

• Objectif

- estimer la densité $p(\mathbf{x})$
- étant donné $D_n = (X_1, X_2, \dots, X_n)$ tirés iid de $p(\mathbf{x})$

• Idée

- la probabilité pour que $\mathbf{x} \in R$: $P = P(R) = \int_R p(\mathbf{x}') d\mathbf{x}'$
- la probabilité pour que k points de D_n tombe dans R :

$$P_k = P_k(R) = \binom{n}{k} P^k (1-P)^{n-k}$$

$$\mathbb{E} \left[\frac{k}{n} \right] = P$$

- si $p(\mathbf{x})$ ne change pas beaucoup sur R :

$$\int_R p(\mathbf{x}') d\mathbf{x}' = p(\mathbf{x}) V(R)$$

- $p(\mathbf{x}) \simeq \frac{k/n}{V(R)}$

• Problèmes

- **fixer R , laisser n croître**: $\frac{k}{n} \rightarrow P(R)$ mais $\frac{P(R)}{V(R)} \not\rightarrow p(\mathbf{x})$
- **fixer n , laisser $V(R)$ décroître**: $\frac{P(R)}{V(R)} \rightarrow p(\mathbf{x})$ mais $\frac{k}{n} \not\rightarrow P(R)$

• Solution théorique

- séquence des régions R_1, R_2, \dots :

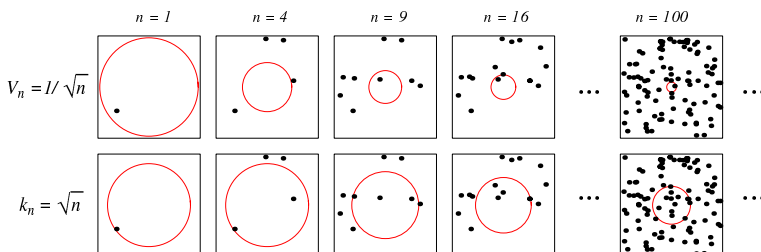
$$p_n(\mathbf{x}) = \frac{k_n/n}{V(R_n)}$$

• Conditions pour que $p_n(\mathbf{x}) \rightarrow p(\mathbf{x})$

- $\lim_{n \rightarrow \infty} V(R_n) = 0$
- $\lim_{n \rightarrow \infty} k_n = \infty$
- $\lim_{n \rightarrow \infty} k_n/n = 0$

• Deux stratégies

- **fenêtres de Parzen**: fixer $V(R_n)$ (e.g. $= 1/\sqrt{n}$)
- **k_n -plus-proches-voisins**: fixer k_n (e.g. $= \sqrt{n}$)



• Fenêtres de Parzen

- R_n est un hypercube: $V(R_n) = V_n = h_n^d$

- **fonction de fenêtre**:

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

- nombre de pointes qui tombent dans l'hypercube centré à \mathbf{x} :

$$k_n = \sum_{i=1}^n \phi \left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n} \right)$$

- l'estimation:

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \phi \left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n} \right)$$

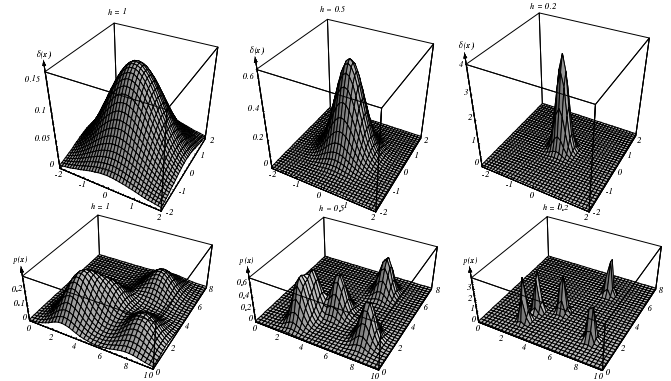
- Conditions pour que $p_n(\mathbf{x})$ soit une densité

- $\phi(\mathbf{u}) \geq 0$
- $\int \phi(\mathbf{u}) d\mathbf{u} = 1$ (avec $V_n = h_n^d$)

