Challenge 2 : apprentissage supervisé

Prévision de la qualité d'une solution d'un problème de transport

Contexte et but du challenge :

L'objectif de ce challenge est de comprendre et prédire la qualité d'une solution d'un problème de tournées de véhicules.

Petit rappel du problème :

« ***Soit un ensemble de véhicules dans un dépôt, et soit un ensemble de clients et leurs demandes, quel véhicule doit livrer quel client, et dans quel ordre, de façon à minimiser la distance totale parcourue par l'ensemble des véhicules ? »***

Le but n'est pas de résoudre le problème lui-même, un algorithme a déjà utilisé pour générer des solutions ! Des milliers de solutions par instance ! En effet, ce problème est un problème difficile et il n'existe pas d'algorithme parfait, qui va nous donner la solution en un temps polynomial en fonction de la taille de l'instance (voir la théorie des problèmes *NP*-complets). Pour obtenir des solutions approchées (parfois optimales, mais pas garanties), nous utilisons des métaheuristiques, familles de méthodes qui vont nous permettre d'obtenir des bonnes solutions.
Nous attaquons ici un problème de *Capacitated Vehicle Routing Problem* (CVRP) avec 100 clients.

- **Qu'est-ce qu'une instance ?** Une instance est une configuration du problème avec une répartition des clients sur une carte, la localisation du dépôt de départ, etc. Nous avons un fichier par instance, nommé par exemple XML100_2113_01.csv
- **Quel algorithme a été utilisé ?** Pour chaque instance, une variante d'algorithme génétique (une métaheuristique) a été appliquée pendant 10min. Ce programme génère ainsi plusieurs milliers voire dizaines de milliers (ou plus) de solutions par instance (par fichier csv).
- **Quelle est la configuration du problème ?** Nous avons environ 100 clients par instance, un nombre illimité de véhicules disponibles avec chacun une certaine capacité : chaque véhicule peut satisfaire 11 unités de demande. Un client cherche à être livré d'un produit.
- **Quel est l'objectif à minimiser ?** L'algorithme cherche à satisfaire la demande des clients tout en minimisant la distance parcourue par l'ensemble de la flotte de véhicules utilisés : chaque véhicule utilisé effectue une tournée de clients et revient au dépôt.
- **But du challenge ?** Nous ne cherchons pas du tout à utiliser du machine learning pour trouver de bonnes solutions (cela aurait été possible mais ici non), nous allons plutôt utiliser les méthodes vues en cours afin d'être capable de prédire la qualité d'une solution sans connaître son détail ni son coût, mais plutôt à partir de certaines caractéristiques liées à la solution (voir section description des données).

- **Pourquoi un tel but ?** Nous pensons qu'être capable de prédire et comprendre les déterminants d'une bonne ou mauvaise solution d'un problème VRP pourrait permettre à terme de mettre en place des méthodes hybrides couplant machine learning et optimisation. Ceci permettrait de tirer parti des expériences du passé (les historiques de solutions à partir desquelles les algorithmes de FML apprennent) afin de guider, et en particulier accélérer, la recherche de solutions du CVRP grâce à des algorithmes d'optimisation. Ceci est un problème de recherche actuel dont nous souhaitions vous faire profiter… Parce que vous le valez bien !

Description des données :

Le dossier à télécharger sur la page Moodle du cours contient 4 sous dossiers, nommés 2113, 2213, 3113 et 3213. Chacun des 4 sous-dossiers correspond à un sous-groupe d'instances, classé (au sens du clustering) selon les caractéristiques suivantes :

2113 : dépôt centré, position des clients totalement aléatoire
2213 : dépôt centré, mais les clients forment plusieurs clusters, grappes géographiques.
3113 : dépôt proche d'un coin de la carte, position des clients aléatoire
3213 : dépôt proche d'un coin, mais les clients forment plusieurs clusters, grappes géographiques.

Dans chaque sous-dossier, il y a respectivement, 27, 27, 26 et 26 instances différentes, pour un total de 106 instances. Chaque instance (chacun des 106 fichiers) contient des milliers de lignes. Chaque ligne correspond à une solution : chaque solution/ligne a été générée pendant l'application d'un algorithme génétique durant 10 minutes.

