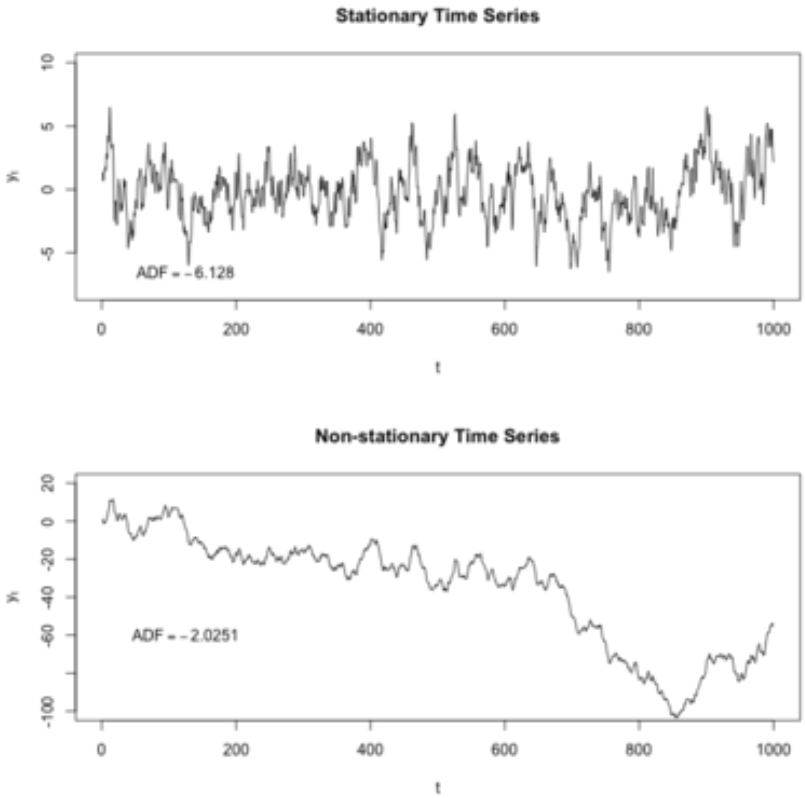
Préparation de données

Stationnarité

Dans le cas de série temporelle, il est possible que les données dépendent du temps, c'est à dire que l'évolution n'est pas stable. Prenons notre exemple de températures moyennes, est ce que la température sur 10 ans évolue de la même façon au XVIIIeme siècle et et XXeme?

Or dans le cas de Machine Learning et de prédiction, nous avons besoin de nous assurer que notre algorithme peut aussi bien prédire la température en 1850 qu'en 2030 mais aussi en 2100. On a donc besoin que nos données soient stationnaires, c'est à dire sans tendance mais aussi sans effet saisonal comme par exemple les températures hivernales. Pour ce faire on peut utiliser le test statistique **adfuller** qui va tester la stationnarité de nos séries temporelles. Si la p-value est inférieur à 0.05 la série est considéré comme stationnaire autrement elle est non stationnaire et nous devons la rendre stationnaire.

Quand une série temporelle est non stationnaire il y a plusieurs façon de la transformer tout en gardant nos la variance de nos données. L'une des façon la plus simple est de faire la différence entre la valeur n et n+1. Cela permet non plus d'avoir une série temporelle en valeur absolue mais d'avoir l'évolution au cours du temps. Il est aussi possible de faire une transformation logarithmique, ou exponentielle dépendant des données



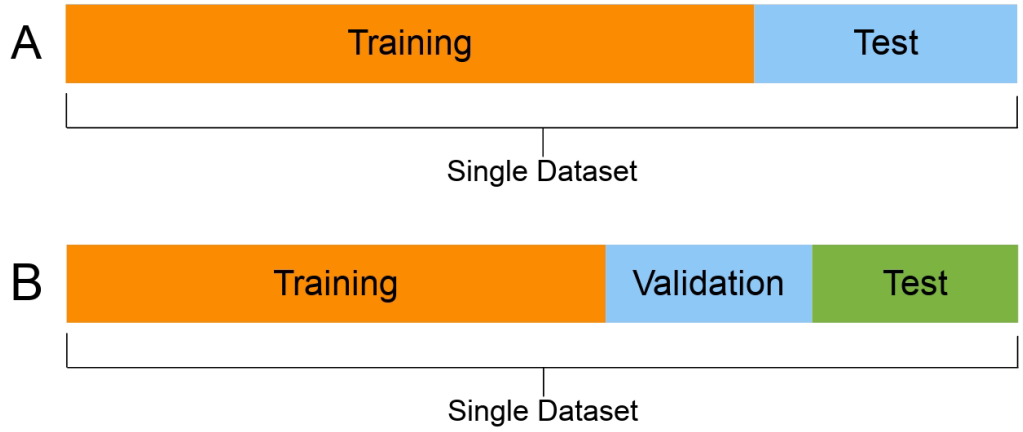
Split le data set

Introduction

Lorsqu'on fait du Machine learning il est indispensable d'entrainer nos algorithmes sur une partie de notre data set et de tester nos prédictions sur une autre partie dites naïve. On a donc un data set total séparer en **training set** et **test set**, en général on considère que le training set représente 70% du data set total. Cela dépend biensur de la taille du data set et peut être supérieur mais rarement inférieur. Il est également parfois possible d'avoir un ***validation set** pour tester nos paramètres, nous en reparlerons plus tard.

Comment faire le split

Il est indispensable de garder une certaine continuité de variance entre le training et test set. C'est à dire qu'on ne peut pas séparer le data set de nimporte quel façon. Prenons notre exemple concret, si l'on veut prédire la température dans les 30 prochaines années, si l'on coupe notre data set à 70% cela veut dire jusqu'en 1932 pour le training set. Certes cela est après la révolution industrielle mais avant la seconde guerre mondiale et l'accélération du réchauffement climatique. On aura donc un training et un test set qui seront très différent et donc **biaisé**.



Boostraping

Il est donc indispensable de "mélanger" le data set grâce au bootstraping, cette méthode permet de mélanger aléatoirement les données. L'aléatoire permet de s'assurer de ne pas avoir de biais conscient mais il peut toujours subsister des biais aléatoires, notamement dans de petits data set comme le notre. On peut donc faire plusieurs folds.