

Tout savoir sur le K nearest neighbors

Partie 1 : La théorie



Présenté par **Morgan Gautherot**



Problème de classification

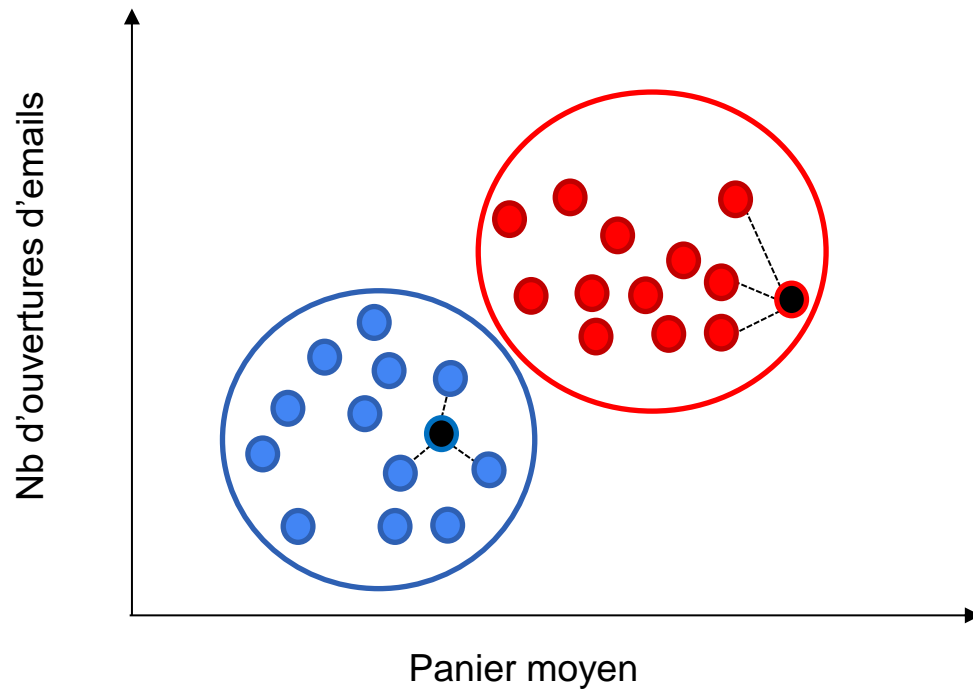


	Nb d'e-mails ouverts (x_1)	Nb de produits achetés (x_2)	Panier moyen (x_3)	Ouverture de l'e- mail (y)
1	12	3	120	1
2	0	1	40	0
3	30	10	1800	1
4	14	5	799	1
...
m	25	2	260	0

Jeu d'entraînement pour la prédiction de prix de maison



Utiliser les distances



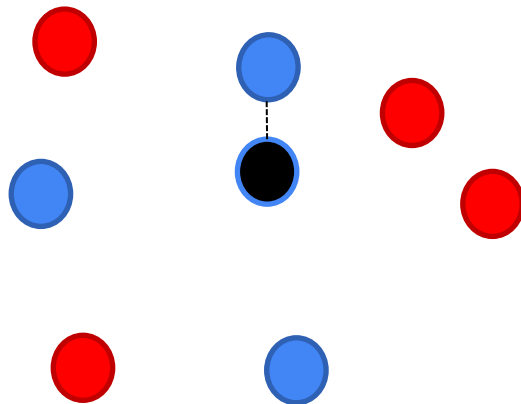


L'algorithme de prédiction

1. Sélectionnez le nombre de k voisins
2. Calculez la distance
3. Prenez les K voisins les plus proches
4. Attribuez la prédiction au nouveau point



