Cours ENSTA
Adrien CHAN-HON-TONG

Compléments théoriques (K moyennes et PCA)

1 : K moyennes

Introduction

Il est important de ne pas confondre *objectif* et solution. Ici, on se donne un problème [1] (objectif) qu'on cherche à résoudre et qui est celui du *clustering*.

Étant donnée $x_1, ..., x_N \in \mathbb{R}^D$ et $K \in \mathbb{N}$ on cherche $c_1, ..., c_K$ tel que

$$c_1, ..., c_K = \min_c \sum_n \min_k ||c_k - x_n||_2^2$$

Sans équation, on cherche à minimiser la somme de la distance minimale entre x_n et un des c_k.

Malheureusement, ce problème est NP complet dès $K = 2$ vis à vis de N, mais aussi, quand K est non borné même si N l'est. Donc, en pratique on ne sait pas résoudre ce problème **exactement**.

Approximation

L'algorithme des K moyennes donne une solution approximative à ce problème très efficace en pratique. Là distinction entre l'objectif et la solution n'est pas triviale car l'algorithme des K moyennes est (peu ou prou) la seule façon d'attaquer ce problème. Mais, il est important d'un point de vue théorique de distinguer les 2.

Maintenant l'algorithme des K moyennes est en fait construit sur un schéma classique d'optimisation alternée : On ne sait **pas** résoudre

$$c_1, ..., c_K = \min_c \sum_n \min_k ||c_k - x_n||_2^2$$

Mais on sait résoudre cette équation si on découple les minimisations !

Par exemple, on sait résoudre

$$c_1, ..., c_K = \min_c \sum_n ||c_\sigma(n) - x_n||_2^2$$

si σ est fixe : en effet cela découple l'optimisation des différents c et donc cela revient à résoudre K problème de la forme

$$\min_c \sum_n ||c - x_n||_2^2$$

1. Bien entendu à ce stade, c'est un très abstrait : on se donne un problème dont on ne voit pas à quoi il sert. C'est l'objet des slides.

Il suffit de poser le problème et de développer le carré pour voir que la solution est $c = \frac{1}{N}\sum_n x_n$.

De même on sait résoudre

$$\min_\sigma \sum_n ||c_\sigma(n) - x_n||_2^2$$

si les c sont fixes : il suffit pour chaque n de choisir son plus proche voisin parmi les c puisque ça découple le problème en N problème trivial.

Algorithme

Il faut réussir à passer au dessus des équations pour réaliser qu'on dit des choses très simples - élémentaires - niveau collège max : on dit simplement que la moyenne c'est ce qui minimise la distance cummulé à un ensemble de point et que étant donné un point et un nuage points, c'est son plus proche voisin qui minimise la distance du point au nuage de points.

Donc combiner on ne sait pas résoudre - mais séparer - on a affaire à 2 problèmes triviaux !

L'idée de l'algorithme des K moyennes est simplement, d'alterner optimisation des c à σ fixe puis optimisation de σ à c fixe, et cela un grand nombre de fois.

Cette boucle composée de 2 étapes est très bien illustrée sur wikipédia (en français et en anglais - les illustrations sont différentes et complémentaires).

Bien entendu, cette algorithme ne convergent pas vers la solution exacte (et même si elle le faisait ce serait en temps exponentiel donc...), mais est très efficace en pratique. L'élément le plus crucial est l'initialisation.

Essayer de prendre des c loin les un des autres est considéré comme pertinent c'est l'initialisation K-means++.

2 : PCA

Pour démystifier la notion de PCA, on va résoudre le problème de la PCA pour $D = 2$ quand on cherche l'axe principale c'est à dire la projection qui maximise la variance.