Explication des fichiers de données :

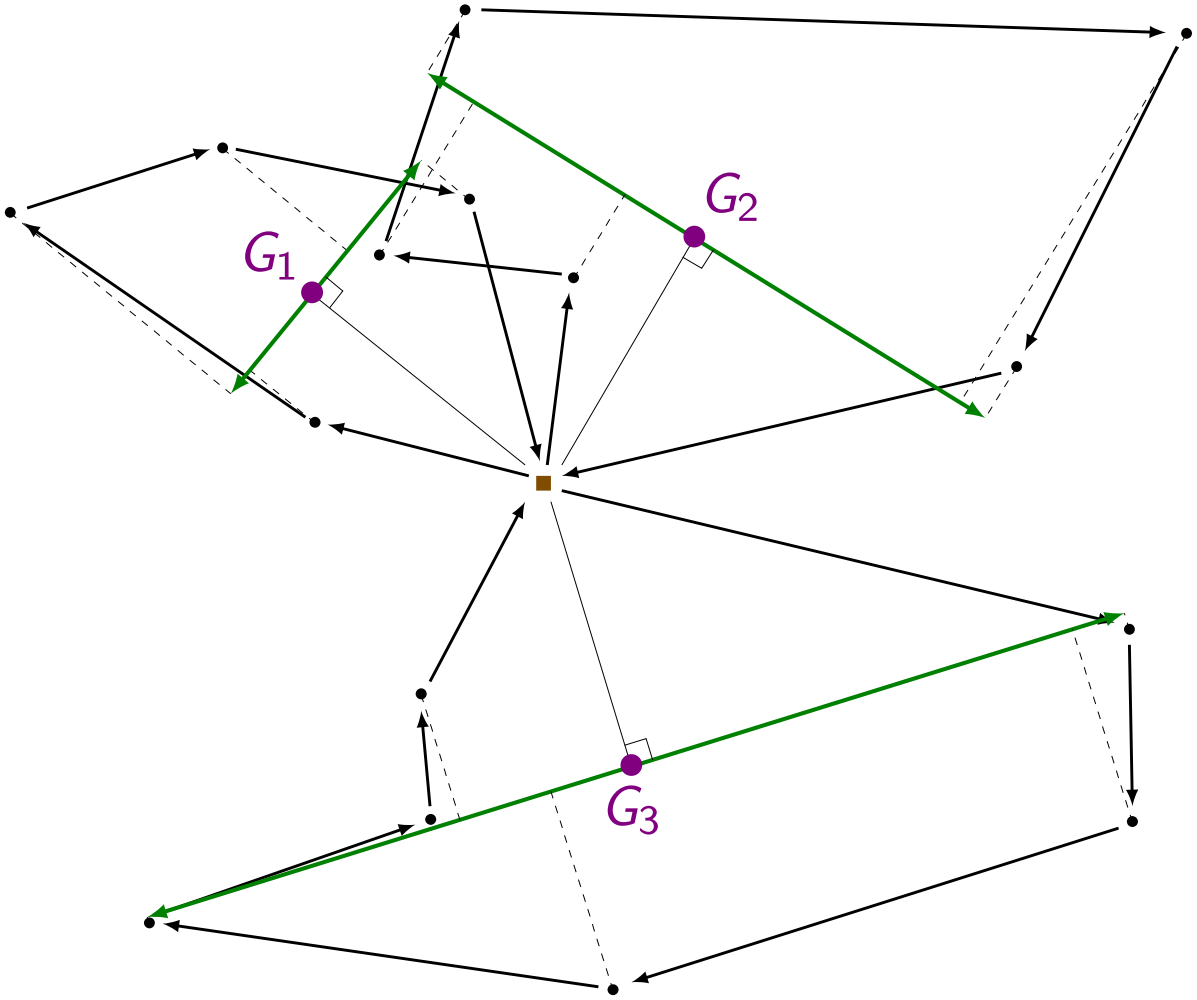
1 ligne = 1 solution
Colonne 1 : Nom de l'instance
Colonne 2 : Coût de la solution
Colonne 3 à 20 : les 18 caractéristiques/features des solutions (que l'on nomme/renommera S1...S18)

Note : Suite à un bug une des features n'a pas été bien calculée. A vous de la détecter et la supprimer.