Déterminer K

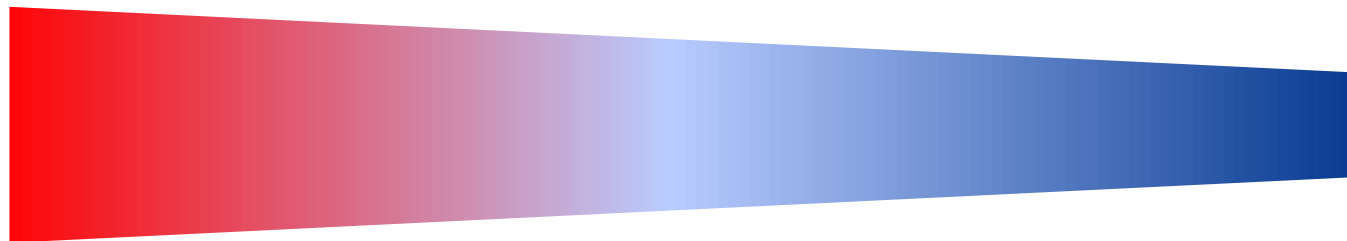




Sur et sous apprentissage

Sur-apprentissage

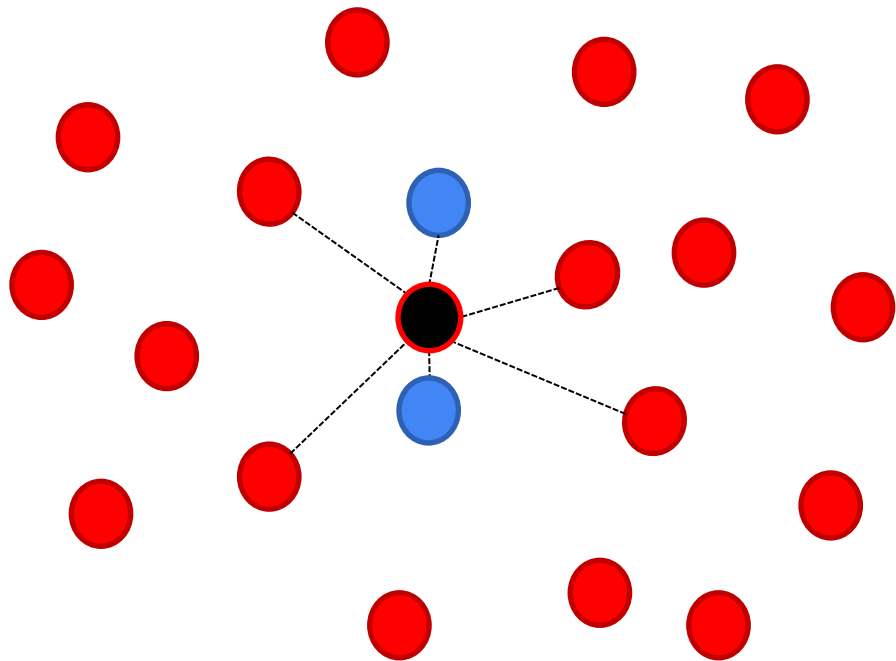
Sous-apprentissage



Augmentation de la valeur de K



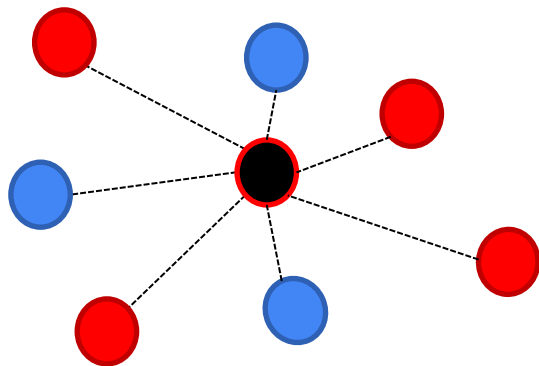
Valeurs aberrantes





K est une valeur impaire

Bleu ou rouge ?



$K = 6$

3 rouge

3 Bleu



L'algorithme de prédiction

1. Sélectionnez le nombre de k voisins
2. Calculez la distance
3. Prenez les K voisins les plus proches
4. Attribuez la prédiction au nouveau point



Qu'est ce qu'une distance ?

