Chapitre 1

L'algorithme des k plus proches voisins

Cet algorithme s'appelle k *Nearest Neighbors* en anglais, nous l'appellerons donc kNN.

1. Des questions concrètes

1. Une entreprise de vente d'articles en lignes collecte les informations de ses clients. Elle en a accumulé un grand nombre, tels que

   – l'âge ;
   – le sexe ;
   – l'adresse ;
   – les revenus moyens mensuels ;
   – *et cætera*.

   En fonction des achats, réguliers ou non, elle a placé ses client.e.s dans des catégories telles que

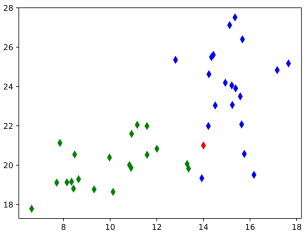
   – client·e fidèle ;
   – client·e à fidéliser ;
   – *et cæetera*.

   Un nouveau client se présente.
   Connaissant son âge, son sexe, son adresse et ses revenus mensuels, dans quelle catégorie l'entreprise va-t-elle le placer ?

2. On a mesuré la largeur et la longueur des pétales et des sépales d'iris de 3 catégories (*iris setosa*, *iris versicolor* et *iris virginica*) On effectue des mesures sur une nouvelle fleur. Dans quelle catégorie va-t-on la placer ?

Pour répondre aux deux questions précédentes on utilise le même algorithme : kNN.

2. Un algorithme pour y répondre



On considère deux nuages de points, l'un vert et l'autre bleu. Le point rouge (appelons-le P) doit-il être considéré comme appartenant au nuage vert ou au nuage bleu ?

Pour répondre à cette question, on va

Méthode : kNN

- choisir un entier k impair (3, 5 ou 7 typiquement) ;
- calculer les distances entre P et tous les autres points ;
- sélectionner les k points les plus proches de P ;
- regarder leurs couleurs.

Puisque k est impair il n'y aura pas de situation d'*ex æquo* et, suivant la couleur majoritaire des « k plus proches voisins » de P, on pourra choisir celle de P.

C'est cela, l'algorithme kNN.

La seule chose qui change selon la situation, c'est la *distance* que l'on utilise. Avec les nuages de points, on utilise la distance euclidienne :

$$d(A,B) = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$$

Mais on peut utiliser une distance différente suivant la situation.
Par exemple, si on ne manipule non plus des points mais des quadruplets $A = (a_0, a_1, a_2, a_3)$, alors on peut poser

$$d(A,B) = \sqrt{(b_0 - a_0)^2 + (b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2}$$

ou encore

$$d(A,B) = |b_0 - a_0| + |b_1 - a_1| + |b_2 - a_2| + |b_3 - a_3|$$

Et d'ailleurs, si c'est la proximité de la première composante qui importe le plus, on peut définir

$$d(A,B) = 50 \times |b_0 - a_0| + |b_1 - a_1| + |b_2 - a_2| + |b_3 - a_3|$$

Pour plus de renseignements sur ce qu'est une distance, tu peux consulter Wikipédia.