

Les k plus proches voisins

Séance 3 – Une méthode sans modèle

Christophe Ambroise 

UEVE, UMR CNRS 8071

5 octobre 2026

1

Du global au local

2

Probabilité par fréquence

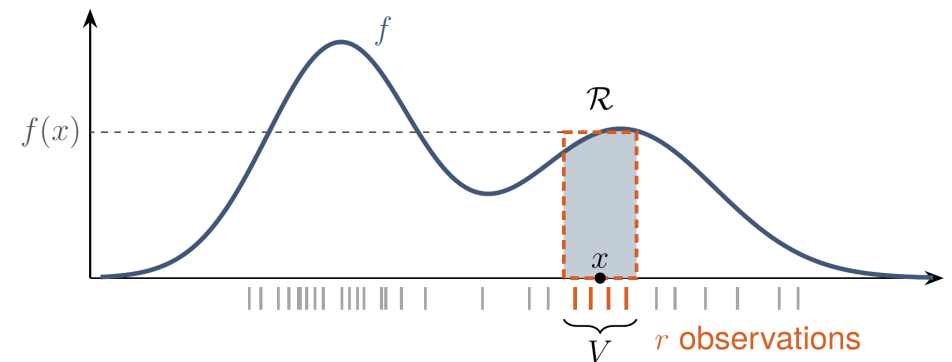
Comment faire sans modèle pour f_k ?

Réponse naïve : observations **proches de x** , puis comptage.

$$\hat{P}(Y = q \mid X = x) = \frac{\text{nombre de voisins de classe } q}{\text{nombre de voisins}}$$

3

Probabilité dans une région



Région $\mathcal{R}$ de volume V autour de x , r observations sur n :
 $f(x) V \approx P(X \in \mathcal{R}) \approx r/n$.

4

L'estimateur local

La densité en x s'estime par la proportion d'observations de la région, rapportée à son volume.

$$\hat{f}(x) = \frac{r}{nV}$$

- r grand : estimation stable, mais biaisée;
- V petit : estimation locale, mais variance élevée.

Deux exigences contradictoires.

5

Volume fixé, ou effectif fixé

Volume fixé

Fenêtre de largeur h autour de x .

Effectif variable.

Zone rare : peu de points.

Effectif fixé

k plus proches, toujours.

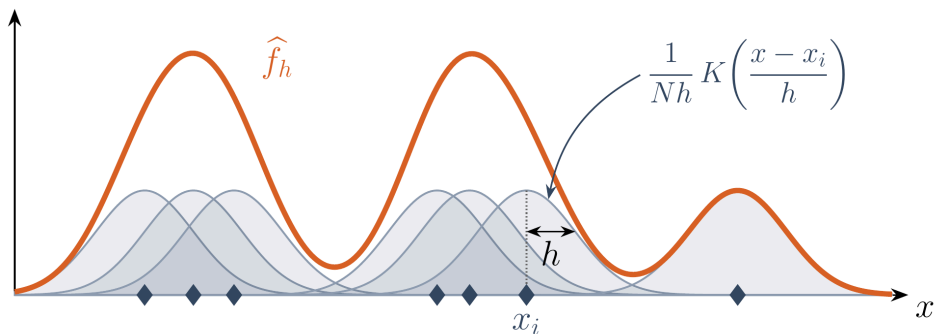
Volume variable.

Zone rare : grand volume.

k -PPV : effectif fixé, **voisinage adaptatif**.

6

Volume fixé : fenêtres de Parzen



Un noyau par observation ; densité = somme des bosses. Largeur h : taille du voisinage.

