Classifications nos supervisées

UFR Environnement - Departement PCMI

PLAN

- CHI- Généralités
- CHII – Analyse en Composante Principale (ACP)
- Classification Ascendante Hiérarchique (CAH)
- K-means clustering

CHIII- K-means clustering

CHIII- K-means clustering

Introduction

K-means est l'approche de clustering partitionnel la plus populaire et la plus simple.

- Chaque cluster est associé à un centroïde (point central). Le Le nombre de clusters K doit être choisi à l'avance.
- Chaque point est affecté au cluster dont le centroïde est le plus proche.

Les deux questions qui se posent avec K-means sont :

1. Comment choisir K.
2. En supposant que l'on ait choisi K, comment procède-t-on au regroupement ?

Nous traiterons d'abord la deuxième question, en attendant la première pour l'instant.

CHIII- K-means clustering

K-means : algorithmes et emarque

I-1 Algorithme du k-means

L'algorithme de base est simple :

1- sélectionner K centroïdes initiaux

2- Repeter

2-1 : assigner chaque point au centroides le plus proche

2-2: recalculer le centroide de chaque cluster

3- Jusqu'à ce qu'aucun centroid ne cange de cluster

CHIII- K-means clustering

I-2 Remarques

Les K centroïdes initiaux sont souvent choisis au hasard. Les clusters produits varient donc d'une exécution de K-means à l'autre.

- Le centroïde est (généralement) la moyenne des points de la grappe (il s'agit de K moyennes, après tout !).
- La proximité est mesurée par toute fonction de distance valide, souvent (mais pas toujours) euclidienne.
- Vous souhaitez généralement (mais pas toujours) mettre à l'échelle/normaliser vos caractéristiques avant d'exécuter les moyennes K.
- Les K-means convergent assez rapidement pour ces mesures de distance courantes. La plupart de la convergence se produit dans les premières itérations.
- La complexité de chaque itération est O(n - K - D) où n est le nombre de points de données, K le nombre de clusters et D le nombre de caractéristiques.

CHIII- K-means clustering

I-3 solutions au problème des "centroïdes initiaux

- Redémarrage multiple (peut aider, mais la probabilité n'est pas de votre côté)
- Sélectionner plus de K centroïdes initiaux, puis sélectionner les centroïdes les plus éloignés les uns des autres parmi ces centroïdes initiaux
- Post-traitement
- Différentes stratégies d'initialisation (par exemple K-means++)
- Échantillonner et utiliser le regroupement hiérarchique pour déterminer les centroïdes initiaux (l'échantillonnage réduit le coût de calcul du regroupement hiérarchique).

CHIII- K-means clustering

II- k-means++

Voici comment fonctionne l'initialisation de K-means++ :

- Initialiser en choisissant 1 centroïde au hasard, m1.
- Puis pour k = 2,...,K :
    1. Pour chaque point, calculer di comme la distance minimale de xi aux centroïdes existants.
    2. Choisir xi comme centroïde initial k avec une probabilité proportionnelle à di^2 .

Ainsi, les points éloignés des centroïdes existants ont plus de chances d'être choisis.

Remarque : il s'agit simplement d'une stratégie d'initialisation différente. Après l'initialisation, le fonctionnement est le même que celui de K-means.

CHIII- K-means clustering

IV- K-means ne trouve qu'un optimum local

K-means tente de minimiser le SSE à l'intérieur des cluster. Mais :

- Il ne trouve jamais d'optimiseur global, sauf pour les problèmes simples.
- Il y a trop de regroupements possibles pour les vérifier tous !

A(N,K) = Nombre d'affectations de N points à K groupes

$$= \frac{1}{K!}\sum_{j=1}^{K}(-1)^{K-j}\cdot\binom{K}{j}\cdot j^{N}$$

$A(10, 4) = 34, 105$

$A(25, 4) \approx 5 \times 10^{13}$

CHIII- K-means clustering

III- K-means : évaluation de l'adéquation à l'échantillon

Soit mk le centroïde du cluster k, et Ck l'ensemble des points du cluster k (c'est-à-dire pour lesquels γi = k). Deux propriétés d'un regroupement "bien adapté" sont :

- La somme des carrés à l'intérieur des grappes doit être faible :

$$\mathrm{SSE}_W = \sum_{k=1}^{K} \sum_{x_i \in C_k} d(m_k, x_i)^2$$

- La somme des carrés entre les clusters doit être élevée :

$$\mathrm{SSE}_B = \sum_{k=1}^{K} d(m_k, \bar{x})^2$$

- où x¯ est la moyenne globale de l'échantillon.

CHIII- K-means clustering

IV Choisir K

Rappelez-vous : K est une donnée d'entrée de K-means, et non un paramètre estimé. Il n'existe pas de "meilleure" méthode généralement acceptée pour choisir K, et il n'y en aura jamais.

