Chapitre II : Analyse Prédictive : (Prédire une quantité : Régression)

II- 1- Régression Linéaire:

contoso

Analyse Prédictive :

consiste a Analyser les données actuelles afin de :
1-faire des hypothèses sur des comportements futurs des individus déjà présents
2-mais aussi sur de nouveaux individus.

Régression Linéaire

La **régression linéaire** est l'un des algorithmes les plus connus et les mieux compris en:

-statistique et

-en apprentissage automatique (Machine Learning).

Définition :

Le Machine Learning Apprentissage Automatique : est un ensemble de techniques puissantes permettant de créer des modèles **prédictifs** à partir de données, **sans avoir été explicitement programmées**.

C'est un domaine au croisement des mathématiques et de l'informatique.

Objectif du Machine Learning

Est de **trouver un modèle** qui effectue une **approximation** de la réalité à l'aide de laquelle on va pouvoir effectuer des prédictions.

La R**égression Linéaire Simple:**

-Il ya deux variables quantitatives X,Y.
-X variable explicative; Y Expliquée
➔**hypothèse** = les données proviennent d'un phénomène qui à la forme d'une **droite**.
-Il existe une relation linéaire entre l' entrée X et la sortie Y

Régression Linéaire

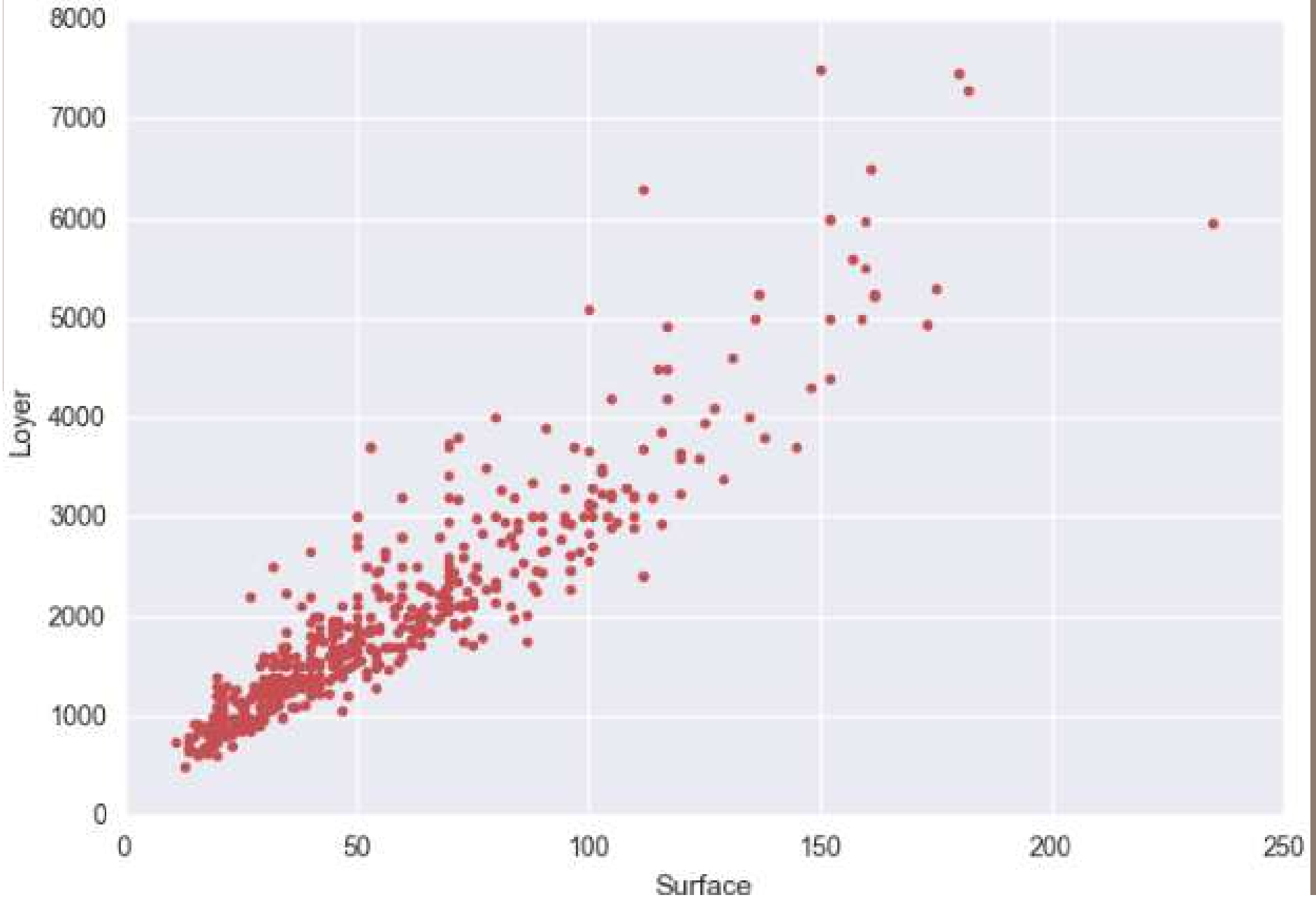
* Lorsqu'il existe une seule variable d'entrée (x), ➔ la méthode est appelée régression linéaire simple.