7

Largeur de fenêtre et lissage

Même échantillon, trois niveaux de lissage

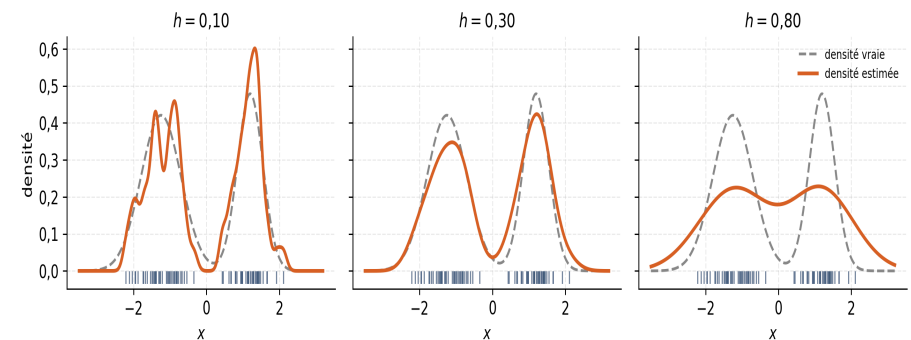


Figure 1

8

La règle

Densité et loi *a posteriori*

$V_k(x)$: volume de la plus petite boule centrée en x contenant k observations.

$$\hat{f}(x) = \frac{k}{n V_k(x)}.$$

Avec k_q voisins de classe q parmi les k , la probabilité *a posteriori* estimée est la fréquence de la classe q parmi les voisins.

$$\hat{P}(Y = q \mid X = x) = \frac{\hat{\pi}_q \hat{f}_q(x)}{\hat{f}(x)} = \frac{k_q}{k}.$$

10

Le vote local

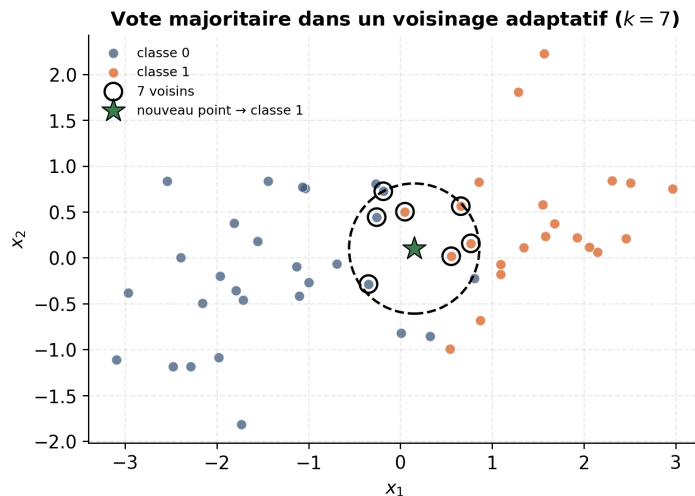


Figure 2

11

L'algorithme

Nouvelle observation x :

1. distances à toutes les observations d'apprentissage ;
2. sélection des k plus petites ;
3. comptage des classes ;
4. classe majoritaire.

Aucun paramètre ajusté. Apprentissage = **stockage** des données.

12

Les deux extrémités de k

$k = 1$

Classe du plus proche voisin.

Frontière irrégulière, sensible au bruit.

$k = n$

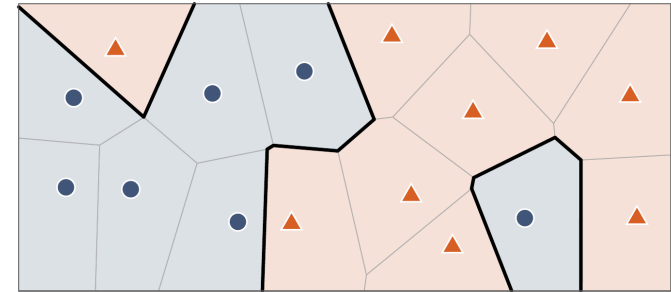
Vote de tous les points, partout.

Classe majoritaire : classifieur constant de la séance 1.

k : toute l'échelle, du modèle le plus souple au plus rigide.

13

Pavage de Voronoï



$k = 1$: une cellule par observation, ensemble des points dont elle est la plus proche.

Frontière (trait épais) : bords entre cellules de classes différentes. Aspect en escalier.

