Curse of dimension & N-p Norm

2017

1 Fléau de la grande dimension

1.1 Introduction & exemple

1.1.1 Introduction

Le fléau de la grande dimension (ou curse of dimensionality, en anglais à l'origine) est un concept inventé et défini pour la première fois par Richard Bellman en 1961 afin de de définir et de présenter les divers conséquences liées à l'analyse de données en grande dimension, phénomènes qui étant propres à ces grands espaces ne sont pas reproductibles dans des dimensions plus petites.

Bien que beaucoup de domaines de la statistique soient concernés par ce phénomène, il en est un qui est particulièrement impacté : L'apprentissage statistique, ou Machine Learning, dont certaines méthodes sont souvent basées sur des métriques afin de déterminer la proximité relative des données, peut s'avérer partiellement si ce n'est totalement inefficace dans des grands espaces. En effet le nombre important de dimensions "éclate" les données et ne permet donc plus de discerner précisément et efficacement leurs écarts puisque les points se retrouvent davantage dispersés au sein de l'espace.

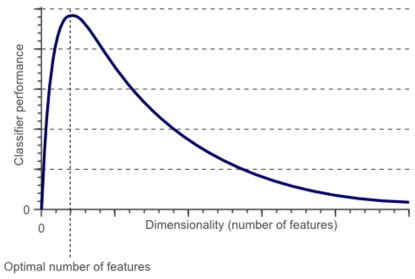

Figure 1: Illustration du fléau de dimension

1.1.2 Exemple

Prenons l'exemple suivant afin de mieux visualiser le problème. Imaginons un algorithme du K-plus proche voisins en dimension 1, avec trois catégories pour chaque variable. On peut représenter les données ainsi, sous formes de 3^1 zones:

Figure 2: 9 données au sein de 3 régions d'un espace unidimensionnel

En considérant désormais un espace de plus grande dimension ($d = 2$ dans un premier temps), on constate une augmentation du nombre de régions : on observe désormais 3^2 soit 9 zones si l'on souhaite conserver la même granularité. Par ailleurs, deux options s'offrent à nous : 1) Choisir de maintenir la même densité de données et donc augmenter le nombre d'observation total, ou 2) Choisir de garder le nombre total de points égal et alors obtenir une nouvelle représentation des données éparse :

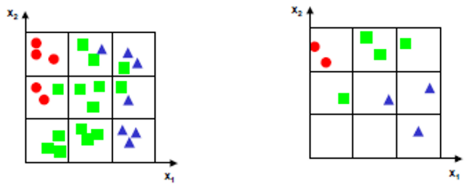

Figure 3: Respectivement 1) & 2)

On remarque par ailleurs qu'en choisissant l'option 1) le nombre données évolue de 9 en 1D à 27 en 2D. Lorsque que l'on passe en 3D, le problème devient encore pire. En effet le nombre de zones passe alors à 3^3, soit 27 régions. En choisissant de maintenir la densité il faudrait près de 81 observations (3 observations en 1D $\times$ 27 Régions en 3D), tandis ce que maintenir le nombre de points constants lors de l'agrandissement de l'espace revient à obtenir un cube pratiquement vide :

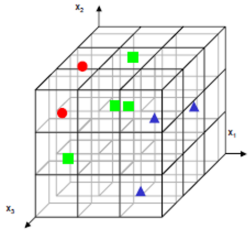

Figure 4: Représentation du cas 2) dans un espace en 3 dimensions

En extrapolant le problème pour des espaces de dimensions très importantes, comparable, par exemple à un problème d'apprentissage basé sur plusieurs centaines, voire milliers de variables, on comprend que le nombre d'observations nécessaires afin d'obtenir une densité de données suffisante au sein de l'hypercube peut s'avérer rédhibitoire compte tenu du nombre nécessairement limité d'observations ayant été réalisées.

