Scikit-learn

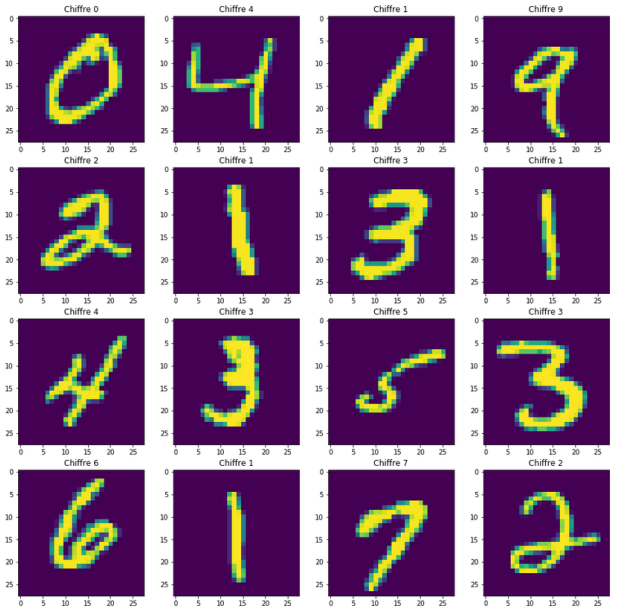
Cours M2 CCN

Modélisation des problèmes scientifiques

2019/2020

Pourquoi l’utiliser ?

- Librairie dédiée à l’apprentissage statistique
- « L'apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité **d' « apprendre » à partir de données**, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. »
- Exemple ;



Attribuer à chaque image le chiffre qui y est représenté. Une tâche sans intérêt particulier pour un humain, qui peut être automatisée.

On construit un modèle de prédiction à l’aide d’une partie des données dont on dispose, **les données d’entrainement** . Puis on teste le modèle sur le reste des données, les **données de test**.

On va jute s’intéresser à une méthode, la régression linéaire. Le reste, vous le verrez dans l’UE d’ intelligence artificielle avec Élisa Fromont!

Régression linéaire

- Établir un lien linéaire entre deux variables quantitatives x et y. On cherche à prédire les valeurs de y à l'aide des valeurs de x.

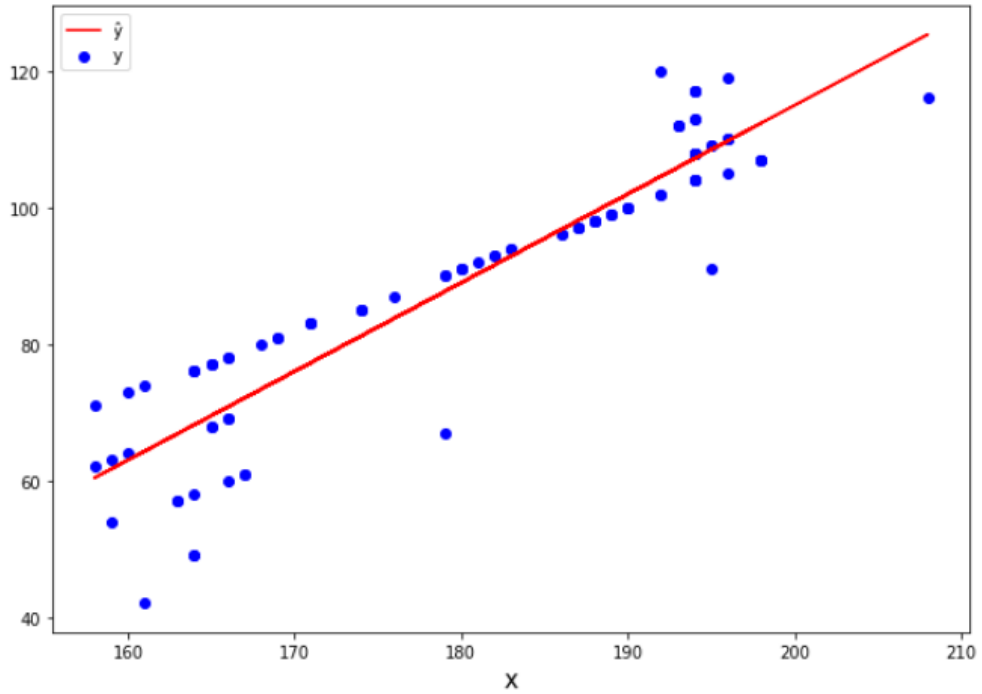
En notant n le nombre d'observations, on veut trouver a_{min} et b_{min} deux réels tels que :

