TD : GENERALITÉS

EXERCICE 2 : On note i le 1er obs et j le 2ème obs.

① Classification → complication respiratoire ou non.
Prédiction → une personne hospitalisée.
Nb obs → 5×100 patients.
Ω → $\underbrace{\mathbb{N}}_{\text{âge}} \times \underbrace{\mathbb{R}_+}_{\text{IMC}} \times \underbrace{\{oui, non\}}_{\text{statut vaccinal}} \times \underbrace{\{oui, non\}}_{\text{complication respiratoire}}$
Distance → $|age_i - age_j| + |IMC_i - IMC_j| + \mathbb{1}_{\{vaccin_i \neq vaccin_j\}} + \mathbb{1}_{\{resp_i \neq resp_j\}}$.

② Regression → prix de vente des maisons.
Inférence → "modéliser"
Nb obs → 2000
Ω → $\underbrace{\mathbb{R}_+}_{\text{surface}} \times \underbrace{\mathbb{N}}_{\text{nb chambres}} \times \underbrace{\{oui, non\}}_{\text{garage}} \times \underbrace{\mathbb{R}_+}_{\text{prix}}$
Distance → $|surf_i - surf_j| + |cham_i - cham_j| + \mathbb{1}_{\{gar_i \neq gar_j\}} + |prix_i - prix_j|$.

③ Regression → concentration quotidienne de PM 2.5.
Inférence → Pas d'indice sur le fait que l'on voudrait prédire.
Nb obs → 365 (sur une année)
Ω → $\mathbb{R}^4$
Distance → $\left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}$.

④ Classification $\longrightarrow$ réponses de 1 à 5.

Inférence $\longrightarrow$ compréhension de la satisfaction.

Nb obs $\longrightarrow$ 2000

$\Omega \longrightarrow \mathbb{N} \times \{QC, ON, \dots\} \times \mathbb{R}_+$
$\times \{\text{sans diplome, secondaire}, \dots\} \times \{1, \dots, 5\}$

Distance $\longrightarrow |age_i - age_j| + \mathbb{1}\{prov_i \neq prov_j\} + |rev_i - rev_j|$
$+ \mathbb{1}\{dipl_i \neq dipl_j\} + \mathbb{1}\{satif_i \neq satif_j\}$

⑤ Classification $\longrightarrow$ Types d'animaux.

Prédiction $\longrightarrow$ "identifier".

Nb obs $\longrightarrow$ 1000

$\Omega \longrightarrow \mathbb{R}_+^3 \times \{\text{castor, ours, cerf, wapitis}\}$

Distance $\longrightarrow \left[\sum_{i=1}^{n} (x_i - y_i)^2\right]^{1/2} + \mathbb{1}\{animal_i \neq animal_j\}$

EXERCICE 3 :

	En % du temps
① Définition de l'objectif	5 %
② Collecte des données	10 %
③ Nettoyage des données	50 %
④ Conception du modèle	10 %
⑤ Implémentation du modèle	5 %
⑥ Analyse des résultats.	20 %

EXERCICE 4 :

① La distance la plus courte est donnée par la distance euclidienne (distance en ligne droite)

$$d(\text{début}, \text{fin}) = \left[(8-0)^2+(6-0)^2\right]^{1/2} = 10$$

② La distance de Manhattan peut être utilisé pour calculer les distances dans NY (car les rues sont quadrillées).

$$d(\text{début}, \text{fin}) = |5-1| + |42-114| = 76.$$

③ Notons $X = \begin{pmatrix}10\\15\end{pmatrix}$, $Y = \begin{pmatrix}18\\22\end{pmatrix}$ et $Z = \begin{pmatrix}5\\8\end{pmatrix}$

- À vol d'oiseau (distance euclidienne) :

$$d(X,Y) = \left[(18-10)^2+(22-15)^2\right]^{1/2} = 10.63$$

$$d(X,Z) = \left[(5-10)^2+(8-15)^2\right]^{1/2} = 8.60$$

À vol d'oiseau, le client le plus proche de l'entrepôt est le 2ème.

- Sur une grille (distance de Manhattan)

$$d(X,Y) = |18-10| + |22-15| = 15$$

$$d(X,Z) = |5-10| + |8-15| = 12.$$

Sur une grille, le client le plus proche de l'entrepôt est le 2ème.

④ $d(\text{"1001"}, \text{"1101"}) = \sum_{i=1}^{4} \mathbb{1}_{\{x_i \neq y_i\}} = 1$

Le système de détection d'erreur ne peut pas détecter la différence entre les deux messages car il n'y a qu'un bit de différence.

(5) $d(u_1, u_2) = 1 + 0 + 1 + 1 + 0 + 1 + 1$

$= 5$

$d(u_1, u_3) = 1 + 1 + 1 + 1 + 0 + 1 + 1 + 1 + 1$

$= 8$

$d(u_2, u_3) = 1 + 1 + 0 + 0 + 0 + 1 + 1 + 0 + 0$

$= 4$

Les utilisateurs les plus proches sont les 2 et 3.