Distance euclidienne

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

Distance de Manhattan

$$d(x, y) = \left(\sum_{i=1}^m |x_i - y_i| \right)$$

Distance de Minkowski

$$d(x, y) = \left(\sum_{i=1}^n |x_i - y_i| \right)^{1/p}$$

⋮

⋮



Standardisation des données

$$x_{std} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

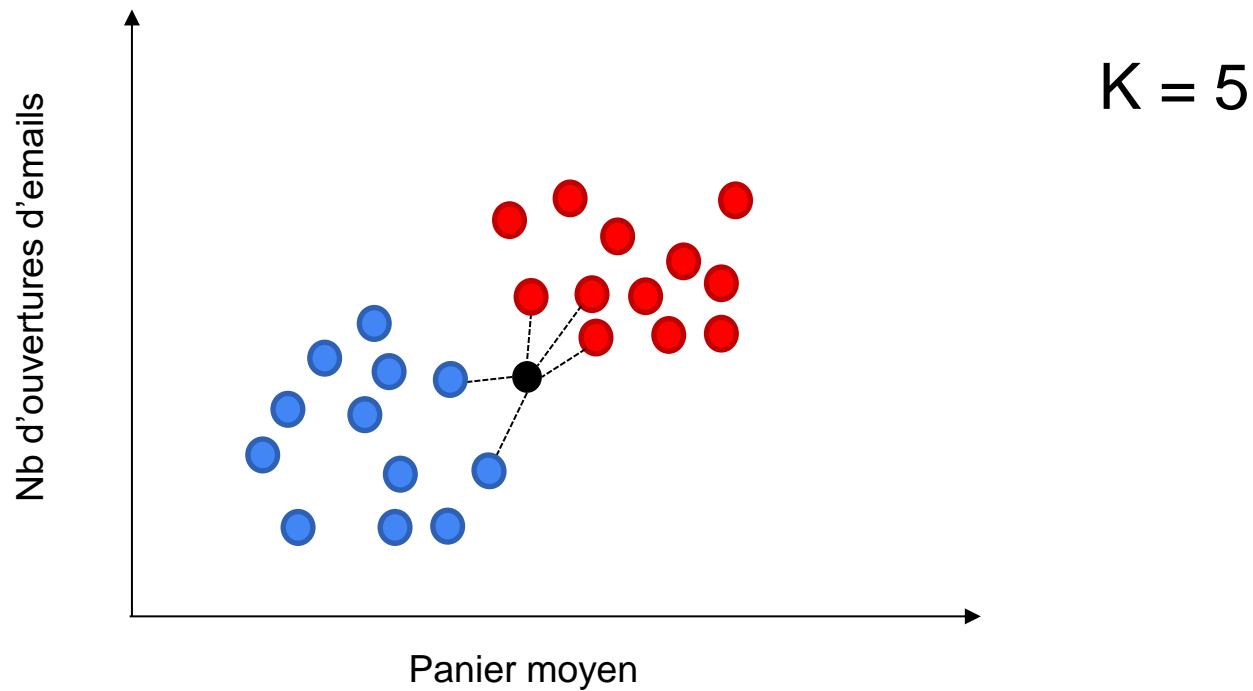


L'algorithme de prédiction

1. Sélectionnez le nombre de k voisins
2. Calculez la distance
3. Prenez les K voisins les plus proches
4. Attribuez la prédiction au nouveau point

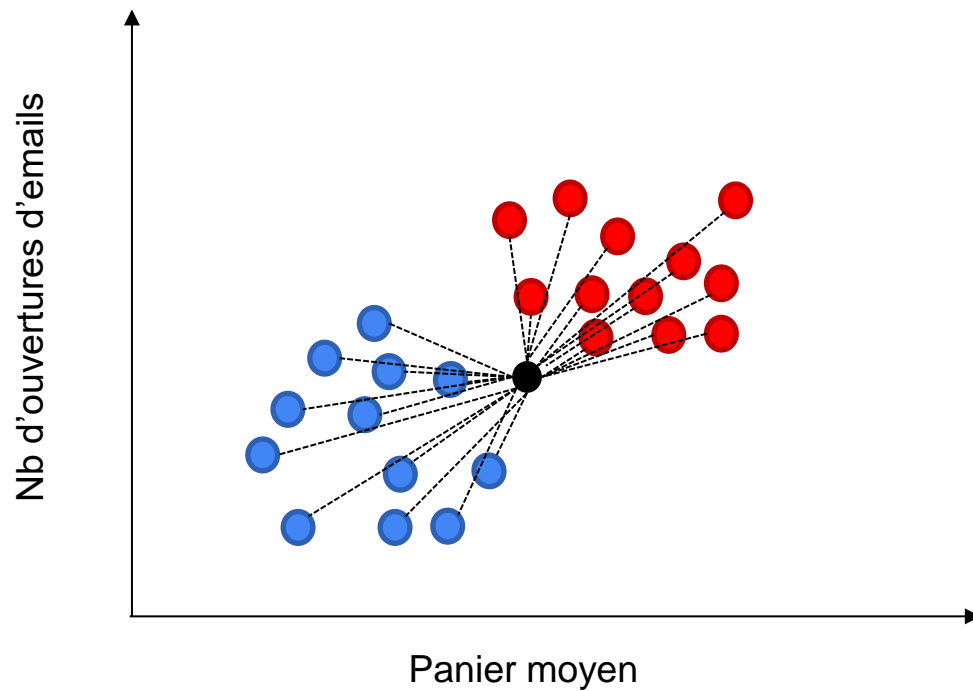


K plus proche voisins





K plus proche voisins




$K = 5$



Alternative au brut force

Multidimensional binary search trees used for associative searching

Author:  Jon Louis Bentley. [Authors Info & Claims](#)

