Formation Machine Learning

Concepts et mise en oeuvre

Giraud François-Marie

3 Juin 2019

ENI Service

À propos de cette formation

Module 0

Votre formateur

Nom Giraud François-Marie

Courriel giraud.francois@gmail.com

Activité Consultant/Formateur en IA

Spécialité Machine Learning

Parcours Master IAD de l'UPMC, Ingénieur de recherche

Votre formation — description

Cette formation présente les **fondamentaux du machine learning** ainsi que les principales **techniques utilisées dans l'industrie**.

Votre formation — profil des stagiaires

Développeurs, ingénieurs informatiques désireux d'utiliser les techniques d'apprentissage automatique pour exploiter les données à leur disposition.

Bon niveau général en informatique, à l'aise en programmation.

Avez-vous un compte google ?

Connaissez-vous le Python ?

Votre formation — objectifs à atteindre

- poser un problème de machine learning
- prétraiter des données
- construire des modèles d'apprentissage pour des données annotées comme non-annotées
- gérer les apprentissages de vos modèles
- extraire des résultats actionnables

Votre formation — programme

- Introduction au machine learning
- Notions de maths
- Fondamentaux
- Régressions linéaire et logistique
- Machine à Vecteurs de Support (SVM)
- Arbres de décision
- Réduction de dimensionnalité et clustering
- Détection d'anomalies
- Réseaux de neurones
- Embeddings
- Système de recommandation

Votre formation — ressources

Les supports utilisées vous seront remis à chaque début de cours.

Tour de table — présentez-vous

- Votre nom
- Votre métier
- Votre société client si appliquable
- Vos compétences dans les domaines liés à cette formation
- Vos objectifs et vos attentes vis-à-vis de cette formation

Horaires

	matin	après-midi
lundi	9h00-12h00	14h00-17h30
mardi	9h00-12h00	14h00-17h30
mercredi	9h00-12h00	14h00-17h30
jeudi	9h00-12h00	14h00-17h30
vendredi	9h00-12h00	14h00-17h30

Pauses

Une pause de 10 minutes pour couper chaque demi-journée.

Introduction au machine learning

Module 1

Objectifs

Objectifs

- cerner ce qu'est le machine learning
- appréhender les différentes facettes du domaine

Machine Learning

Machine learning

Qu'est-il pour vous ?

Machine learning

- pas de définition exacte
- idée transversale : éviter la programmation **explicite**.
- création de programmes qui utilisent des données ou des algorithmes généraux pour apprendre à réaliser leurs tâches

Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)

Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)
- les modèles (arbres, grammaires, automates, réseaux de neurones ...)

Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)
- les modèles (arbres, grammaires, automates, réseaux de neurones ...)
- les données (tabulaire, image, texte, vidéo, graphe, ...)

Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)
- les modèles (arbres, grammaires, automates, réseaux de neurones ...)
- les données (tabulaire, image, texte, vidéo, graphe, ...)
- les techniques (statistiques, symboliques, probabilistes, ...)

Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)
- les modèles (arbres, grammaires, automates, réseaux de neurones ...)
- les données (tabulaire, image, texte, vidéo, graphe, ...)
- les techniques (statistiques, symboliques, probabilistes, ...)
- les contraintes (real time, embarqué, big data, multilingue, ...)

