# TD : Généralités

## 1 Exercice 1 : Preuve de la distance de Jaccard

Dans cette exercice, on se propose de finir la preuve que la distance de Jaccard est bien une distance. On s'intéresse en particulier à la preuve de l'inégalité triangulaire. Pour cela, l'idée est de réécrire la distance de Jaccard comme une distance entre des ensembles.

Considérons deux observations $x$ et $y$ de $K$ variables binaires. Notons $X = \{i \in \{1, \dots, K\} \mid x_i = 1\}$ et $Y = \{i \in \{1, \dots, K\} \mid y_i = 1\}$.

1. Faire un dessin montrant $M_{11}$, $M_{10}$ et $M_{01}$ à l'aide des ensembles $X$ et $Y$.
2. Écrire l'indice de Jaccard entre $x$ et $y$ à l'aide des ensembles $X$ et $Y$. En déduire la distance de Jaccard en fonction des ensembles $X$ et $Y$. On pourra noter $X \Delta Y = (X\ Y) \cup (B\ A)$.
3. Soit $A = \{1, 2\}$, $B = \{2, 3\}$ et $C = \{3, 4\}$. Calculer les distances de Jaccard entre $A$ et $B$, $B$ et $C$ et $A$ et $C$ en utilisant la définition ensembliste de cette distance.
4. Soit une troisième observation $z$, notons $Z = \{i \in \{1, \dots, K\} \mid z_i = 1\}$. Réécrire l'inégalité triangulaire pour la distance de Jaccard en utilisant la définition ensembliste.
5. Montrer que $X \Delta Y \subseteq (X \Delta Z) \cup (Y \Delta Z$.
6. Montrer que $X \cup Y \subseteq (X \cup Z) \cup (Y \cup Z)$.
7. Conclure sur l'inégalité triangulaire. On pourra supposer que, pour tous nombres positifs $a_1, a_2, b_1, b_2$, on a $\frac{a_1+a_2}{b_1+b_2} \leq \frac{a_1}{b_1} + \frac{a_2}{b_2}$.

## 2 Exercice 2 : Différent scénarios

Pour chacun des scénarios suivants, dire si c'est un problème de classification ou de régression et si c'est un problème d'inférence ou de prédiction, donner le nombre d'observations et l'ensemble mathématique dans lequel ces observations vivent, donner une distance possible pour comparer les observations.

1. On voudrait savoir si une personne hospitalisée au Québec à une chance de développer une complication respiratoire. Pour cela, on se rend dans 5 hôpitaux de la province et on demande à 100 patients dans chacun des hôpitaux s'ils ont eu une complication respiratoire, ainsi que leur âge, leur IMC et leur statut vaccinal.
2. On cherche à modéliser le prix de vente des maisons à Québec en fonction de leur surface, du nombres de chambres et de la présence d'un garage. Il y a 2000 ventes dans la base de données.
3. On s'intéresse à la concentration quotidienne de particules fines (PM2.5) dans l'air à Québec sur une année. De plus, on mesure la température, le vent et l'humidité.
4. Un sondage est mené auprès de 2000 résidents canadiens pour évaluer leur niveau de satisfaction vis-à-vis des services publics (santé, éducation, transport, etc.). On recueille les réponses sur une échelle de 1 (très insatisfait) à 5 (très satisfait), ainsi que des informations démographiques comme l'âge, la province de résidence, le revenu et le niveau d'éducation.
5. On veut identifier l'espèce d'un animal à partir de mesures morphologiques (longeur des pattes, poids, longueur du museau) collectées dans les parcs nationaux. On observe des castors, des ours noirs, des cerfs et des wapitis. En tout, on observe 1000 animaux.

## 3 Exercice 3 : Une mise en situation

Une chercheuse en sciences politiques souhaite étudier les caractéristiques de certains élus canadiens. Elle a monté une base de données avec 151 élus. Pour chaque élus, la base de données contient des variables qualitatives (genre, langue, parti politique, …) et ordinales (niveau de scolarité, tranche d'âge au moment de l'élection, …). Elle souhaite visualiser ses données pour mieux comprendre les caractéristiques des élus et les regrouper selon leurs caractéristiques.

Proposer un plan de travail : identifier toutes les étapes de l'analyse (en commençant par la définition de l'objectif) et associer un nombre d'heures approximatif à chaque étape.

## 4 Exercice 4 : Calcul de distance

1. Un robot commence à la position $(0, 0)$. Calculer la distance la plus courte que le robot doive parcourir pour aller à la position $(8, 6)$.
2. Un taxi New-Yorkais dois aller de l'angle de la 5th avenue et 42nd street (New York public library) à l'angle de la 1st avenue et 114th street (Thomas Jefferson park). Calculer la distance que le taxi va faire.
3. Une entreprise de transport utilise les coordonnées GPS pour calculer les distances entre son entrepôt et deux de ses clients. Les coordonnées de l'entrepôt sont $(10, 15)$, les coordonnées du premier client sont $(18, 22)$ et les coordonnées du deuxième client sont $(5, 8)$. Quel est le client le plus proche de l'entrepôt à vol d'oiseau. Est-ce le même client

qui est le plus proche de l'entrepôt pour un camion qui suivrait des routes, réparties en grille.

4. Considérons un système de communication qui envoie des messages encodés sur 4 bits. Ce système envoie les messages suivants : "1001" et "1101". Quelle est la distance de Hamming entre ces deux messages ? Si un système de détection d'erreurs peut détecter des erreurs de 2 bits, est-ce que la différence entre les deux messages est détectable ?
5. Trois utilisateurs ont aimé les films suivants : {Titanic, Avatar, Star Wars, Matrix, Inception}, {Avatar, Matrix, Batman, Superman, Inception} et {Inception, Avengers, Spiderman, Superman, Batman}. Quels sont les utilisateurs les plus proches basés sur la distance de Hamming ?