UNIVERSITÉ DE SHERBROOKE

DÉPARTEMENT D'INFORMATIQUE

IFT 599/IFT799 : Science de données

TP #3 — Automne 2021

Analyse des données boursières par le clustering

À remettre le vendredi 3 décembre 2021

Le TP3 porte sur l'analyse des séries chronologiques. L'ensemble (le tableau) de données contient 288 colonnes dont la première représente le temps et les 287 autres représentent des différentes séries chronologiques correspondant aux différents indices, ETFs, actions, etc. Il y a exactement une valeur (prix de clôture) par jour ouvrable. La longeur de chaque série est de 5221 correspondant aux 5221 jours ouvrables du 2000-12-29 au 2021-03-23. Il n'y a pas de problème de données manquantes avec cet ensemble de données.

Ce TP comprend plusieurs tâches à effectuer, de même que des tâches supplémentaires pour les étudiantes et étudiants de l'IFT799. Il y aura des remises séparées pour l'IFT599 et l'IFT799. Si une personne de l'IFT599 et une personne de l'IFT799 font l'équipe, vous devez clairement indiquer cette information dans la page de couverture de votre TP. Vous devez le soumettre une seule fois, dans le groupe de l'IFT799. Ce TP compte pour 10% dans les notes finales.

Objectifs du TP

Les objectifs de ce TP sont multiples.

- Se familiariser avec des algorithmes de clustering.
- Se familiariser avec la détermination automatique du nombre de clusters.
- Suivie des trajectoires de changement dans les données financières.
- Découvrir et comprendre des phénomènes des régimes du marché.

Tâches de clustering

Pour ce TP, l'analyse de clustering doit être appliquée sur chacune des fenêtres de données que vous allez extraire. Une fenêtre de données correspond typiquement à un interval de 21 jours (= un mois) ou de 63 jours (= 3 mois). Une fenêtre de 21 jours contient donc

287 vecteurs (objets) de 21 composantes. L'analyse à effectuer sur une fenêtre de données inclut la détermination du nombre optimal de clusters et l'obtention de ces clusters. Les deux tâches se font ensemble dans la pratique.

Il est possible de générer beaucoup de fenêtres de données. Par exemple, si nous adoptons la taille de 21, les 5221 jours dans l'ensemble nous permettent de générer potentiellement 5201 fenêtres différentes en déplacant la fenêtre d'un jour à la fois. Si nous faisons le déplacement par 21 jours (pour éviter complètement la superposition entre deux fenêtres), nous pouvons quand même obtenir 248 ou 249 fenêtres. Dans ce cas, la détermination du nombre optimal de clusters et l'obtention de ces clusters devraient être effectuées 248 ou 249 fois. La complexité de calcul et le besoin en mémoire pourraient devenir extrêmement grands et hors de la porté de votre ordinateur personnel. C'est pourquoi, **pour ce TP, vous n'êtes pas obligés d'utiliser toutes les 287 séries, ni toutes les historiques de 5221 jours**. Voici quelques suggestions concernant les données à utiliser :

- Autant que possible, utilisez toutes les séries. Si vous décidez d'en utiliser moins, il ne faut pas descendre en-dessous d'une centaine.
- La longeur des séries doit correspondre au moins à une période de 5 ans consécutifs, c'est-à-dire, environ 1260 jours. Vous avez la liberté de choisir le période qui vous intéresse, mais c'est plus intéressant d'étudier les périodes où il y avait des grands movements sur le marché boursier, ex. la période couvrant 2008.
- Une fois la période et les séries choisies, vous devez normaliser les séries à moins que vous adoptiez une mesure de similitude qui est indépendante de l'amplitude pour le clustering.
- Il suffit de travailler avec des fenêtres de 21 jours. Vous pouvez vous contentez de déplacements de 21 jours (c'est-à-dire, pas de superposition). Idéalement, appliquez vos algorithmes sur des fenêtres superposées, par exemple par le déplacement de 10 jours.

Pour le clustering des données de chaque fenêtre, tout le monde doit implanter une solution comme suit.

- Le "k-means" OU le "FCM" avec la sélection aléatoire des centres initiaux de cluster. Si vous choisissez l'algorithme FCM, vous devez implanter un processus de "défuzzification", c'est-à-dire de transformer le membership flou en membership binaire en affectant à la fin du clustering chaque objet au cluster dont la valeur de membership est maximale. Pour l'un ou l'autre l'algorithme, la fonction de distance est supposée d'être la distance Euclidienne. Mais, il est tout à fait permis d'utiliser d'autres fonctions de distance si vous le jugez nécessaire.
- Comme indice de validation, implanter le "Silhouette", voir la page
  https://en.wikipedia.org/wiki/Silhouette_(clustering)
  L'indice de validité pour chaque valeur du nombre de clusters k est la moyenne de $s(i)$,

tandis que $s(i)$ est la valeur "silhouette" de chaque objet. Une des raisons principales de choisir la "silhouette" est qu'elle se base uniquement sur la distance entre les objets et peut être utilisée avec une grande variété d'algorithmes de clustering.