On a $x_1, ..., x_N \in \mathbb{R}^2$, tel que $\sum_n x_n = \mathbf{0}$ et on cherche $u \in \mathbb{R}^D$ tel que

$$u = \max_u \sum_n \frac{(u^T x_n)^2}{u^T u}$$

Notons, $A = \sum_n x_{n,1}^2$, $B = \sum_n x_{n,2}^2$ et $C = \sum_n x_{n,1}x_{n,2}$ et paramétrons $u = (1 \quad t)$, on obtient l'équation :

$$\max_t \frac{1}{1+t^2}\sum_n x_{n,1}^2 + x_{n,2}^2 t^2 + 2x_{n,1}x_{n,2}t = \frac{1}{1+t^2}(A + Bt^2 + 2Ct)$$

Posons $f(t) = \frac{1}{1+t^2}(A + Bt^2 + 2Ct)$
$f'(t) = \frac{1}{1+t^2}(2Bt + 2C) - \frac{2t}{(1+t^2)^2}(A + Bt^2 + 2Ct)$
$f'(t) = 0 \Leftrightarrow (1 + t^2)(Bt + C) - t(A + Bt^2 + 2Ct) = 0$
$\Leftrightarrow Bt + C + Bt^3 + Ct^2 - At - Bt^3 - 2Ct^2 = 0$
$\Leftrightarrow -Ct^2 + (B - A)t + C = 0$ si $C = 0$, alors $t = 0$ sinon
$\Leftrightarrow t^2 - \frac{B-A}{C}t - 1 = 0$
$\Leftrightarrow t = \frac{B-A}{2C} \pm \sqrt{1 + (\frac{B-A}{2C})^2}$

La PCA c'est une généralisation quand la dimension est quelconque, et, qu'on cherche pas uniquement 1 vecteur mais plusieurs... Cf les slides.

Soit une famille de fonctions notée $\mathcal{H}$ de dimension de Vapnik–Chervonenkis $VC(\mathcal{H})$ fini.

Si on considère un problème donnée par une distribution P et qu'on tire une base B de N éléments selon P alors pour tout $h \in \mathcal{H}$, on a

$$P\left(E_{reel}(h) \leq E_{emp}(h, B) + \frac{1}{\sqrt{N}}\sqrt{VC(\mathcal{H})\log(\frac{2N}{VC(\mathcal{H})}) - \log(1/4\delta)}\right) \leq 1 - \delta$$

Notamment, si on considère h^* une fonction de $\mathcal{H}$ qui fait 0 erreur sur B (on suppose qu'elle existe), on a

$$P\left(E_{reel}(h^*) \leq \frac{1}{\sqrt{N}}\sqrt{VC(\mathcal{H})\log(\frac{2N}{VC(\mathcal{H})}) - \log(1/4\delta)}\right) \leq 1 - \delta$$

Maintenant si on considère la famille de fonction restreinte à h^* c'est à dire $\mathcal{T} = \{h^*\}$. Mais, il me parait faux [2] de considérer qu'on peut applique la borne à $\mathcal{T}$: avec $VC(\mathcal{T}) = 1$ et dont le seule élément a une erreur empirique sur B de 0 ($VC(\mathcal{T}) = 1$ puisque n'importe quel problème à 1 point non consistant à h^* ne peut être scattered par $\mathcal{T}$) sinon on aurait

$$P\left(E_{reel}(h^*) \leq \frac{1}{\sqrt{N}}\sqrt{\log(2N) - \log(1/4\delta)}\right) \leq 1 - \delta$$

Cependant, quand on va valider un système, on sera toujours plus ou moins dans ce dernier cas puisqu'en machine learning, on définie l'algorithme à la lumière du problème et non a priori comme c'est le cas quand on veut valider une hypothèse en stat.

Moins caricatural : si l'industriel a testé arbre de décision, VGG et ResNet, la VC dimension c'est celle de la solution retenu ou des 3 ?

2. Enfin, en soit cette inégalité n'est pas vraiment fausse mais inutile : on est dans le cas où on est sure d'être dans le δ c'est à dire le cas où $E_{reel}(h^*) > \frac{1}{\sqrt{N}}\sqrt{\log(2N) - \log(1/4\delta)}$.