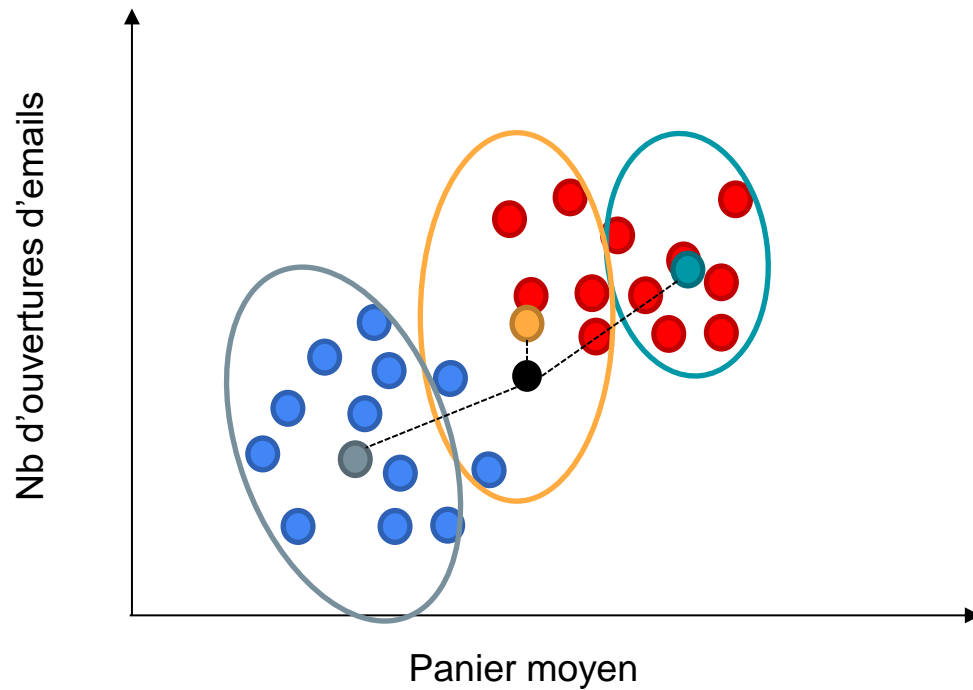
Communications of the ACM, Volume 18, Issue 9 • Sept. 1975 • pp 509–517 • <https://doi.org/10.1145/361002.361007>

Five Balltree Construction Algorithms

Title	Five Balltree Construction Algorithms
Publication Type	Technical Report
Year of Publication	1989
Authors	Omohundro, S.
Other Numbers	562
Abstract	Balltrees are simple geometric data structures with a wide range of practical applications to geometric learning tasks. In this report we compare 5 different algorithms for constructing balltrees from data. We study the trade-off between construction time and the quality of the constructed tree. Two of the algorithms are on-line, two construct the structures from the data set in a top down fashion, and one uses a bottom up approach.
URL	http://www.icsi.berkeley.edu/ftp/global/pub/techreports/1989/tr-89-063.pdf



