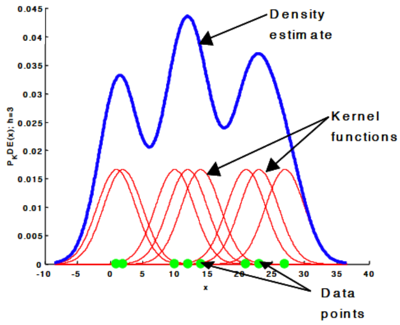



La même estimation pour $h$ grand : la densité est estimée de manière plus lisse.





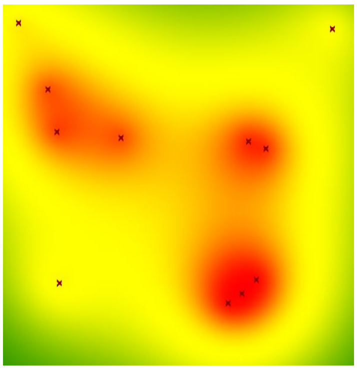
## Fenêtres de Parzen (noyaux gaussiens)








## Estimation de densité par noyau (2D)







## Estimation de densité par noyau

- L'estimation de densité par noyau est une somme de "bosses"
- La fonction noyau détermine la forme des "bosses"
- Le paramètre de lissage $h$ (**smoothing parameter** ou **bandwidth**) détermine leur largeur

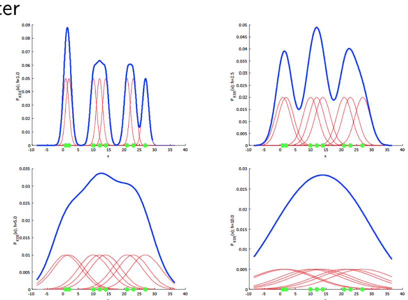




## Choix du paramètre de lissage $h$

**Le choix de $h$ est crucial en estimation de densité par noyau.**

- Si $h$ est trop grand, la densité estimée est trop lissée et masque la structure des données
- Si $h$ est trop petit, la densité estimée est hérissée de pointes et difficile à interpréter





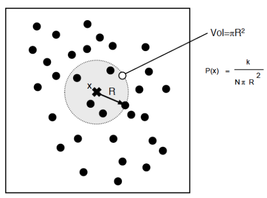


## Estimation de densité par K-NN

- Dans l'approche des $K$-plus proches voisins, le volume autour du point d'estimation $x$ grandit jusqu'à contenir $K$ points de l'ensemble de données.
- L'estimation de la densité devient alors :

$$p(x|\omega) = \frac{K}{mV_K^d(x)}$$

où $V_K^d(x)$ est le volume de la sphère en dimension $d$ dont le rayon est la distance entre le point d'estimation $x$ et son $K$-ième plus proche voisin dans l'ensemble de données.



**Algorithme des K-plus proches voisins**

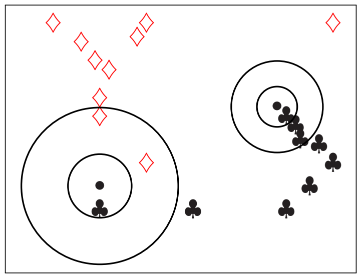
---

**Début**
**pour** chaque exemple $(\boldsymbol{y}, \omega)$ de l'ensemble d'apprentissage **faire**
calculer la distance $D(\boldsymbol{y}, \boldsymbol{x})$ entre $\boldsymbol{y}$ et $\boldsymbol{x}$
**fin pour**
Dans les $K$ points les plus proches de $\boldsymbol{x}$
compter le nombre d'occurences de chaque classe
Attribuer à $\boldsymbol{x}$ la classe qui apparaît le plus souvent
**Fin**

---



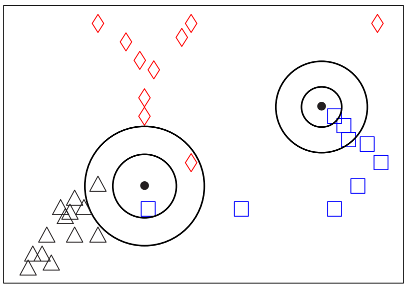




*Décision par 1-PPV et 3-PPV pour deux classes.*







*Décision par 1-PPV et 3-PPV pour trois classes*





## Les K-ppv : validité (1)

**Revendication**

La règle des K-ppv approxime la décision bayésienne, car elle fait implicitement une estimation comparative de toutes les densités de probabilités des classes apparaissant dans le voisinage de $\boldsymbol{x}$ et choisit la plus probable.