Quelques heuristiques utilisées dans la pratique :

- le choisir à l'avance sur la base de connaissances ou d'utilités préalables
- trouver le "coude" sur un graphique de SSE_W en fonction de K.
- optimiser l'un des nombreux critères quantitatifs de sélection des modèles (indice CH, AIC, BIC, gap statistique. . .)

Mais le principe le plus populaire est probablement le suivant :

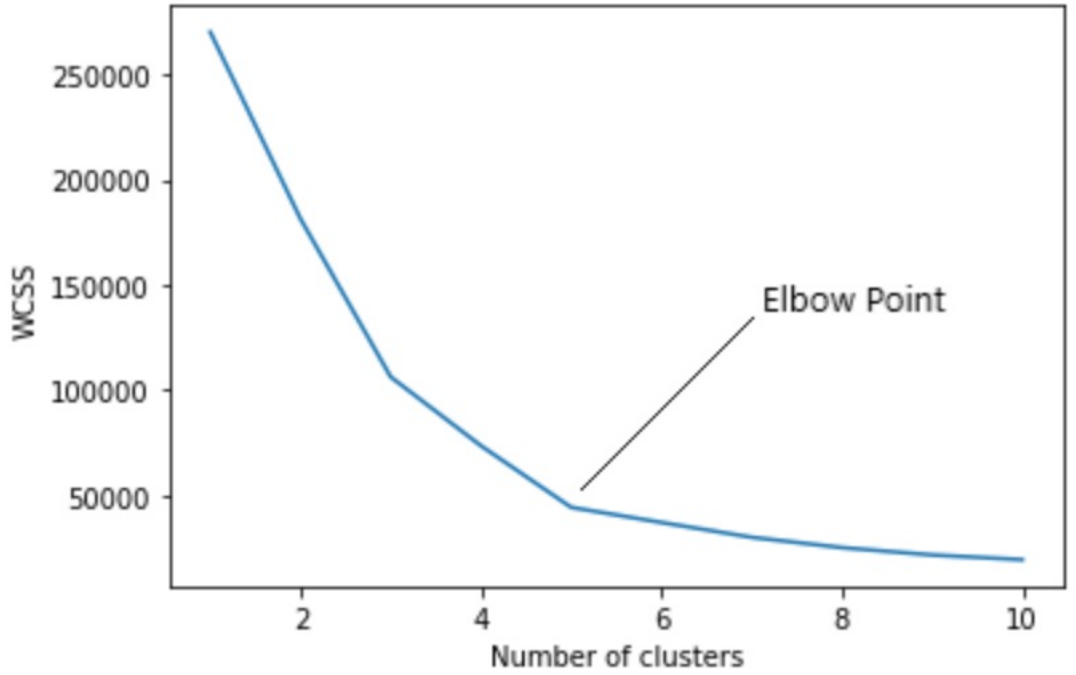
- satisfaire, ne pas optimiser : choisir une valeur de K qui donne des regroupements que vous et vos parties prenantes pouvez interpréter, et s'en tenir là.
- Il n'y a absolument aucune honte à cela, et les gens intelligents le font tout le temps.

CHIII- K-means clustering

V Tracé du coude (Elbow)

La regle du Elbow ou méthode du “coude” peut contribuer à déterminer le nombre optimal de cluster.

Elle consiste à tracer le nombre de clusters sur l’axe des abscisses et la distance moyenne des points de données aux centres de leurs clusters respectifs (WCSS) sur l’axe des ordonnées. On obtient alors une courbe en « coude » comme sur le graphique suivant :



CHIII- K-means clustering

V Tracé du coude (Elbow)

WCSS - Il est défini comme la somme des distances carrées entre les centroïdes et chaque point.

Le nombre de clusters optimal est généralement considéré comme étant celui qui correspond au “coude” de la courbe. C’est-à-dire que pour un nombre de clusters inférieur à celui du coude, la distance moyenne des points de données aux centres de leurs clusters diminue rapidement, tandis que pour des valeurs supérieures, cette diminution est moins prononcée.

Il faut cependant noter que la méthode du coude n’est pas infaillible et que son résultat dépend de la façon dont les données sont distribuées. Il est donc recommandé de vérifier les résultats obtenus avec d’autres méthodes ou indicateurs pour s’assurer de la pertinence du nombre de clusters choisis.

CHIII- K-means clustering

VI Indice CH

L'indice CH n'est que l'un des nombreux critères de sélection de modèles proposés pour choisir K dans le regroupement K-means. Il tente de trouver un équilibre entre l'adéquation et la simplicité.