S01 (colonne numéro 3) Largeur moyenne des tournées

Cette caractéristique est égale à la taille moyenne des arêtes vertes de la figure ci-dessous.

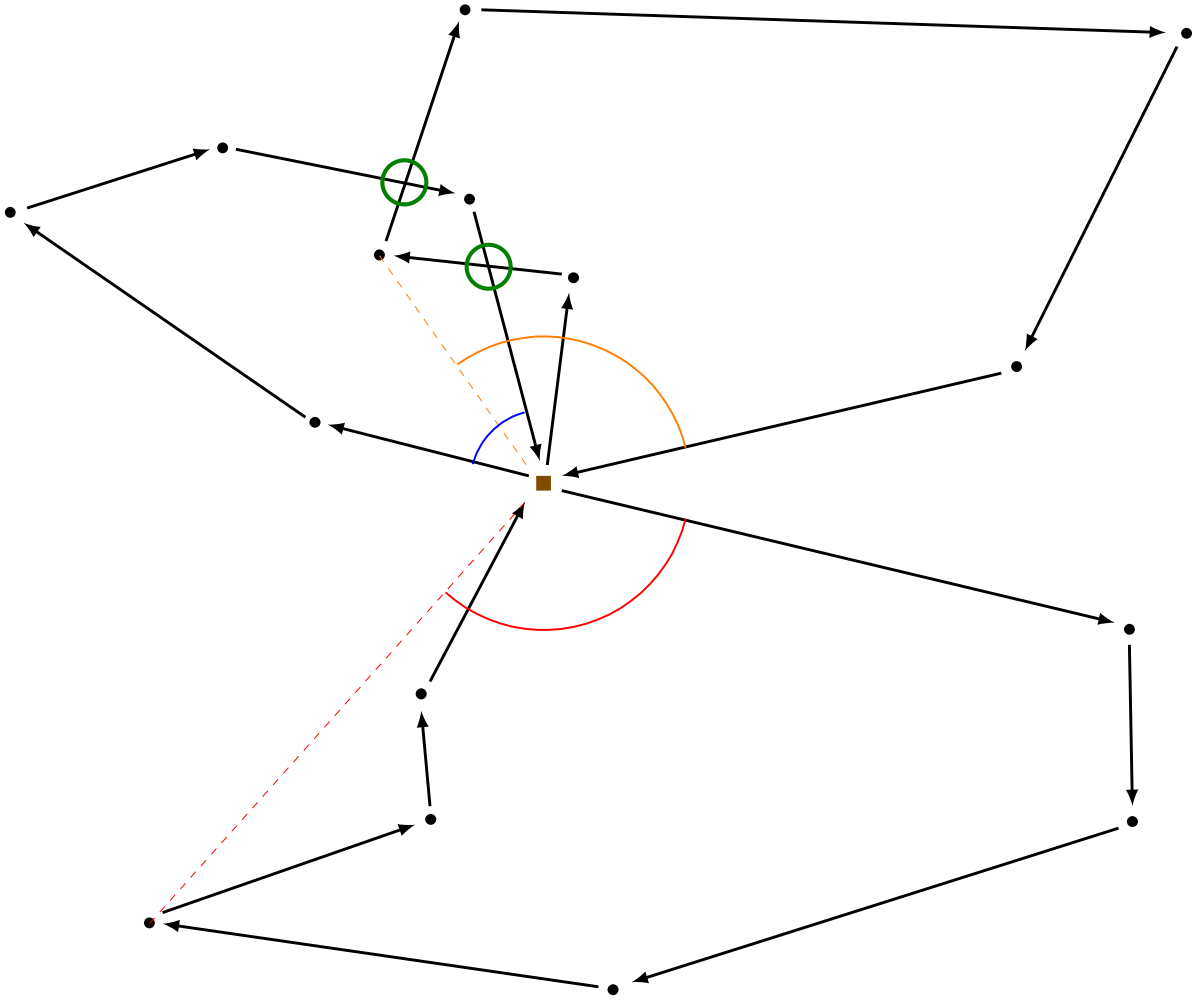
Celles-ci correspondent aux distances maximales de deux clients d'une même tournée, sur la projection sur l'axe perpendiculaire à l'axe dépôt-centre de gravité.