$$(a_{min}, b_{min}) = argmin_{(a,b)} | \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 |$$

$$= argmin_{(a,b)} | \sum_{i=1}^{n} (y_i - (a * x_i + b))^2 |$$

a_{min} et b_{min} sont alors respectivement le coefficient directeur et l'ordonnée à l'origine de la droite de régression.

Graphiquement, on essaye de trouver la droite ŷ qui capture le mieux la tendance du nuage de points bleus y (ici, la doite rouge, correspondant à $\hat{y} = a_{min} * x + b_{min}$)

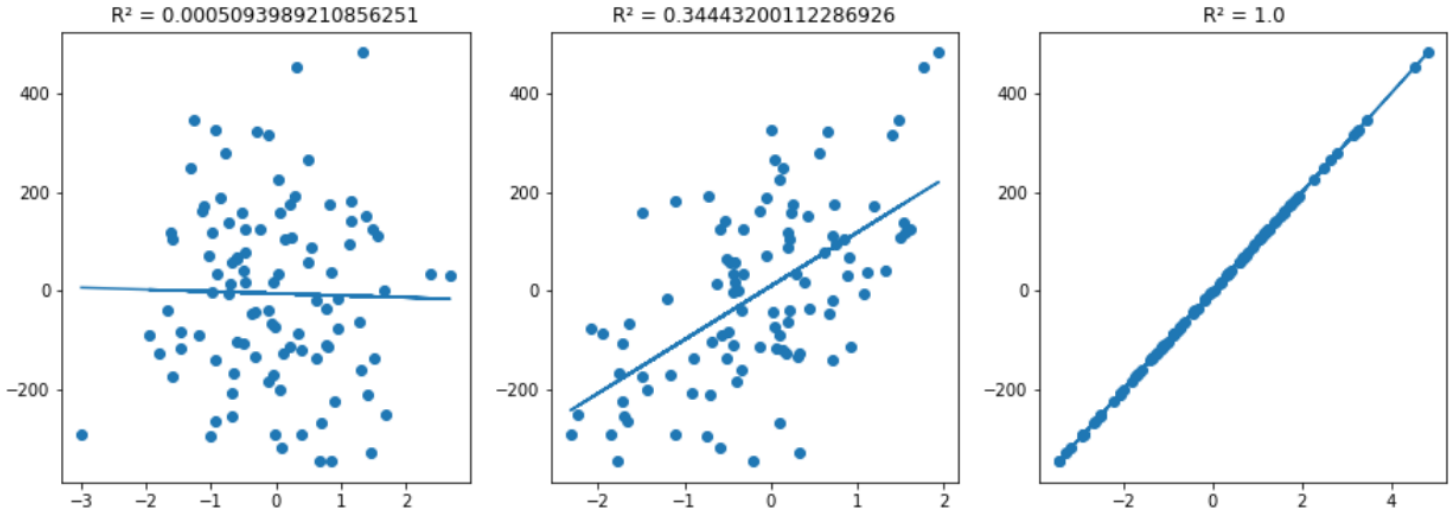


Régression linéaire (2)

- On peut mesurer la qualité d'ajustement de la droite aux données à l'aide du R^2

  Le R^2 est un indicateur compris entre 0 et 1. Plus il est proche de 1, plus la droite de régression capte la variance du nuage de points.