* Lorsqu'il y a plusieurs variables d'entrée, ➔ la méthode est appelée régression linéaire multiple.

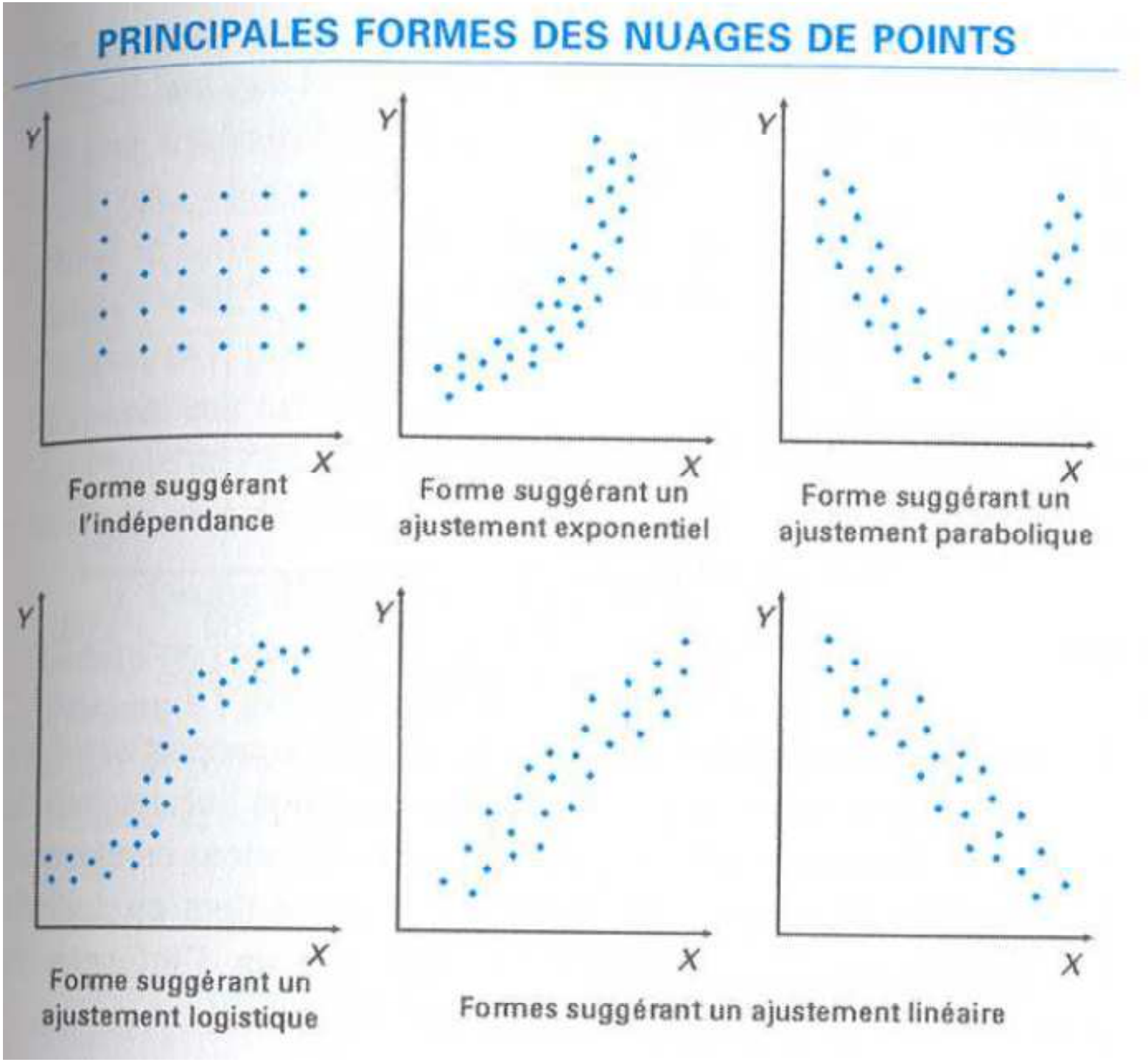
➔Un **graphique de corrélation permet de vérifier rapidement l'existence d'un lien.**