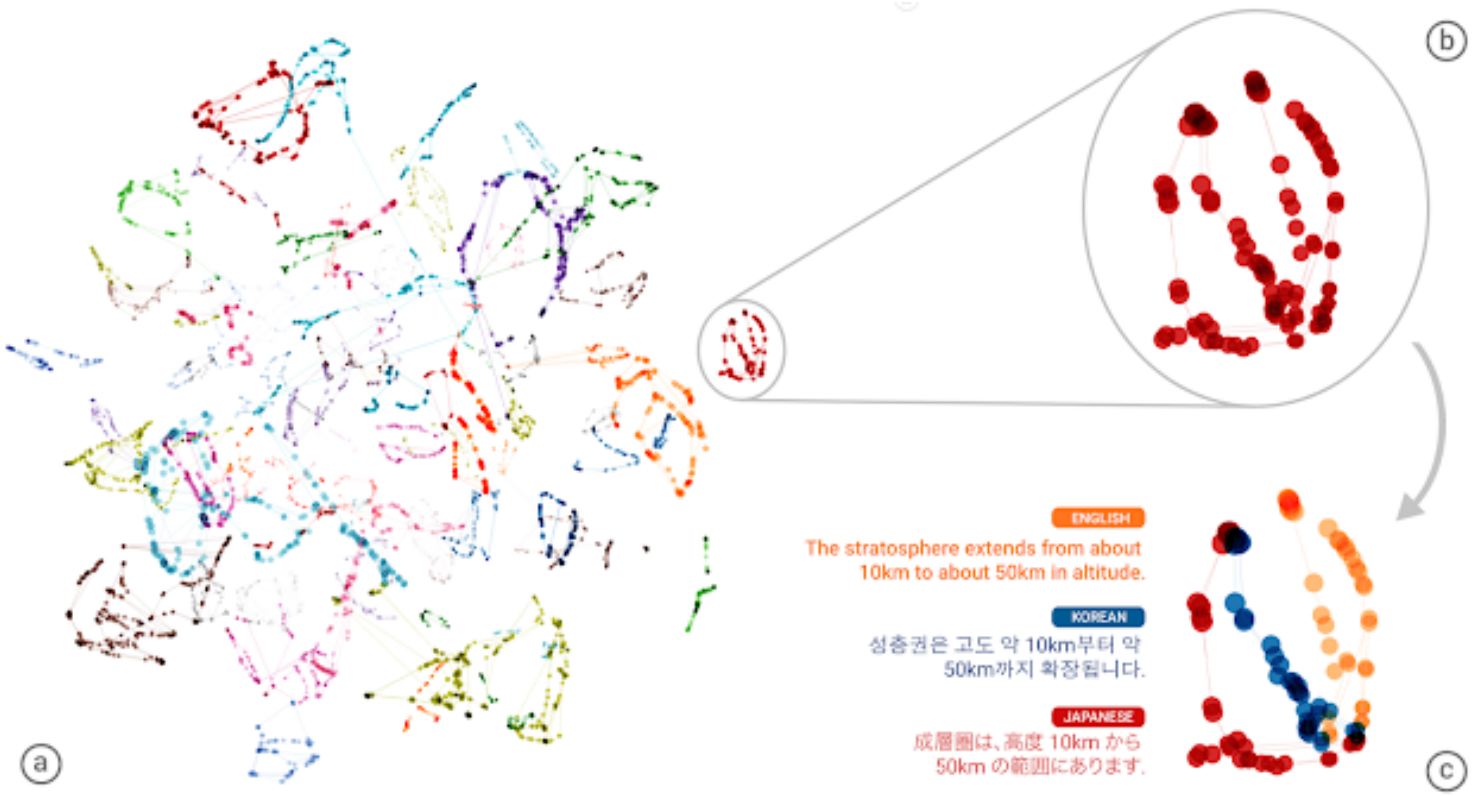
Points de vue

Beaucoup de façons de voir le machine learning. Basées sur :

- les paradigmes (supervisé, non supervisé, renforcement, en ligne, ...)
- les modèles (arbres, grammaires, automates, réseaux de neurones ...)
- les données (tabulaire, image, texte, vidéo, graphe, ...)
- les techniques (statistiques, symboliques, probabilistes, ...)
- les contraintes (real time, embarqué, big data, multilingue, ...)

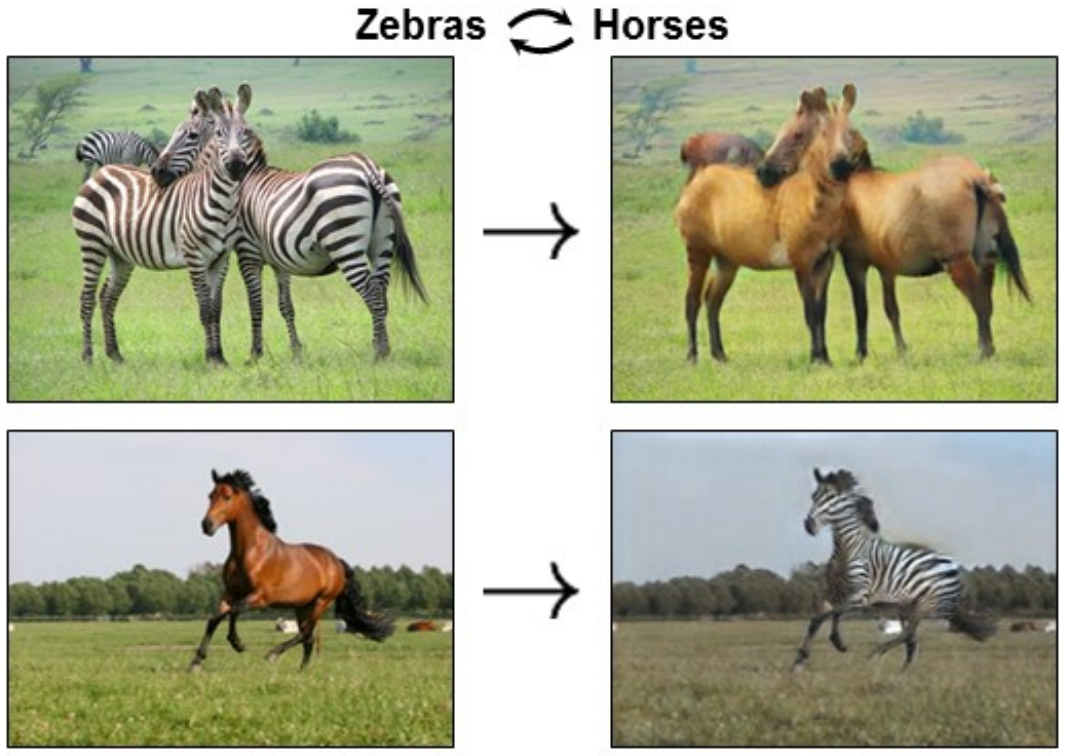
→ Domaine **extrêmement** vaste.