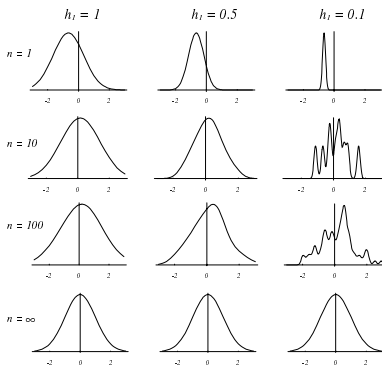
- L'effet de la largeur de fenêtre h_n

- $\delta_n(\mathbf{x}) = \frac{1}{V_n} \phi\left(\frac{\mathbf{x}}{h_n}\right)$
- $p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \delta_n(\mathbf{x} - \mathbf{x}_i)$

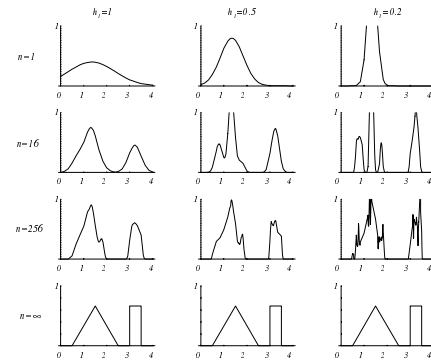
- L'effet de la largeur de fenêtre h_n



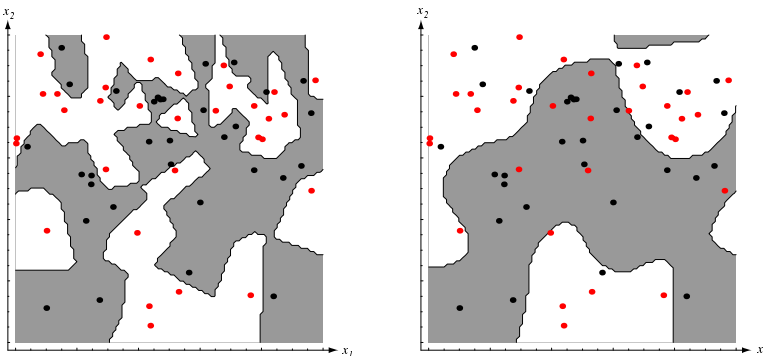
- Exemple: $p(x) \sim N(0, 1)$, $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$



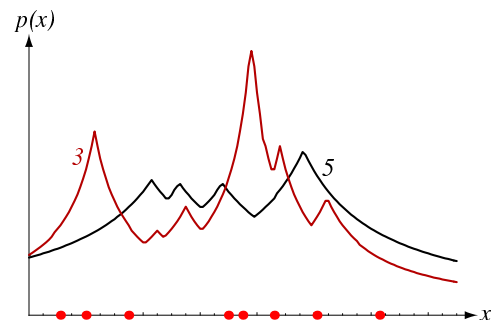
- Exemple: $p(x) \sim \text{triangle} + \text{uniform}$, $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$



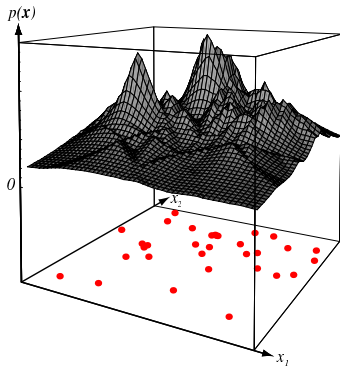
- Exemple de classification



- Estimation k_n -plus-proches-voisins



• Estimation k_n -plus-proches-voisins



• Règle de classification

- probabilité jointe:

$$p_n(\mathbf{x}, C_i) = \frac{k_i/n}{V}$$

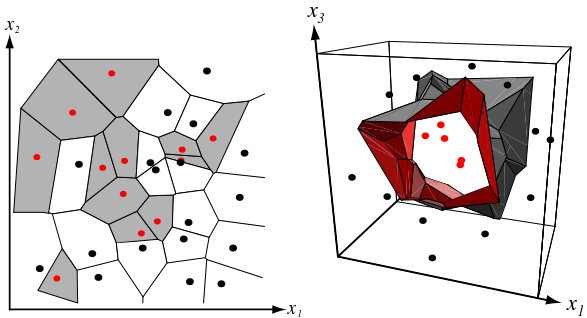
- probabilité a-posteriori:

$$p_n(C_i, \mathbf{x}) = \frac{p_n(\mathbf{x}, C_i)}{p_n(\mathbf{x})} = \frac{\frac{k_i/n}{V}}{\frac{k/n}{V}} = \frac{k_i}{k}$$