Il a donc fallu trouver des solutions statistiques afin de pouvoir résoudre ce problème. C'est tout l'enjeu des méthodes parcimonieuses :

1.2 Parcimonie

Il existe plusieurs types de méthodes parcimonieuses afin de réduire l'espace d'étude statistique :

- Les méthodes par extraction de variables
- Les méthodes par sélection de variables

1.2.1 Méthodes par extraction de variables

L'objectif de ces méthodes consiste à transformer les variables corrélées en nouvelles variables décorrélées les unes des autres. L'une des méthodes des plus classiques dans ce domaine est l'analyse en composante principale (ACP).
Le principe est d'obtenir une représentation approchée du nuage des n individus de l'espace F (où $\dim F = p$) dans un sous-espace vectoriel de dimension plus faible. Cette méthode repose sur le principe de projection :
En effet les n individus de F sont projetés M-orthogonalement (où M définie la métrique choisie) sur un sous espace vectoriel, noté F_k avec $dim F_k = k$ et k petit. Le choix du sous-espace vectoriel de projection F_k s'effectue selon le critère suivant : Il faut que l'inertie du nuage des individus projetés sur le sous-espace F_k soit maximale (en effet en projection les distances ne peuvent que diminuer), ainsi la structure du nuage des individus est conservée au maximum après projection sur F_k.
L'inertie n'est par ailleurs qu'une mesure de la dispersion des individus d'un nuage de points autour d'un axe défini (par exemple le centre de gravité du nuage) :

Définition 1 *On définit l'inertie I_g comme la dispersion de n individus autour de leur centre de gravité, noté g, tel que :*

$$I_g = \sum_{i=1}^{n} p_i \parallel x_i - g \parallel_M^2$$

Avec p_i les poids de chaque individu x_i et g le barycentre.

Remarque :
Elle est une généralisation de la variance.

Par la suite afin de déterminer comment construire le meilleur sous espace F_k, on peut utiliser le théorème suivant :

Théorème 1 *Le sous-espace F_k de dimension k optimal est engendré par les k vecteurs propres de la matrice de corrélation de X (équivalent à la matrice de variance-covariance pour des données centrées-réduites) associés aux k plus grandes valeurs propres.*

Où X représente la matrice des données initiales telle que :

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^j & \dots & x_1^p \\ x_2^1 & x_2^2 & \dots & x_2^j & \dots & x_2^p \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_i^1 & x_i^2 & \dots & x_i^j & \dots & x_i^p \\ x_n^1 & x_n^2 & \dots & x_n^j & \dots & x_n^p \end{bmatrix}$$

Remarque :
La matrice de corrélation de X est symétrique de dimension p et semie-définie positive, elle possède donc p vecteurs propres orthogonaux deux à deux et ses valeurs propres sont toutes positives ou nulles.

Ainsi on peut en déduire simplement que dans le cadre de données centrées-réduites, l'inertie I_g du nuage des individus par rapport au centre de gravité g peut s'écrire :

$$I_g = \sum_{j=1}^{p} \lambda_j = p$$

Où λ_i représente la ième valeur propre associée à la matrice de variance-covariance des données centrées-réduites.

En utilisant l'ACP il est possible de "compresser" un ensemble de p variables aléatoires en utilisant leurs représentations issus des k premiers axes de l'analyse (où $k \leqslant p$). Ces axes étant considérés comme un meilleur choix du point de vue de l'inertie ou de la variance.
L'ACP permet donc de dégager les axes permettant d'expliquer au mieux la dispersion des données. Ainsi à titre d'exemple simple et concret une ACP permet de déterminer le meilleur plan permettant de représenter une structure en plusieurs dimensions tout en conservant le maximum d'informations.
Avec cet exemple on voit une représentation d'un chameau selon les deux premiers axes principaux (en rouge) issus d'une ACP, on remarque que l'on peut facilement distinguer l'animal à l'aide de ces deux plans malgré la réduction de dimension car le maximum d'informations a été conservé lors des deux projections. En utilisant deux autres axes il aurait été plus difficile si ce n'est impossible de déterminer l'animal en question puisque davantage d'information fondamentale aurait été sacrifiée lors de la réduction de dimension.

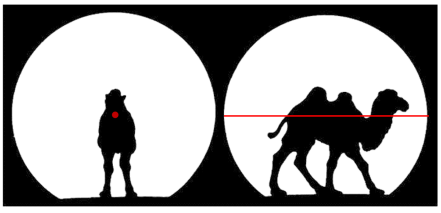

Figure 5: Exemple de projection sur deux axes principaux

Il existe par ailleurs d'autres méthodes factorielles pour analyser d'autres types de données, notamment l'analyse factorielle des correspondances (AFC) ou encore l'Analyse des correspondances multiples (ACM). Une ACP peut être très facilement réalisé grâce au package R FactoMineR.