**TP : Introduction au Machine Learning**

**Partie I : ML par Renforcement (30 min)**

1. Quels sont les paramètres :
   1. du modèle ?

Shape (8 genes, 1 per vertex)

Wheel size (2 genes, 1 per wheel)

Wheel position (2 genes, 1 per wheel)

Wheel density (2 genes, 1 per wheel) darker wheels mean denser wheels

Chassis density (1 gene) darker body means denser chassis

* 1. de l’algorithme ?

|  |  |  |
| --- | --- | --- |
| Mutation rate: |  | |
| Mutation size: | |

Elite clones:

* 1. de l’environnement ?

Floor

Gravity

1. Quelle est la récompense ?

Top scores : Distance parcourue et hauteur

1. Quels sont les effets attendus des paramètres de l’algorithme sur le résultat

Augmentation top score mais pas jusqu’à convergence

Si augmente trop, modifications trop différentes, non optimisation

**Partie II : Régression linéaire (30 min)**

Dans un tableur, réaliser une régression linaire sur chaque jeu de donnée.

1. Quelle est l’équation de la régression linaire pour chaque série ?

0.5x+3

1. Quel est le coefficient de corrélation pour chaque série ?

0.67

1. Que remarque-t-on ? Que peut-on proposer ?

Toutes les courbes sont identiques bien que les données soient très différentes

1. Tracer la différence entre la prédiction et la valeur y de chaque série par rapport à x
2. Que peut-on remarquer

Courbe 1 aléatoire donc bon signe, courbe 2 suit une loi donc compléxifiable, courbe 3 valeur abbérante à traiter et courbe 4 no-sens

**Partie III : WEKA (30 min)**

1. Exploration des données :

* Combien y a-t-il d’exemples ?

3772

* Combien y a-t-il de classes ?

4

Cliquez sur les différents noms de la liste

* Combien y-a-t-il de descripteurs et quel est le type des données ?

30, label ou stat

* Ces données sont-elles complètes ? Quels problèmes cela pourraient poser ?

Non car on peut voir dans selected attribute que la partie missing n’est pas toujours nulle. Par exemple pour le renseignement du sex, 150 personnes n’ont pas répondu soit environ 4%. Cela peut biaiser les résultats statistiques ou des fois si missing l’algo peut ne pas marcher.

* Si l’on regarde la répartition des classes pour chaque descripteur que peut-on remarquer ?

Les descripteurs sont toujours en deux parties, les classes statistiques elles sont réparties en fonction de la valeur de distinct.

* Quels problèmes cela pourraient poser ?

Une comparaison entre classes plus difficile voir non-appropriée

* Si je veux réaliser une validation croisée avec 10-fold, combien d’exemple d’entrainement ai-je dans chaque fold ?

10-fold : on divise par 10 Full training set donc la totalité ?

Classe 4 on divise par 10 alors que y’en a que 2 donc pb

1. Visualisation :

Aller dans l’onglet « Visualize », et modifier la sélection des descripteurs, la taille des points