➔La forme du **nuage de points** obtenus détermine la nature de la Liaison statistique entre deux variables.

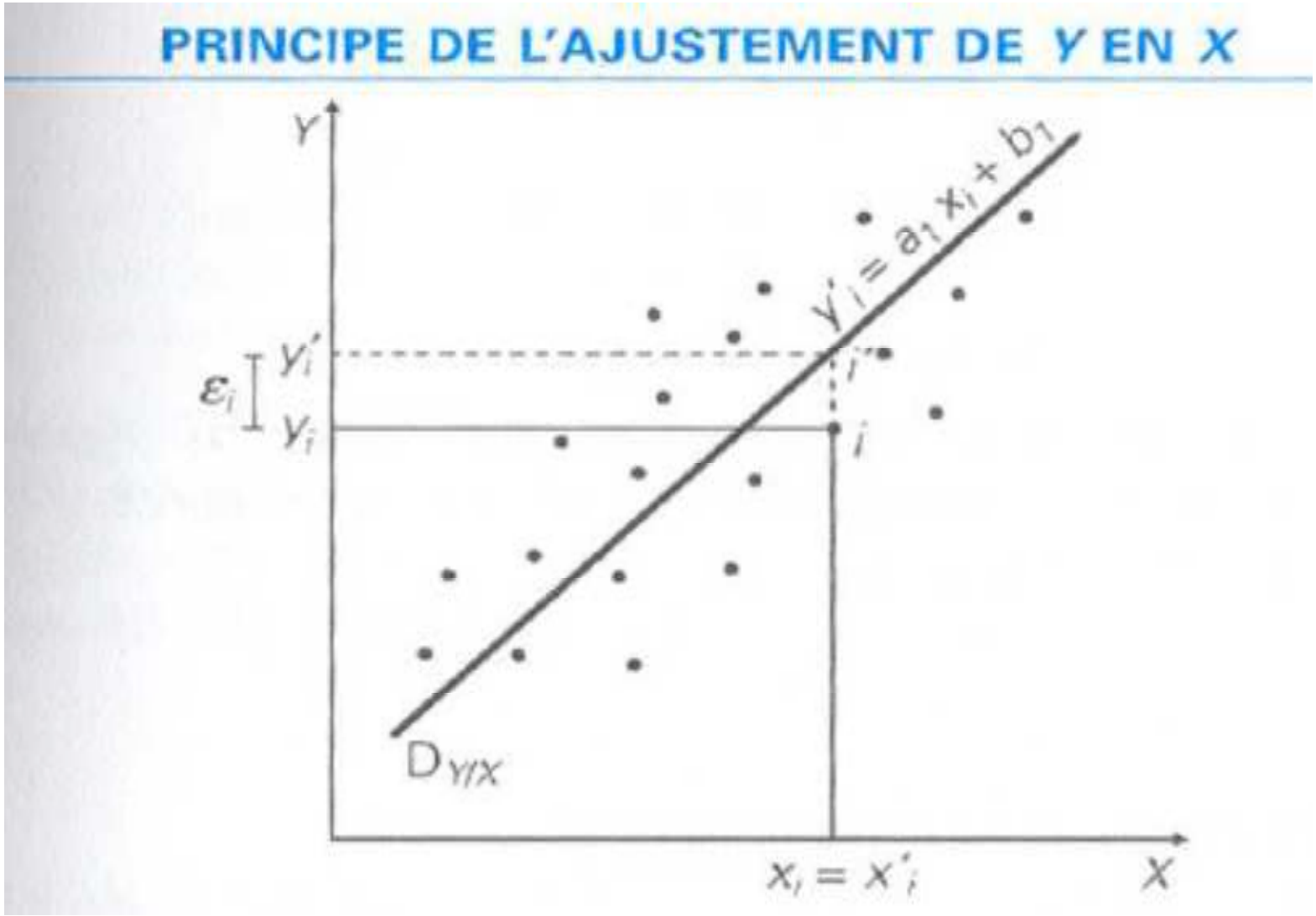
Régression Linéaire Simple



PRINCIPALES FORMES DES NUAGES DE POINTS



Régression Linéaire Simple



Régression Linéaire

Différentes techniques permettent de préparer ou d'entraîner l'équation de régression linéaire à partir de données, la plus courante étant appelée

➔ **Méthode des Moindres Carrés Ordinaires. MMCO**

Principe de MMCO:

-On calcule la distance entre chaque point des données et la ligne de régression.

-Cette opération engendre **des résidus ei par rapport à Y ,** Les valeurs de x restent inchangées ;

nous calculons la distance et la somme de toutes les erreurs au carré: $\sum(ei)^2$

C'est la quantité que les moindres carrés ordinaires cherchent à minimiser.

Min ($\sum(ei)^2$)

Représentation du modèle de régression linéaire

La représentation de la régression linéaire est une équation linéaire qui combine un ensemble de valeurs d'entrée (x) à la sortie (y) prévue.

Exemple : $Y = a1x + a2$

$$Y = a1x1 + a2x2 + a3$$

a1=facteur d'echelle; a3=biais

En tant que telles, les valeurs d'entrée (x) et valeur de sortie (Y)sont **Quantitatives**.

contoso

Représentation du modèle de régression linéaire

Il est courant de parler de la **complexité** d'un modèle de régression comme la régression linéaire == **Nombre de coefficients** utilisés dans le modèle.

Régression linéaire Apprentissage du modèle