S02 Écart type sur la largeur des tournées

S03 Envergure moyenne des tournées

Caractéristique égale à la valeur moyenne des angles maximums obtenus entre deux clients d'une tournée, et le dépôt, représentés ci-dessous



S04 Écart type sur l'envergure des tournées

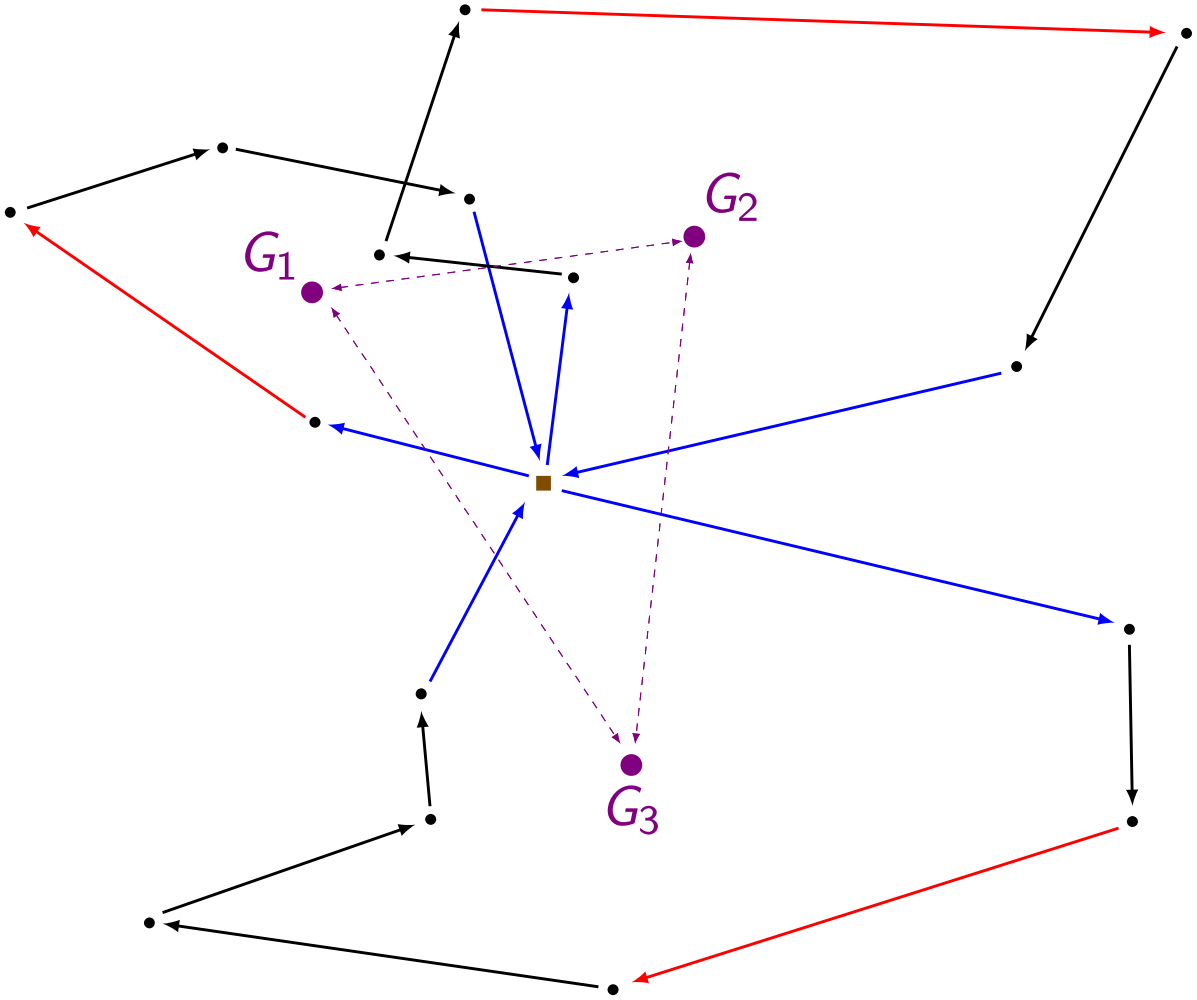
S05 Profondeur moyenne des tournées

Correspond à la distance moyenne par tournée entre le client le plus éloigné du dépôt et le dépôt.