Supposons que les $m$ points de l'ensemble d'apprentissage comportent $m_i$ points de la classe $\omega_i$ et que sur les $K$ plus proches voisins de $\boldsymbol{x}$, il y a $K_{m_i}$ points de cette classe.

On a :

$$\widehat{p_m}(\boldsymbol{x} \mid \omega_i) = \frac{K_{m_i}/m_i}{V_m}$$





## Les K-ppv : validité (2)

Comme $m_i/m$ est un estimateur de $P(\omega_i)$, la probabilité *a priori* de la classe de rang $i$, on peut écrire : $m_i/m = \widehat{P_m}(\omega_i)$.

Donc :

$$\widehat{p_m}(\boldsymbol{x} \mid \omega_i).\widehat{P_m}(\omega_i) = \frac{K_{m_i}}{m_i} \cdot \frac{1}{V_m} \cdot \frac{m_i}{m}$$

On en déduit :

$$K_{m_i} = \widehat{p_m}(\boldsymbol{x} \mid \omega_i).\widehat{P_m}(\omega_i).m.V_m$$

Par conséquent, la classe qui maximise $K_{m_i}$ maximise aussi :

$$\widehat{p_m}(\boldsymbol{x} \mid \omega_i).\widehat{P_m}(\omega_i)$$

et donc, par la règle de Bayes, maximise aussi :

$$\widehat{P_m}(\omega_i \mid \boldsymbol{x}).p(\boldsymbol{x})$$





## Les K-ppv : validité (3)

Cette classe est donc conforme au choix de la règle de classification bayésienne, puisqu'elle maximise :

$$\widehat{P_m}(\omega_i \mid \boldsymbol{x})$$

Pour finir il faut démontrer que cette méthode répond aux conditions imposées plus haut. Pour $K$ fixé et $m \to \infty$, on a pour chaque classe :

- $V_m \to 0$
- $K_m/m \to 0$

La probabilité d'erreur $E_{K-ppv}$ de la règle des $K-ppv$ converge vers l'erreur bayésienne quand $m$ augmente.

## Les K-ppv en tant qu'algorithme d'apprentissage

- Les K-ppv sont considérés comme un algorithme d'apprentissage paresseux (**lazy** learning algorithm)
  - Les données ne sont pas traitées avant de recevoir un exemple non labellisé à classer
  - La réponse consiste à combiner les données d'apprentissage stockées
- Cette stratégie s'oppose aux autres algorithmes d'apprentissage (**eager** learning algorithm) qui
  - compilent les données en une description compressée ou un modèle
  - écartent les données d'apprentissage une fois le modèle construit





## Les K-ppv : considérations pratiques

Choix de $K$ ?

Diverses considérations théoriques et expérimentales mènent à l'heuristique suivante :

$$K \simeq \sqrt{m/C}$$

$m/C$ est le nombre moyen de points d'apprentissage par classe. On remarquera que $d$, la dimension de l'espace de représentation, n'apparaît pas dans cette formule.





## Les K-ppv : considérations pratiques

Quelle décision prendre en cas d'égalité ?

- On peut augmenter $K$ de 1 pour trancher le dilemme, mais l'ambiguïté peut subsister.
- Une bonne solution consiste à tirer au hasard la classe à attribuer au point ambigu.
- On peut aussi pondérer les "votes" de chaque exemple par sa distance au point $y$.





## Les surfaces séparatrices de la règle de décision K-ppv

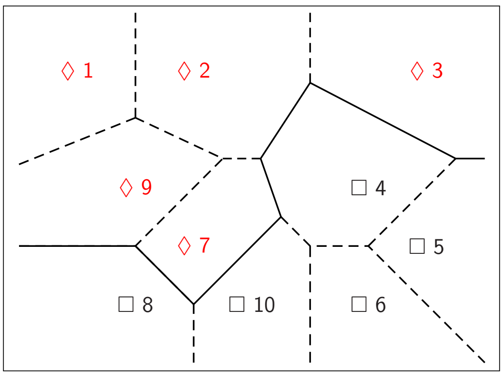
On appelle **zone de Voronoï** d'un exemple le lieu des points de $\mathbb{R}^d$ qui sont plus proches de cet exemple que de tout autre exemple.