Apprendre un modèle de régression linéaire signifie = estimer les valeurs des coefficients utilisés dans la représentation avec les données disponibles.

Régression linéaire Apprentissage du modèle

1. **Régression linéaire simple**

Avec une régression linéaire simple lorsque nous avons une seule entrée, nous pouvons utiliser des statistiques pour estimer les coefficients.

(moyenne, écart type, corrélations et covariance.

C'est un exercice amusant, mais pas vraiment utile dans la pratique.

contoso

Régression linéaire Apprentissage du modèle

$$a_1 = \frac{COV_{xy}}{V_x} = \frac{1/n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{1/n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\bar{y} = a_1 \bar{x} + b_1 \Leftrightarrow b_1 = \bar{y} - a_1 \bar{x}$$

2. **Méthode Moindres Carrés Ordinaires:**

Lorsque nous avons plus d'une entrée, nous pouvons utiliser les moindres carrés ordinaires pour estimer les valeurs des coefficients.

La procédure des moindres carrés ordinaires cherche à minimiser la somme des résidus au carré. $\sum(ei)^2$

Régression linéaire Apprentissage du modèle

Définition : un Minimum d'une fonction de plusieurs variables se produit au point ou les dérivées partielles par rapport a ses inconnus (a,b) s'annulent.

On peut démontrer que cette condition est vérifiée si le coefficient directeur de la droite vaut :

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Régression linéaire Apprentissage du modèle

Cette approche traite les données sous forme de matrice et utilise des opérations **d'Algèbre Linéaire** pour estimer les valeurs optimales des coefficients.