Pour les formules, voir le notebook méthodes statistiques

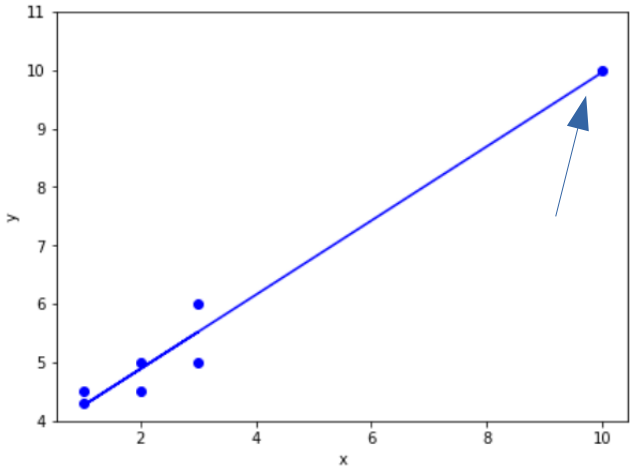


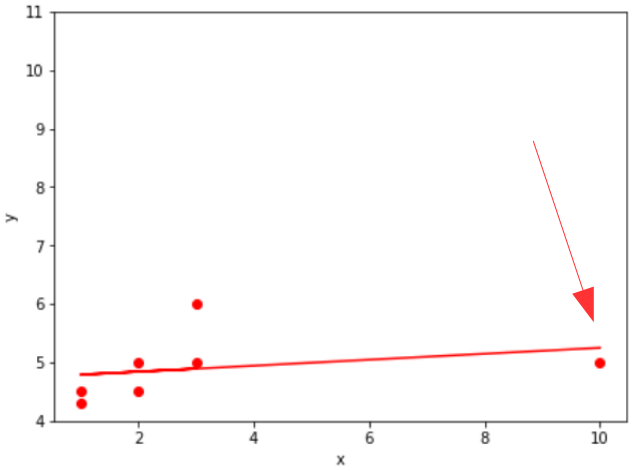
Régression linéaire (3)

Quelques pièges à éviter :

- Les points leviers

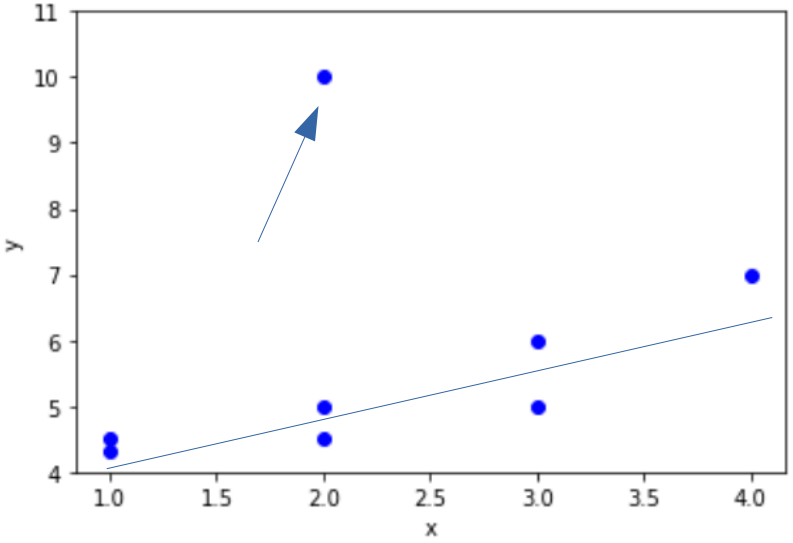
Un point qui fait varier énormément le coefficient de la droite si on l'enlève ou on le change





- Les points aberrants ou outliers

Un point qui ne correspond pas à la tendance générale et fait chuter le R^2 du modèle



Régression linéaire (4)

- Si on veut expliquer y avec deux variables quantitatives x1 et x2, on peut voir la régression comme un produit matriciel

- $Y = BX$

  avec $Y = (y_i)$, $1<i<n$,

  $B = (b_i)$, $1<i<n$,

  $X = (x1_i, x2_i)$, $1<i<n$

On peut généraliser avec p variables…

$B = X'(X'X)^{-1}Xy$

On projette y sur l'espace engendré par les données X

Pour aller plus loin

- http://fermin.perso.math.cnrs.fr/Files/Chap3.pdf Cours de régression linéaire
- Tests sur les coefficients (Student, Fisher)
- Régression par morceaux
- Régression pénalisée ; LASSO, Ridge, ElasticNet
- Régression logistique (linéaire + composition par la fonction logistique)
- Construction de modèle ; forward, backward, stepwise avec critère de sélection de variables (AIC, BIC)
- https://scikit-learn.org/stable/ La doc
- UE machine learning → sujet très large ;-)