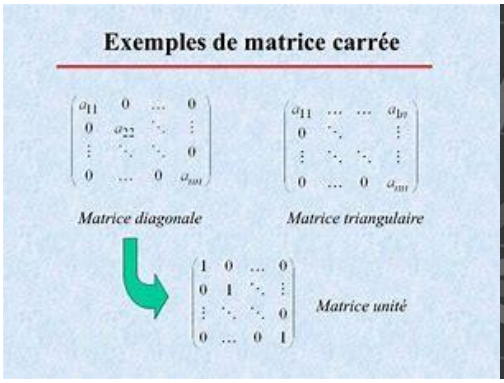
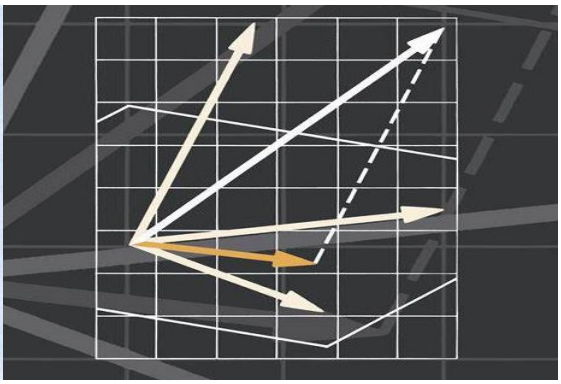
Chapitre IV Outil mathématiques utilisé en Data Science : Algèbre Linéaire





I. Les outils mathématiques utiles en Sciences de données

En sciences de données, une vaste gamme d'outils mathématiques est utilisée pour analyser, modéliser et interpréter des données. Voici une liste des principaux domaines et concepts mathématiques utiles en sciences de données :

1. Statistiques et Probabilités

Ces outils sont essentiels pour comprendre les données, en modéliser les distributions et prendre des décisions fondées sur l'incertitude.

- **Statistiques descriptives** : Moyenne, médiane, variance, écart-type, quantiles, corrélation.
- **Statistiques inférentielles** : Tests d'hypothèses, intervalles de confiance, p-value.
- **Probabilité** : Espaces probabilistes, lois de probabilité (normale, binomiale, Poisson, etc.).
- **Modèles probabilistes** : Régression logistique, réseaux bayésiens.
- **Estimation** : Maximum de vraisemblance, estimation bayésienne.

2. Algèbre linéaire

L'algèbre linéaire est la pierre angulaire de nombreux algorithmes de machine learning et de traitement de données.

- **Vecteurs et matrices** : Manipulation des données multidimensionnelles.
- **Décompositions matricielles** : Décomposition en valeurs propres (SVD, PCA).
- **Espaces vectoriels** : Dimensions, bases, sous-espaces.
- **Transformations linéaires** : Rotation, projection, changement de base.
- **Produits scalaires et tensoriels** : Utilisés dans les réseaux de neurones et les calculs complexes.

3. Calcul différentiel et intégral

Essentiel pour comprendre les algorithmes d'optimisation et les changements continus.

- **Dérivées et gradients** : Optimisation des fonctions de coût, descente de gradient.
- **Différentiation multivariée** : Calcul de Jacobienne et Hessienne.
- **Intégrales** : Modélisation des probabilités continues, calcul des aires et des volumes.
- **Équations différentielles** : Modélisation de systèmes dynamiques.

4. Théorie des graphes

Utilisée pour modéliser et analyser des relations dans des données complexes.

- **Représentation des graphes** : Listes d'adjacence, matrices d'adjacence.
- **Propriétés des graphes** : Connexité, centralité, degrés.
- **Algorithmes de graphes** : Dijkstra, PageRank, recherche en profondeur/largeur.
- **Applications** : Réseaux sociaux, détection de communautés, graphe de connaissances.

5. Optimisation

Utile pour ajuster les modèles et améliorer leur performance.