14

Erreur d'apprentissage du 1-PPV

Erreur **presque** toujours nulle : chaque point, son propre voisin.

Exception : deux fleurs d'espèces différentes, mêmes mesures.

15

Trois valeurs de k sur Iris

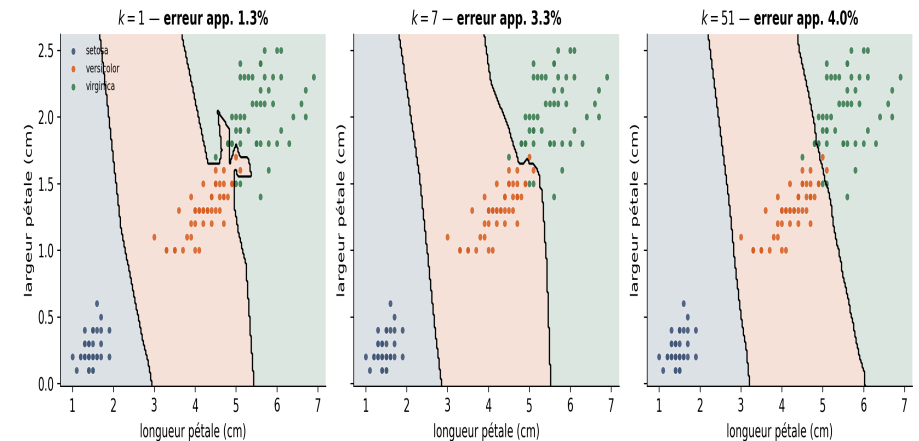


Figure 3

$k = 1$ au-dessus de 0 % : **48 lignes en double** sur les deux mesures de pétale, dont certaines d'espèces différentes.

16

Des probabilités à gros grain

Vote uniforme :

$$\hat{P}(Y = q \mid X = x) \in \left\{0, \frac{1}{k}, \frac{2}{k}, \dots, 1\right\}.$$

$k = 7$: sept valeurs possibles. Seuil et courbe ROC de la séance 1 : résolution grossière.

Vote **pondéré** par $1/d(x, x_i)$: échelle plus fine, calibration toujours imparfaite.

17

Les égalités

Deux classes, k pair : égalité possible.

18

En théorie

19

Une question légitime

Quelle performance sans modèle ?

1. Plus proche voisin avec **une infinité** de données : quel risque ?
2. Conditions d'accès au **risque de Bayes** pour les k -PPV ?

20

1-PPV et Cover et Hart (1967)

Soit $\eta(x) = P(Y = 1 \mid X = x)$,

sachant $X = x$ et $X_{(1)} = x'$, les labels Y et Y' sont indépendants (couples i.i.d.), avec $Y \sim \mathcal{B}(\eta(x))$ et $Y' \sim \mathcal{B}(\eta(x'))$:

$$P(Y \neq Y' \mid X = x, X_{(1)} = x') = \eta(x)(1 - \eta(x')) + (1 - \eta(x))\eta(x').$$

Deux erreurs possibles, et si η est continue, $\eta(X_{(1)}) \rightarrow \eta(x)$ et le risque conditionnel tend vers $2\eta(x)(1 - \eta(x))$

Avec $n \rightarrow \infty$, le plus proche voisin de x tend vers x et sa classe suit un tirage selon $\eta(x)$

$$R_{1\text{-PPV}}^\infty = \mathbb{E}[2\eta(X)(1 - \eta(X))].$$

21

L'encadrement de Cover et Hart

Soit $r(x) = \min(\eta(x), 1 - \eta(x))$, le risque de Bayes est $R^* = \mathbb{E}[r(X)]$.

$$2\eta(1 - \eta) = 2r(1 - r) \in [r, 2r].$$

Comme $u \mapsto 2u(1 - u)$ est concave, l'inégalité de Jensen donne la borne de Cover–Hart :

$$R^* \leq R_{1\text{-PPV}}^\infty \leq 2R^*(1 - R^*) \leq 2R^*.$$

Cas d'égalité $R^\infty = R^*$:

- $\eta \in \{0, 1\}$ p.s. : il n'y a pas de bruit, et les deux risques sont nuls ;
- $\eta = 1/2$ p.s. : les deux risques valent $1/2$. . .