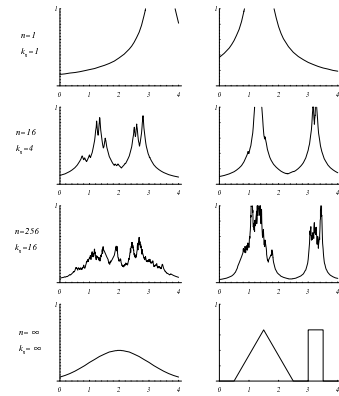
- décision par vote de majorité:

$$f(\mathbf{x}) = C_i \text{ si } k_i \geq k_j, j = 1, \dots, N$$

• Partition de Voronoi



• Estimation k_n -plus-proches-voisins



• Règle de plus-proche-voisin

$$\mathbf{x}'(\mathbf{x}) = \mathbf{x}' = \arg \min_{\mathbf{x}_i \in D_n} d(\mathbf{x}, \mathbf{x}_i)$$

$$f(\mathbf{x}) = y'$$

• L'erreur de plus-proche-voisin

- le risque:

$$R(f) = \mathbb{E}_X[L(f, X)] = R(f) = P(f(X) \neq Y) = P(e)$$

- l'erreur espéré:

$$P_n(e) = \mathbb{E}_{D_n}[R(f_{D_n})]$$

- l'erreur espéré asymptotique:

$$P = \lim_{n \rightarrow \infty} P_n(e)$$

• L'erreur de plus-proche-voisin

- décision de Bayes:

$$C_m = C_m(X) = \arg \max_{C_i} P(Y = C_i | X) = \arg \max_{C_i} P(C_i | X)$$

- consistance statistique:

$$\begin{aligned} P = P^* &= \int P^*(e|X)p(X)dX \\ &= \int P(f^*(X) \neq Y|X)p(X)dX \\ &= 1 - \int P(C_m|X)p(X)dX \end{aligned}$$

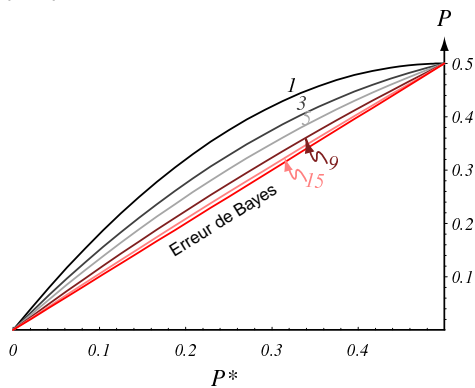
• L'erreur de plus-proche-voisin

$$P = \lim_{n \rightarrow \infty} P_n(e) = \int \left[1 - \sum_{i=1}^N P^2(C_i|X) \right] p(X)dX$$

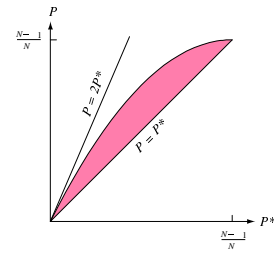
$$1 - \sum_{i=1}^N P^2(C_i|X) \leq 2P^*(e|X) - \frac{N}{N-1} P^{*2}(e|X)$$

$$\int P^{*2}(e|X)p(X)dX \geq P^{*2}$$

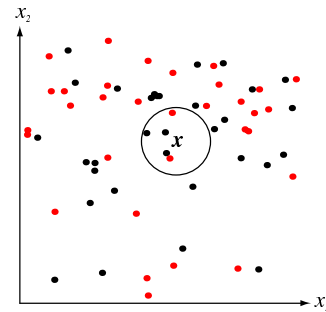
• L'erreur de k-plus-proche-voisin



• L'erreur de plus-proche-voisin $P^* \leq P \leq P^* \left(2 - \frac{N}{N-1} P^* \right)$