- **Optimisation convexe** : Fonctions convexes, conditions de Karush-Kuhn-Tucker (KKT).
- **Algorithmes d'optimisation** : Descente de gradient (et ses variantes comme Adam, RMSProp), programmation quadratique, méthode des moindres carrés.
- **Problèmes de contrainte** : Lagrangiens, dualité.

6. Théorie de l'information

Elle permet de mesurer l'incertitude et de comprendre les données.

- **Entropie** : Mesure d'incertitude dans une variable.
- **Divergence de Kullback-Leibler** : Distance entre distributions de probabilité.
- **Mutual information** : Relation entre deux variables aléatoires.
- **Codage** : Compression et transmission des données.

7. Apprentissage statistique et mathématiques du Machine Learning

- **Régressions** : Linéaire, polynomial, logistique.
- **Classification et clustering** : k-NN, SVM, K-means, DBSCAN.
- **Réseaux neuronaux** : Perceptrons, backpropagation, fonctions d'activation.
- **Méthodes bayésiennes** : Processus de Markov, Monte-Carlo, Gibbs sampling.
- **Validation et évaluation** : Courbes ROC, F1-score, cross-validation.

8. Traitement des séries temporelles

- **Modèles ARIMA, SARIMA** : Analyse des séries temporelles.
- **Transformée de Fourier** : Analyse fréquentielle.
- **Analyse des signaux** : Fenêtre glissante, convolution.

9. Géométrie et Topologie

- **Réduction dimensionnelle** : t-SNE, UMAP.
- **Distances et métriques** : Euclidienne, Manhattan, cosine.
- **Visualisation des données** : Graphes multidimensionnels, cartographie topologique.

Ces outils mathématiques forment une base solide pour résoudre une variété de problèmes en sciences de données, de l'analyse exploratoire à la modélisation prédictive avancée. Maîtriser ces concepts permet de comprendre et de développer des algorithmes efficaces pour tirer des insights des données.

II. Introduction à l'algèbre linéaire en sciences de données

L'algèbre linéaire est une branche fondamentale des mathématiques qui traite des espaces vectoriels, des matrices et des transformations linéaires. Elle joue un rôle central en sciences de données, car elle fournit les outils nécessaires pour manipuler, transformer et analyser des données structurées sous forme de vecteurs ou de matrices. Dans ce contexte, l'algèbre linéaire est omniprésente dans des domaines tels que le machine learning, le traitement des images, la compression de données, et l'analyse des réseaux.

Pourquoi l'algèbre linéaire est-elle essentielle en sciences de données ?

1. **Représentation des données** :
    - Les données en sciences de données sont souvent organisées sous forme de **matrices** ou de **vecteurs**. Par exemple :
        - Une base de données avec m échantillons et n caractéristiques peut être représentée par une matrice X de dimensions m×n.
        - Une image en niveaux de gris peut être stockée sous forme de matrice où chaque entrée correspond à l'intensité d'un pixel.
2. **Manipulation et transformation des données** :
    - L'algèbre linéaire permet d'effectuer des transformations comme la **normalisation**, la **réduction de dimension** (PCA), ou encore les **rotations** et **translations** dans l'espace multidimensionnel.
3. **Modèles de machine learning** :
    - De nombreux algorithmes de machine learning reposent sur des opérations d'algèbre linéaire. Par exemple :
        - La régression linéaire utilise la résolution de systèmes d'équations linéaires.
        - Les réseaux de neurones s'appuient sur des produits matriciels et des sommes pondérées.
4. **Calcul efficace** :
    - L'algèbre linéaire offre des méthodes puissantes pour travailler avec de grandes quantités de données grâce à des algorithmes optimisés comme la **décomposition matricielle**.

L'algèbre linéaire est un outil incontournable pour comprendre et résoudre des problèmes en sciences de données. Elle permet non seulement de structurer et manipuler les données, mais aussi de construire des modèles et d'implémenter des algorithmes efficaces.