Équipe 3 – Notre implémentation de la régression linéaire

Ce document vise à brièvement décrire la méthode et l'idée algorithmique que nous avons mis en place vis-à-vis de l'ajout de régression linéaire à nos graphiques en « nuages de points », dans le cadre du récent projet pour la SAÉ combinée s2.04/s2.05.

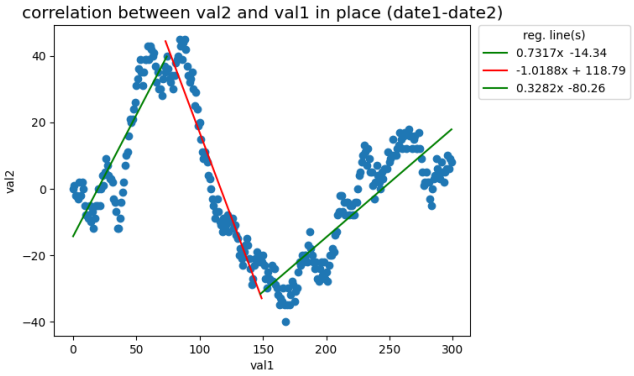
Premièrement, il est certain qu'utiliser une simple régression linéaire sur un ensemble de données ne permettrait que rarement d'obtenir un résultat satisfaisant. Par exemple, si l'ensemble présente une tendance à la forte croissance dans un premier temps, puis une forte chute, une droite de régression sur tout l'ensemble aurait un coefficient directeur proche de 0 (ce qui donnerait une « ligne quasi parallèle à l'axe des abscisses »). Il vaut mieux alors diviser l'ensemble en deux sous-ensembles et tracer les deux droites de régression linéaires correspondant mieux aux tendances des données.

Ainsi, nous avons décidé d'appliquer un raisonnement récursif à cette création de régression linéaire. Voici donc l'algorithme utilisé :

- Lors de l'appel sur un ensemble de données (en réalité, une paire d'ensembles de même taille, appelons-les x et y respectivement), la droite de régression et le coefficient de régression de cet ensemble est calculé.

> Si le coefficient est considéré « suffisant » *(supérieur à une certaine valeur déterminée selon la profondeur actuelle de récursion, de telle sorte que l'on soit « plus exigeant » au début, et de moins en moins à mesure que l'on est sur des ensembles plus découpés)*, ou bien qu'il est de « trop petite taille », l'on trace la droite calculée et arrête là.

> Sinon, on divise x et y chacun en deux autres sous-ensembles (x1, x2, y1, y2) contenant respectivement la première ou la deuxième moitié de l'ensemble parent. L'on rappelle alors la fonction récursive pour (x1, y1) et (x2, y2), avec une profondeur augmentée de 1.



Exemples du rendu de l'application de cette méthode sur des données fictives générées aléatoirement, de tailles différentes et de profils plus ou moins différents.

Comme il est ici visible, nous avons rajouté, pour aider à la lecture, une couleur verte aux droites de coefficient directeur positif, et une couleur rouge aux droites de coefficient directeur négatif

De plus, les équation approximatives des droites tracées sont mises en légende, avec une correspondance entre l'ordre gauche → droite des lignes du graphe et celui haut → bas des équations.