S06 Écart-type sur la profondeur des tournées

S07 Longueur de la première et de la dernière arête de chaque tournée

arêtes bleues dans la figure ci-dessous, divisée par la longueur totale de la tournée.

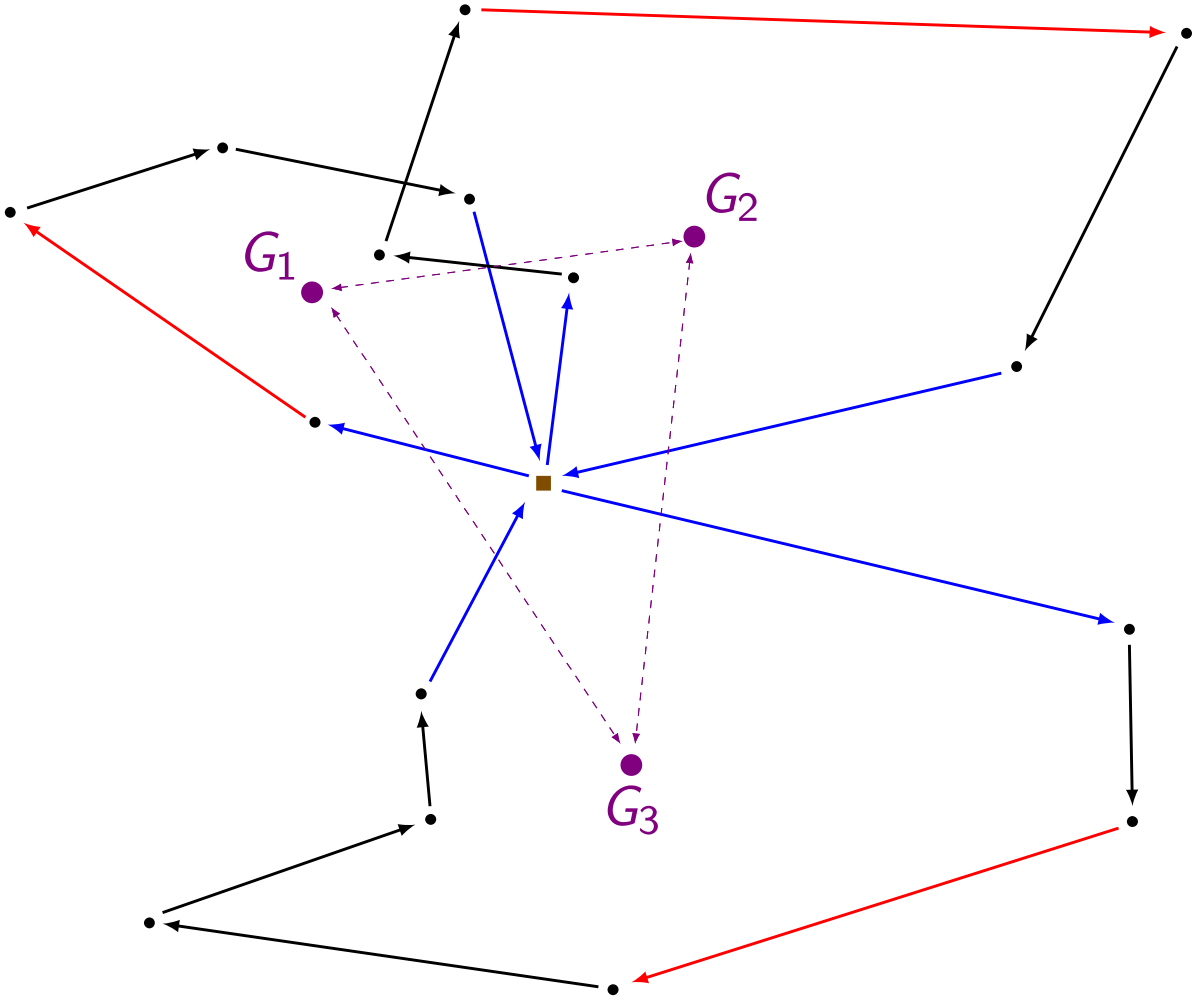


S08 Longueur moyenne de la plus grande arête de chaque tournée
S09 Longueur de la plus grande arête de chaque tournée, divisée par la longueur de la tournée

Cette caractéristique indique donc la proportion du temps de trajet utilisé pour le plus long déplacement "dépôt vers client", "client vers client" ou "client vers dépôt" de la tournée.