• Règle de k-plus-proche-voisin



• Complexité computationnelle de k-plus-proche-voisin

- méthode naïve: $T(n, k, d) = O(nkd) = O(n^2d)$

- méthode de distances partielles:

$$d_r(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^r (a_i - b_i)^2 \right)^{1/2}, r \leq d$$

- méthode d'arbre de recherche

• Complexité computationnelle de k -plus-proche-voisin

- méthode de suppression/émondage (editing/pruning/condensing)

```

ÉMONDAGEDEPLUSPROCHEVOISIN( $D_n$ )
1  construire le diagramme de Voronoi complet de  $D_n$ 
2  pour  $j \leftarrow 1$  à  $n$  faire
3    pour tout les voisins de Voronoi  $x'$  de  $x_j$  faire
4      si  $y_i \neq y'$  alors
5        marquer  $x_i$ 
6  pour  $j \leftarrow 0$  à  $n$  faire
7    si  $x_j$  n'est pas marqué alors
8      supprimer  $x_i$ 
    
```

• $T(n, d) = O(d^3 n^{d/2} \ln n)$

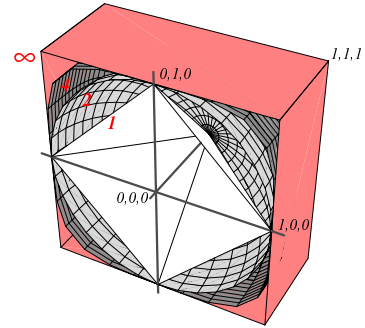
• Propriétés d'une métrique

- positivité: $d(\mathbf{a}, \mathbf{b}) \geq 0$
- réflexivité: $d(\mathbf{a}, \mathbf{a}) = 0$
- symétrie: $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$
- inégalité de triangle: $d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c}) \geq d(\mathbf{a}, \mathbf{c})$

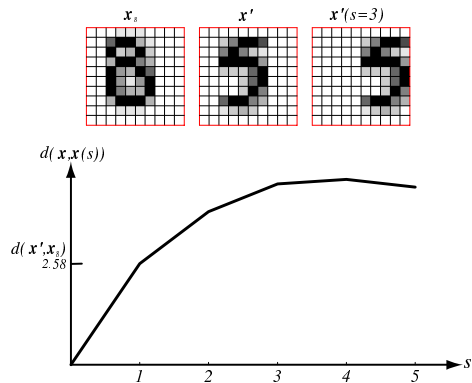
• Exemples des métriques

euclidienne	L_2	$d(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^d (a_i - b_i)^2 \right)^{1/2}$
Manhattan	L_1	$d(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^d a_i - b_i $
	L_∞	$d(\mathbf{a}, \mathbf{b}) = \max a_i - b_i $
Minkowski	L_p	$d(\mathbf{a}, \mathbf{b}) = \left(\sum_{i=1}^d a_i - b_i ^p \right)^{1/p}$
Tanimoto	L_{Tanimoto}	$d(S_1, S_2) = \frac{ S_1 + S_2 - 2 S_1 \cap S_2 }{ S_1 + S_2 - S_1 \cap S_2 }$

• La métrique de Minkowski



• Les limitations de la métrique euclidienne

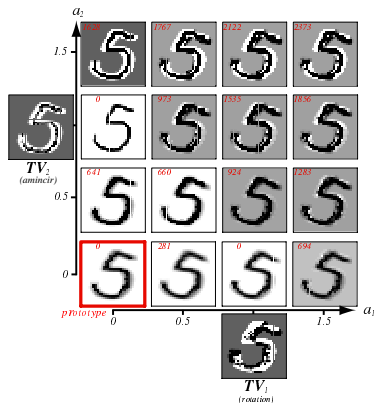


• La distance tangente

- capturer l'invariance de certaines transformations:

$$TV_i = \mathcal{F}_i(x'; a_i) - x'$$

- La distance **tangente**



- La distance **tangente**: $d_{tan}(\mathbf{x}', \mathbf{x}) = \min_{\mathbf{a}} \|\mathbf{x}' + \mathbf{T}\mathbf{a} - \mathbf{x}\|$