Soit nj le nombre de points dans le cluster j, et définissons:

Variance intra.

$$B(K) = \sum_{j=1}^{K} n_j d(m_j, \bar{x})^2$$

Variance inter

$$W(K) = \sum_{j=1}^{K} \sum_{x_i \in C_j} d(x_i, m_j)^2$$

L'indice CH est défini comme suit

$$CH(K) = \frac{B(K)}{W(K)} \cdot \frac{N-K}{K-1}$$

- Le premier terme, B(K)/W(K), est d'autant plus grand que K est grand (meilleur ajustement).
- Le second terme, (N - K)/(K - 1) est d'autant plus petit que K est grand (modèle plus complexe).

Choisissez donc K^ =argmaxCH(K).

Si vous connaissez les tests F dans la régression et l'ANOVA, cette statistique peut vous sembler familière.

CHIII- K-means clustering

VII- Gap statistic

Une autre méthode courante pour choisir K est le gap statistique :

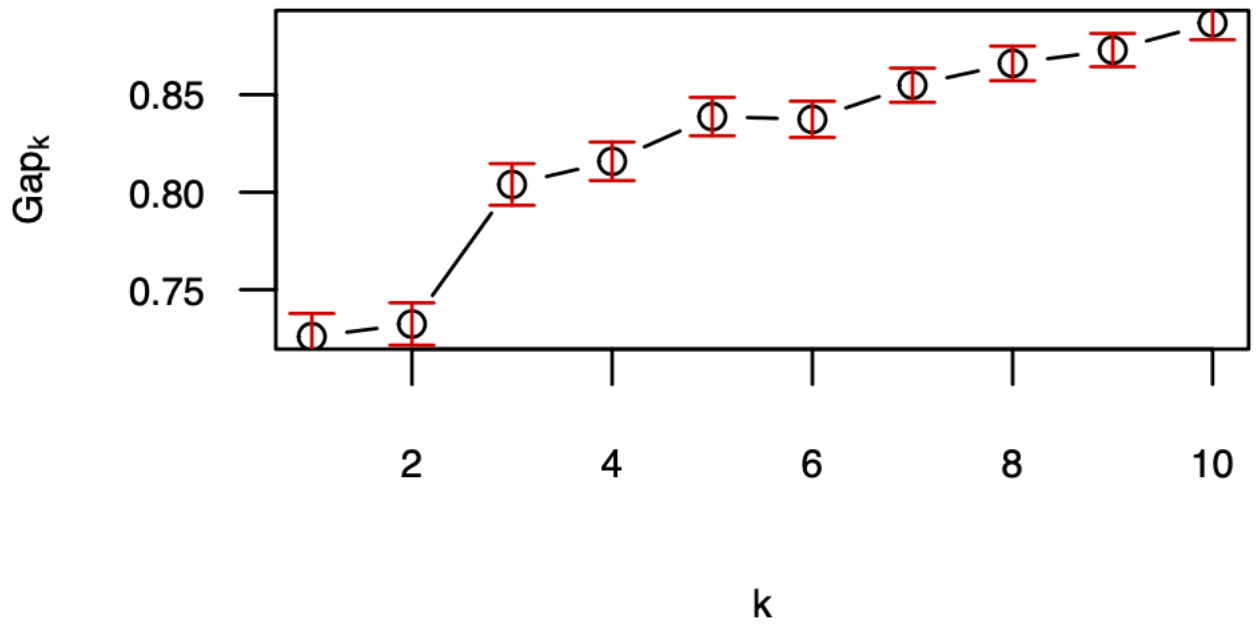
- trouver un moyen de normaliser la comparaison de WK (ou en fait, de log WK) avec une distribution de référence ou de référence "nulle" des données, c'est-à-dire une distribution sans regroupement évident.
- Plus le log W_k est éloigné de cette courbe de référence, meilleur est le regroupement. Cette information est contenue dans la formule suivante pour le gap statistique :

$$\mathrm{Gap}_n(K) = E_n^{\star}[\log W_K] - \log W_K$$

- Ici, la valeur attendue est prise sous l'hypothèse "nulle" d'absence de regroupement, c'est-à-dire lorsque les points de données sont distribués uniformément à l'intérieur de la boîte de délimitation des données d'origine.

CHIII- K-means clustering

VII- Gap statistic



La méthode de sélection par défaut se contente de rechercher le premier pic local, jusqu'à l'erreur standard dans l'estimation de E_n^*[log WK].

Ici, c'est à K=5.

CHIII- K-means clustering

RESUMER

Avantages et inconvénients :

- Bon pour les clusters de forme convexe, mais moins bon pour les autres formes
- Il faut choisir K et il n'y a pas de méthode évidente pour le faire
- Super rapide, super évolutif, et en quelque sorte efficace
- Généralement la première chose que les gens essaient dans un problème de clustering

NB : J'ai généralement eu plus de chance avec le gap statistique comme heuristique de sélection de modèle.

Mais n'hésitez pas à choisir une valeur de K qui vous semble raisonnable !