Un seul voisin : jamais plus du **double** du meilleur risque possible.

22

Stone (1977) : atteindre R^*

1-PPV : pas de convergence vers R^* , bruit de l'étiquette toujours présent.

Les k -PPV convergent vers le risque de Bayes si le nombre de voisins croît sans limite, mais moins vite que l'échantillon.

$$k_n \rightarrow \infty \quad \text{et} \quad \frac{k_n}{n} \rightarrow 0 \quad \implies \quad R_{k_n\text{-PPV}} \rightarrow R^*.$$

$k_n = \lceil \sqrt{n} \rceil$: convergence. k fixe, $k = n$: échec.

23

Les trois régimes

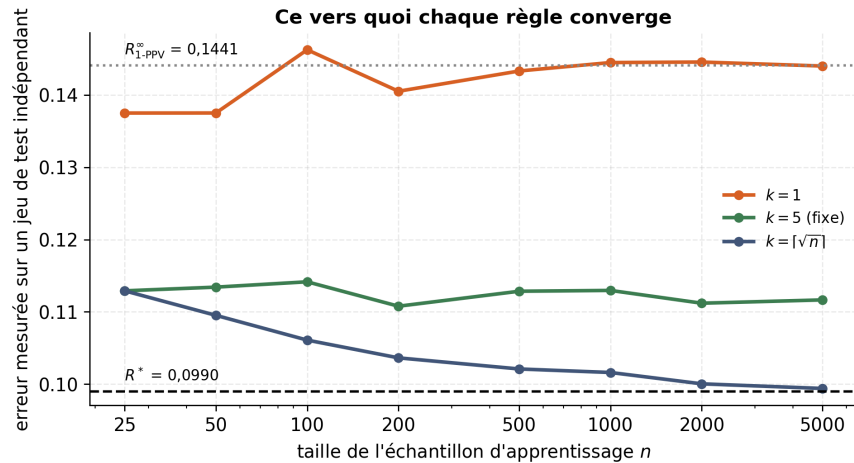


Figure 4

24

Les trois régimes

- $k = 1$: palier à **0,1441**, valeur de Cover et Hart.
- $k = 5$ **fixe** : plafond vers **0,112**, biais local persistant.
- $k = \lceil\sqrt{n}\rceil$: descente jusqu'à **0,0990**, risque de Bayes.

k -PPV **universellement constants** : optimum atteint sans hypothèse sur la loi. Aucune garantie de ce type en paramétrique.

25

Biais et variance en k

k petit

Voisinage étroit : biais faible.

Peu de votants : variance forte.

Courbe en U de la séance 1, axe de complexité $1/k$.

k grand

Voisinage large, zones hétérogènes moyennées : biais fort.

Beaucoup de votants : variance faible.

26

Choix de distance

27

Hypothèse implicite

$$d(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$

Contribution de chaque variable : **carré** de son écart. Étalement × 10 → poids × 100.

Choix d'une unité = choix implicite d'un modèle.

