GGplot2

Benjamin GUIGON

January 2021

1 Lien

Article écrit par Jiayue LIU, Soukaina ELGHALDY

https://github.com/liu-jiayue/psbx

2 Introduction

GGplot2 est la librairy d'affichage de graphe la plus connue et la plus utilisée dans le monde de la data.

L'une des principales fonctionnalités de GGplot2 va être le travail des données avant d'appliquer un modèle de machine learning, c'est le pré-processing (*exemple : as.factor* ou *ggplot*).

On peut aussi retrouver des fonctions de regressions qui permettent d'avoir des a priori sur le dataset en utilisant *geom_smooth(method=lm)* ou *lm* veut dire linear model mais on peut aussi retrouver une méthode non linéaire avec *Loess*.

Il existe ensuite toute une panoplie de commande pour gerer les nuages de points avec les intervalles de confiance, les droites de regressions par nuages, les formes/tailles/couleurs des points de chaque nuage.

Un nuage de point peut avoir des tendances qu'on ne voit pas à l'oeil nu. L'ajout de densité, permet de reporter directement sur les axes les densités de point. Sur un nuage très isotrope, on pourrait voir se dessiner une constance dans l'écart des points qui ne serait pas évidente sans ces marques.

Toutes ces commandes sont assez basiques, je vais m'attarder sur les bouts de code qui utilise la densité en 2 dimension avec les regroupement par ellipses. En utilisant *geom = polygon* on peut arriver à partimenter l'espace un peu comme une carte avec des reliefs. On peut aussi utilliser des ellipses pour créer des cluster de point grace à la commande *stat_ellipse()*, on peut évidemment

modulé cette fonctionnalité pour changer les formes/tailles/couleurs pour une meilleure ergonomie.

3 Avis

Le travail est très complet et constitu une sorte de glossaire pour GGplot2. C'est assez pratique, car souvent on ne connait pas forcément par coeur les fonctionnes d'affichages et avoir toutes ces fonctions résumées dans un seul document représente un super raccourci.

De plus l'article est didactique, il expliquer exactement toutes les lignes de codes sans laisser des flous. Il y a des affichages basiques ainsi que des affichages plus axés pour la data science avec les modèles linéaires et les densités 2D.