S10 Longueur de la plus grande arête intérieure de chaque tournée (arête non connectée au dépôt), divisée par la longueur de la tournée

Cette caractéristique indique donc la proportion du temps de trajet utilisé pour le plus long déplacement "client vers client" de la tournée, et est représentée par les arêtes rouge de la figure ci-dessous.



S11 Longueur moyenne de la première et de la dernière arête de chaque tournée
(arêtes bleues dans la figure précédente.
S12 Demande du premier et dernier client de chaque tournée, divisée par la charge du véhicule

Indique la proportion moyenne de charge du véhicule dû au premier et au dernier client.

S13 Demande du client le plus éloigné du dépôt, pour chaque tournée, divisée par la charge du véhicule

Indique la proportion moyenne de charge du véhicule dû au client éloigné

S14 Écart type sur la demande du client le plus éloigné du dépôt
S15 Écart type sur la longueur de chaque tournée
S16 Distance (euclidienne) moyenne entre les centre de gravité tournées
i.e : entre les moyenne de coordonnées de client + dépôt, indiqués par les points G1, G2 et G3 de la figure précédente.
S17 Écart type sur le nombre de client de chaque tournée
S18 Degré de chaque voisinage moyen des clients

Un client qui sera livré après son plus proche voisin et avant son 3eme plus proche voisin aura un degré de voisinage de 2.

Travail à effectuer

Le travail consiste à mettre en place une méthodologie complète d'apprentissage supervisé appliquant les méthodes vues en cours. Nous proposons de suivre les étapes suivantes et indiquons un barème indicatif pour chacune d'entre elles :

Statistiques descriptives et feature engineering (2 points)

1. Statistiques uni- et multi-dimensionnelles : évaluation de la qualité des données, compréhension de la structure, des liens entre variables
2. Recodage des variables, transformation, création éventuelle de nouvelles variables

Benchmark des méthodes de régression pour prédire le coût d'une solution d'un problème CVRP (3 points)

Le but de cette partie est de prédire la variable « coût », soit la deuxième colonne du jeu de données, en formulant donc le problème sous la forme d'une régression. Vous suivrez une méthodologie classique d'apprentissage supervisé (train/test), en appliquant différentes méthodes vues en cours :

- Régression logistique
- kNN Regression
- Support Vector Regression
- Arbres de régression et random forest
- Gradient Boosting
- Réseaux de neurones et deep learning

Il s'agit de proposer le meilleur modèle possible en termes de généralisation sur un ensemble de test, c'est-à-dire d'être capable de prédire le plus précisément possible le coût d'une solution, à partir des variables fournies, ou d'un sous-ensemble de variables sélectionnées. **Le choix de l'ensemble test et les subtilités d'évaluation vont revêtir une importance dans notre appréciation.**

Transformation en un problème de classification (3 points)

C'est la partie créative du challenge, qui départagera les meilleures équipes. Pour chaque instance (chaque fichier de départ), il est évident que la valeur de coût minimum correspond à la valeur de la meilleure solution connue. Il est alors envisageable de créer une nouvelle colonne qui peut prendre la forme d'une variable binaire (ex : bonne/mauvaise solution) ou catégorielle (ex : mauvaise/moyenne/bonne/excellente). Questions de seuil, que vous rencontrerez dans votre carrière d'ingénieur…. Vous transformez alors le problème de régression en un problème de classification et pourrez appliquer les méthodes vues en cours. Vous proposerez un modèle capable de prédire la qualité d'une nouvelle solution et l'évaluerez en généralisation, sur un ensemble test. Vous pourrez appliquer si besoin une méthode de rééquilibrage de classes et calculerez les métriques de performances habituelles (matrices de confusion, F1 score, etc).

Qualité du code et de l'analyse (2 points)

Une attention particulière sera portée sur la qualité de votre analyse, vos idées sur le problème proposé. Un notebook très bien commenté sera le rendu minimum.