C'est l'intersection de $m-1$ demi-espaces, définis par les hyperplans médiateurs entre cet exemple et tous les autres.

Pour $k=1$, la surface séparatrice entre deux classes est la surface séparatrice entre les deux volumes obtenus en faisant l'union des surfaces de Voronoï des exemples de chaque classe.








*Un ensemble de points et leurs zones de Voronoï ($k=1$).*





## Exercice 2

On considère de nouveau le jeu de données "Tennis" rappelé dans la table ci-après.
On souhaite prédire la classe de la donnée $\boldsymbol{x}$=(Ensoleillé, Fraîche, Elevée, Fort), en utilisant une méthode de classification basée sur la règle de Bayes.

1. Exprimer la règle de décision utilisée.
2. Calculer les probabilités a priori de chaque classe.
3. En appliquant l'hypothèse de Bayes naïve, estimer les vraisemblances de $\boldsymbol{x}$ pour chaque classe. Quelle est la classe de $\boldsymbol{x}$ ?

## Exercice 2

| Jour | Ciel | Température | Humidité | Vent | Jouer au tennis ? |
|---|---|---|---|---|---|
| 1 | Ensoleillé | Chaude | Elevée | Faible | Non |
| 2 | Ensoleillé | Chaude | Elevée | Fort | Non |
| 3 | Couvert | Chaude | Elevée | Faible | Oui |
| 4 | Pluie | Tiède | Elevée | Faible | Oui |
| 5 | Pluie | Fraîche | Normale | Faible | Oui |
| 6 | Pluie | Fraîche | Normale | Fort | Non |
| 7 | Couvert | Fraîche | Normale | Fort | Oui |
| 8 | Ensoleillé | Tiède | Elevée | Faible | Non |
| 9 | Ensoleillé | Fraîche | Normale | Faible | Oui |
| 10 | Pluie | Tiède | Normale | Faible | Oui |
| 11 | Ensoleillé | Tiède | Normale | Fort | Oui |
| 12 | Couvert | Tiède | Elevée | Fort | Oui |
| 13 | Couvert | Chaud | Normale | Faible | Oui |
| 14 | Pluie | Tiède | Elevée | Fort | Non |





## Exercice 3

On considère à présent le jeu de données pour lequel les valeurs de certains attributs sont numériques (table ci-après). En l'absence d'information supplémentaire, on suppose que la distribution des valeurs des attributs numériques est normale.
On souhaite prédire la classe de la donnée $\boldsymbol{x}$=(Ensoleillé, 18, 90, Fort),

1. Estimer la densité de probabilité de chacun des attributs numérique.
2. Sous l'hypothèse de Bayes naïve, estimer les vraisemblances de $\boldsymbol{x}$ pour chaque classe. Quelle est la classe de $\boldsymbol{x}$ ?





## Exercice 3

| | Ciel | Temp. | Humidité | Vent | Jouer au tennis ? |
|---|---|---|---|---|---|
| 1 | Ensoleillé | 27.5 | 85 | Faible | Non |
| 2 | Ensoleillé | 25.0 | 90 | Fort | Non |
| 3 | Couvert | 26.5 | 86 | Faible | Oui |
| 4 | Pluie | 20.0 | 96 | Faible | Oui |
| 5 | Pluie | 19.0 | 80 | Faible | Oui |
| 6 | Pluie | 17.5 | 70 | Fort | Non |
| 7 | Couvert | 17.0 | 65 | Fort | Oui |
| 8 | Ensoleillé | 21.0 | 95 | Faible | Non |
| 9 | Ensoleillé | 19.5 | 70 | Faible | Oui |
| 10 | Pluie | 22.5 | 80 | Faible | Oui |
| 11 | Ensoleillé | 22.5 | 70 | Fort | Oui |
| 12 | Couvert | 21.0 | 90 | Fort | Oui |
| 13 | Couvert | 25.5 | 75 | Faible | Oui |
| 14 | Pluie | 20.5 | 91 | Fort | Non |