28

Une fleur, deux unités

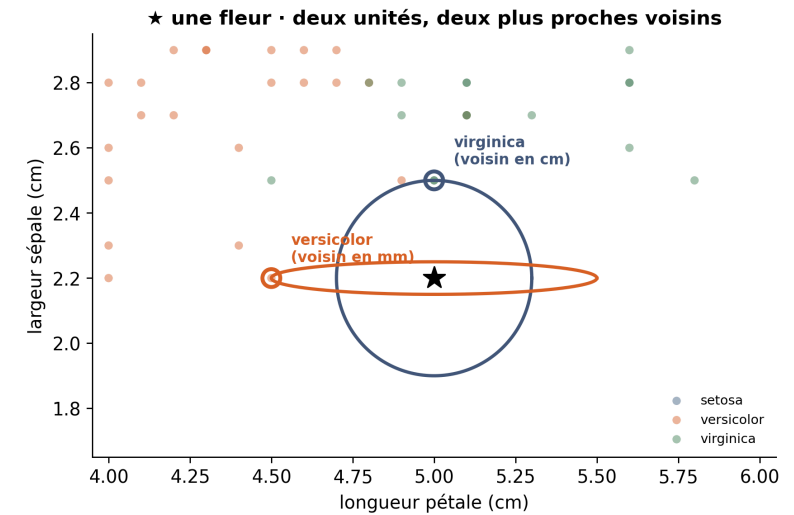


Figure 5

Même fleur, mêmes données. Largeur de sépale en millimètres : autre plus proche voisin, autre espèce.

Le poids réel de chaque variable

Le poids d'une variable dans la distance suit son étalement

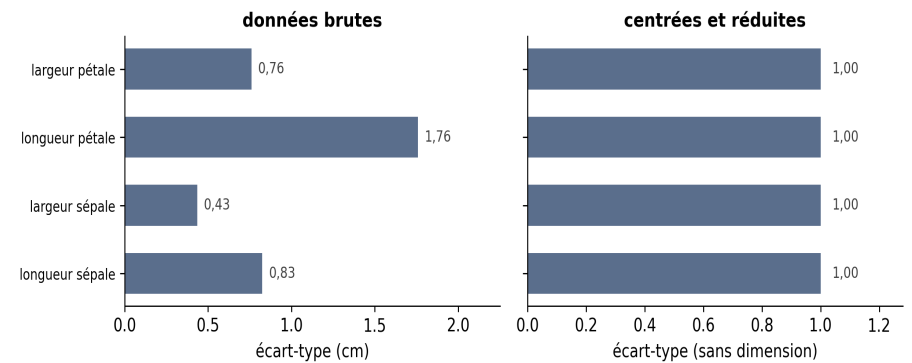


Figure 6

29

30

Standardiser

$$z_j = \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}$$

Poids égal pour chaque variable. Choix par défaut raisonnable, pas vérité : variable connue pour compter davantage → poids supérieur.

31

Où estimer $\hat{\mu}_j$ et $\hat{\sigma}_j$

Apprentissage seul, jamais le jeu complet.

Standardisation avant découpage : information du test dans l'apprentissage, fuite de la séance 1.

Effet faible sur Iris, massif sur un jeu déséquilibré ou petit. Règle non négociable.

32

Une chaîne indivisible

```
1 from sklearn.pipeline import make_pipeline
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.neighbors import KNeighborsClassifier
4
5 modele = make_pipeline(StandardScaler(), KNeighborsClassifier(n_neighbors=7))
```

make_pipeline : prétraitement lié au classifieur. Validation croisée → standardisation réestimée **dans chaque pli**.

33

Et Mahalanobis ?

Séance 2 : $d_{\Sigma}^2(x, \mu) = (x - \mu)^{\top} \Sigma^{-1} (x - \mu)$.

Standardisation = $\Sigma = \text{diag}(\sigma_j^2)$: étalements corrigés, corrélations ignorées.

Distance euclidienne après blanchiment complet = distance de Mahalanobis. Métrique = modèle.

34

Choisir k

36

protocole classique

1. **grille** de valeurs de k , fixée avant tout résultat ;
2. erreur de chaque k par **validation croisée** ;
3. k retenu, puis un **seul** passage sur le jeu de test.

Erreur d'apprentissage inutile : décroissante avec $1/k$, quasi nulle en $k = 1$.