Toutes les données doivent être disponibles et il faut disposer de suffisamment de mémoire pour les adapter et effectuer des opérations de matrice.

3. Descente de Gradient

Lorsqu'il y a une ou plusieurs entrées, vous pouvez utiliser un processus d'optimisation des valeurs des coefficients en réduisant de manière itérative l'erreur du modèle.

Cette opération s'appelle Gradient Descent et commence par des valeurs aléatoires pour chaque coefficient.

Régression linéaire Apprentissage du modèle

En pratique, cette méthode est utile lorsque vous avez un très grand ensemble de données, que ce soit en nombre de lignes ou en nombre de colonnes qui peuvent ne pas tenir dans la mémoire.

Le calcul de l'inverse d'une matrice prend beaucoup de temps.

Algorithme descent de Gradient

Début

Iteration t=1

η : le pas (stepsize)

While not converged

W(t+1)⟵w(t) - η(dg(w)/dw)

T⟵t+1

Fin

Régression linéaire Apprentissage du modèle

Algorithme Descent de Gradient:

Un taux d'apprentissage η est utilisé comme facteur d'échelle et les coefficients sont mis à jour dans le sens d'une minimisation de l'erreur.

Le processus est répété jusqu'à ce qu'une erreur de somme au carré soit atteinte ou qu'aucune amélioration supplémentaire ne soit possible.

contoso

Régularisation

Il existe des extensions de la formation du modèle linéaire appelées méthodes de régularisation.

Celles-ci cherchent à la fois à **minimiser** la somme de l'erreur au carré **++** à **réduire** la complexité du modèle (comme le nombre ou la taille absolue de la somme de tous les coefficients du modèle). .

Régression linéaire Apprentissage du modèle

Régularisation

Deux exemples populaires de procédures de régularisation pour la régression linéaire:

- **Lasso Regression**: les moindres carrés ordinaires sont modifiés pour minimiser également la somme absolue des coefficients (appelée régularisation L1).

Régularisation

- Ridge Regression : les moindres carrés ordinaires sont modifiés pour minimiser également la somme absolue au carré des coefficients (appelée régularisation L2).

Régularisation

Ces méthodes sont efficaces lorsqu'il existe une **colinéarité** dans vos valeurs d'entrée et que des moindres carrés ordinaires surchargeraient les données d'apprentissage.

Faire des prédictions avec la régression linéaire

Faire des prédictions:

Étant donné que la représentation est une équation linéaire, faire des prédictions est aussi simple que de résoudre l'équation pour un ensemble spécifique d'entrées.

Faire des prédictions avec la régression linéaire

Exemple:

Imaginons que nous prédisons le poids (y) à partir de la taille (x).

La représentation du modèle de régression linéaire pour ce problème serait: $y = B0 + B1 * x1$

ou

poids = B0 + B1 * hauteur

Faire des prédictions avec la régression linéaire

Où B0 est le coefficient de biais et B1 est le coefficient de la colonne de hauteur.

Par exemple, utilisons B0 = 0.1 et B1 = 0.5.
- le poids (kg) d'une personne d'une hauteur de 182 centimètres.

poids = 0,1 + 0,05 * 182

poids = 91,1

Préparation des données pour la régression linéaire

En pratique, il faut **préparer les données** avant d'appliquer MMCO afin que les prédictions soient corrects.

On peut utiliser ces règles davantage comme règles empiriques .

Essayez différentes préparations de données en utilisant ces heuristiques et voyez ce qui convient le mieux à votre problème.

Préparation des données pour la régression linéaire

- **Supprimer le bruit**: La régression linéaire suppose que vos variables d'entrée et de sortie ne sont pas bruyantes.

Pensez à utiliser des opérations de nettoyage des. Ceci est très important pour la variable de sortie et vous souhaitez supprimer les valeurs aberrantes dans la variable de sortie (y) si possible.

Faire des prédictions avec la régression linéaire

- **Supprimer la colinéarité.** La régression linéaire surchargera vos données lorsque vous avez des variables d'entrée hautement corrélées.

Pensez à calculer des corrélations par paires pour vos données d'entrée et à supprimer les plus corrélées.