K plus proche voisins



$K = 5$

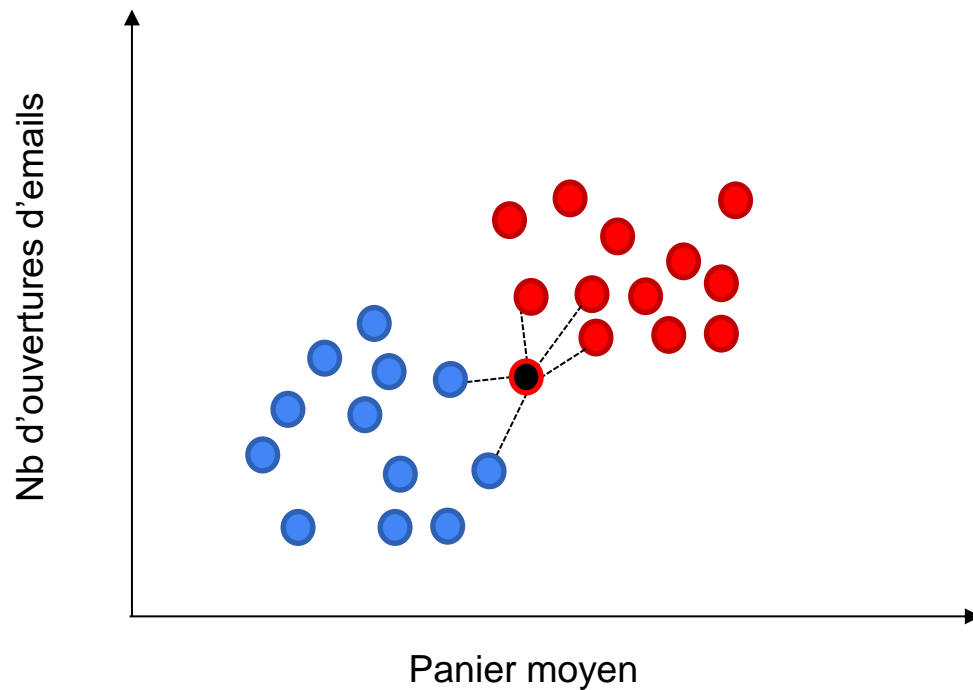


L'algorithme de prédiction

1. Sélectionnez le nombre de k voisins
2. Calculez la distance
3. Prenez les K voisins les plus proches
4. Attribuez la prédiction au nouveau point



La prédiction pour la classification



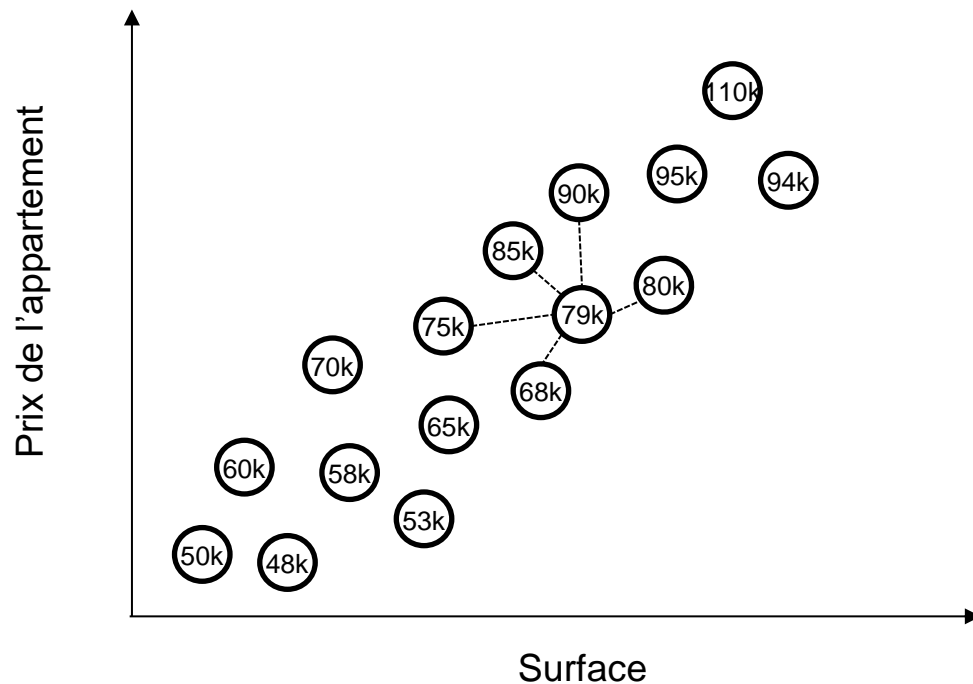
$K = 5$

3 rouge

2 Bleu



La prédiction pour la régression



$$K = 5$$

$$\begin{array}{r} 90k \\ + 85k \\ + 80k \\ + 75k \\ + 68k \\ \hline = 398k / 5 = 79k \end{array}$$



Avantages et inconvénients



- Facile à comprendre
- Facile à adapter
- Très peu d'hyperparamètre



- Ne scale pas bien
- Souffre de la curse of dimensionality
- Surentraîne facilement