- Pour déterminer le nombre de clusters, vous devez aussi pré-déterminer le nombre minimal et le nombre maximal. Choisissez 2 comme le nombre minimal. Pour le nombre maximal, vous devez faire un peu d'essaie-et-erreur. 10 n'est probablement pas un mauvais choix.

Pour les étudiantes et les étudiants de l'IFT799, vous devez concevoir et implanter un deuxième algorithme de clustering. Voici les informations pertinentes à cette partie de travail :

- La fonction de distance à utiliser est le "dynamic time warping (DTW)". Voir https://en.wikipedia.org/wiki/Dynamic_time_warping
  Cette "distance", qui n'est pas une vraie distance mathématiquement parlant, permet de mesurer plus efficacement la différence entre deux séries.

- Quand utiliser DTW pour calculer la distance entre deux séries, on ne doit pas utiliser la moyenne des séries pour calculer un représentant d'un cluster. Vous devez proposer un algorithme de type k-means ici qui n'a pas besoin du centre (représentant) pour chaque cluster.

- Cependant, vous pouvez utiliser exactement le même indice de validation, le "Silhouette", implantée précédemment. Il y a d'ailleurs moyens d'optimiser vos codes pour le clustering et pour le calcul de l'indice comme les deux sont basés sur le calcul des distances entre des objets.

- Pour déterminer le nombre de clusters aussi, vous pouvez profiter les démarches que vous avez effectuées précedemment pour l'algorithme de k-means ou de FCM.

Analyse des résultats de clustering

Les travaux de clustering ci-dessus vous permettent d'obtenir des clusters à chaque fenêtre de données. Les voir dans le temps vous permet d'avoir une séquence d'ensembles de clusters. Chaque ensemble pourrait avoir une étiquette de temps que vous pouvez lui donner en choisissant le jour du centre de la fenêtre, le jour du début de la fenêtre, ou le jour de la fin de la fenêtre. Chacun des clusters dans un ensemble peut être considéré comme un "régime". On peut donc découvrir des informations comme, à une fenêtre ou à un moment, une telle série (ETF, ou indice, ou action) se trouve dans un tel régime.

Il y a deux tâches d'analyse à effectuer. **La tâche 1** consiste à evaluer comment la taille du plus grand cluster (régime) évolue :

- Pour chaque fenêtre, trouver la taille du plus grand cluster et dessiner la fonction de la plus grande taille comme une courbe dans le temps.

- Identifier des points dans cette courbe qui sont les plus succeptibles d'être des "outliers". Il y a toujours des points qui sont plus succeptibles que d'autres d'être "outliers". La question ici n'est pas d'identifier le plus possible des "outliers" (discussions en classe sur de différentes méthodes possibles).

- Marquer (souligner) le moment de ces outliers. Optionnellement, si vous connaissez un peu d'historiques de crash boursier, observer s'il y a une corrélation entre les moments de ces outliers et les moments des crashes.

La tâche 2 consiste à evaluer comment les régimes (clusters) évoluent globalement :

- Générer une courbe de la mesure "rand" en prenant l'actuel ensemble de clusters de la fenêtre courante comme "résultat du clustering" et l'ensemble de clusters de la fenêtre précédente comment "clusters originaux". Cette courbe donne une bonne idée si les régimes changent ou pas dans le temps.

- Optionnellement, vous pouvez aussi évaluer si les régimes (clusters) eux-même changent dans le temps. Il s'agit de comparer le profil (représentants) des clusters à un moment avec celui à un autre moment.

Pour les étudiantes et les étudiants de l'IFT799, comme vous implantez deux algorithmes de clustering, les tâches 1 et 2 de cette partie doivent être effectuées pour les résultats de chacun des dexu algorithmes. Si vous observez une différence importante dans les résultats de ces deux tâches, rapportez-les.

Comme pour le TP1, vous avez toujours beaucoup de liberté pour développer vos propres solutions. Vous devez faire preuve d'imagination. Il n'y a pas une solution parfaite et ce n'est pas le but du TP non plus. Le développement d'un esprit critique, la recherche de solutions et le savoir de se justifier sont bien plus important. C'est pourquoi, vous devez mettre du temps et de l'energie pour bien rédiger votre rapport. Votre rapport doit décrire clairement les différentes étapes de traintement incluant les prétraitements effectués, les résultats obtenus, vos commentaires et vos conclusions.

N'oubliez pas de remettre aussi vos données si elles sont modifiées des données originales. Normalement, le correcteur ne regarde pas les prétraitements s'ils ne sont pas intégrés dans le programme principal.

Le TP est à remettre le vendredi 3 décembre 2020. Votre remise doit comprendre un rapport et les programmes que vous aurez développés pour ce TP, de même que des données modifiées ou nouvelles. La remise doit se faire par **turnin** sur https://turnin.dinf.usherbrooke.ca/.