36

La courbe de validation

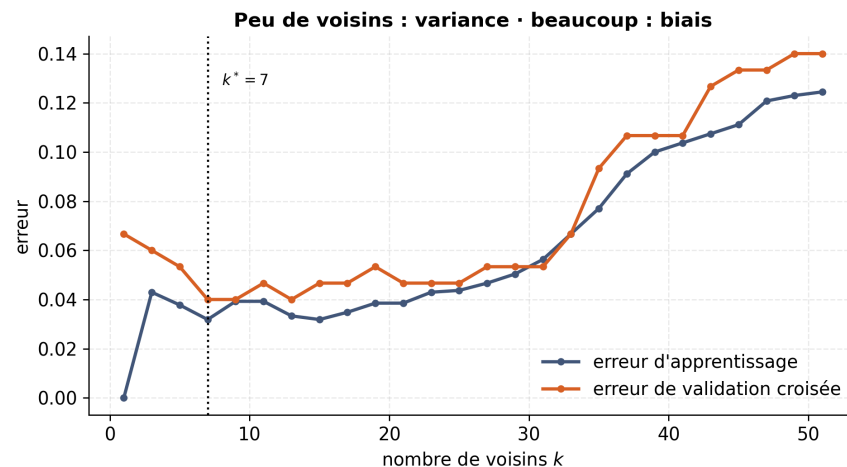


Figure 7

37

Le fléau de la dimension

38

Coût d'une variable inutile

$$d_{d+1}^2(x, z) = d_d^2(x, z) + (x_{d+1} - z_{d+1})^2$$

Terme ajouté **positif**, sans information sur Y . Même bruit dans toutes les distances, écart informatif noyé.

Aucune compensation : pas de coefficient à annuler, contrairement à LDA.

39

Bruit ajouté à Iris

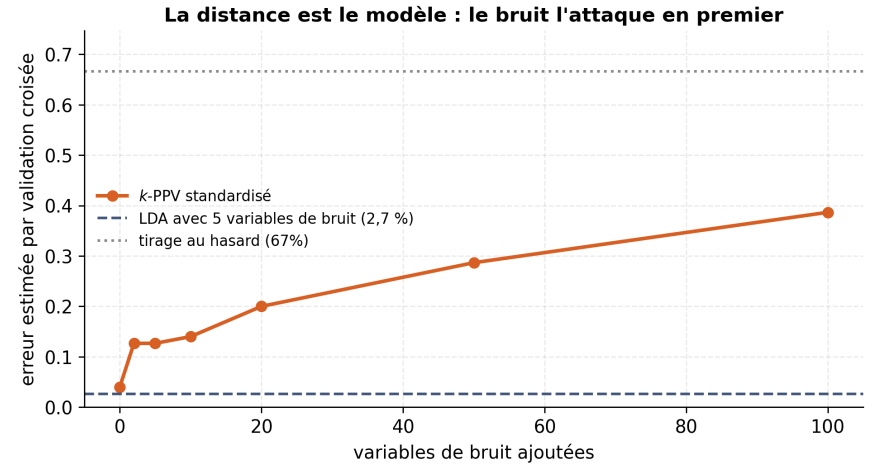


Figure 8

40

Réponses pratiques

- variables choisies par connaissance du domaine ;
- variables quasi constantes écartées ;
- sélection **dans** la validation croisée, jamais avant ;
- réduction de dimension — ACP de la séance 6 ;
- distance pondérée plutôt que subie.

41

Le coût de calcul

Apprentissage

Stockage des données. Coût nul.

Prédiction

n distances en dimension d par point : $O(nd)$.

Index spatiaux (**kd_tree**, **ball_tree**) : gain en dimension modérée, perdu au-delà d'une vingtaine de variables — même fléau.

Apprentissage gratuit, prédiction coûteuse : inverse des modèles paramétriques.

42

En bref

1. k -PPV : $P(Y = q \mid x)$ par fréquence dans un voisinage adaptatif, sans modèle de densité.
2. 1-PPV : jamais plus du double du risque de Bayes ;
 $k_n \rightarrow \infty, k_n/n \rightarrow 0 \rightarrow$ risque de Bayes atteint.
3. Complexité en n/k ; choix de k par validation croisée.
4. Distance = modèle : standardisation, dans chaque pli.
5. Exposition maximale aux variables inutiles.