Facettes du machine learning — supervised learning



Demande beaucoup de données, parfois couteuses. Modèles performants en sortie.

Facettes du machine learning — unsupervised learning



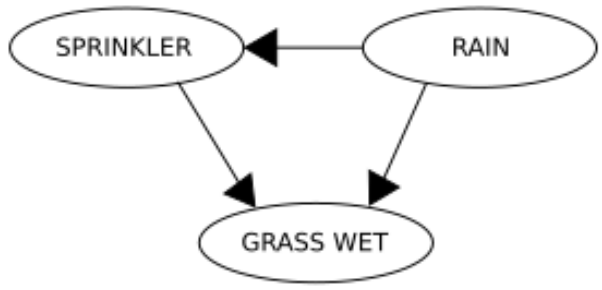
Pas besoin d'annotation → données moins couteuses. Limite les possibilités des modèles.

Facettes du machine learning — reinforcement learning

Paradigme de d'acquisition des données différent. Modèles potentiellement extrêmement performants.

Facettes du machine learning — bayesian networks

RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

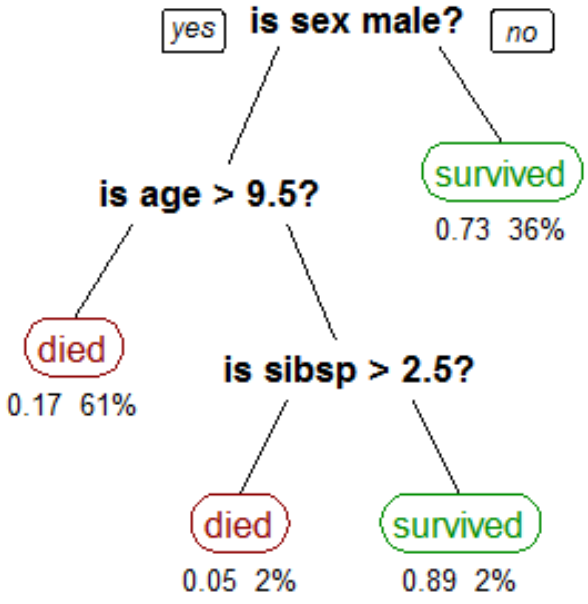


RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Très interprétable, requêtable.

Facettes du machine learning — decision trees



Assez interprétable, robuste, couteau-suisse du machine learning tabulaire.

Calculabilité vs expressivité

Un modèle facilement calculable est souvent peu expressif.

Inversement, un modèle peu calculable est souvent expressif (sinon mauvais modèle).

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données
- difficulté du problème à résoudre

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données
- difficulté du problème à résoudre
- besoin d'interprétabilité

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données
- difficulté du problème à résoudre
- besoin d'interprétabilité
- contraintes techniques

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données
- difficulté du problème à résoudre
- besoin d'interprétabilité
- contraintes techniques
- contraintes de délai

Choisir la bonne facette

Critères pour s'orienter dans les approches de machine learning :

- quantité de données à disposition
- qualité du signal d'apprentissage dans les données
- difficulté du problème à résoudre
- besoin d'interprétabilité
- contraintes techniques
- contraintes de délai
- ... et d'autres en fonction des domaines métiers

Conclusion

- le machine learning est un champ vaste.
- il existe sûrement un modèle/paradigme pour vos besoins
- l'important est de définir les bons critères

Discussion

- à quelles données allez-vous appliquer le machine learning ? À quels besoins ?
- aurez-vous besoin de modèles interprétables ou simplement très performant en prédiction ?
- quelles sont vos contraintes ?

Mathématiques pour le machine learning

Module 2

Objectifs

Objectifs

- exprimer des transformations de données grâce à l'algèbre linéaire
- minimiser des fonctions analytiquement
- décrire l'incertain
- décrire des données

Algèbre linéaire

Utilité

- décrire des transformations simples sur un dataset entier avec des mécanismes adaptés
- comprendre les possibilités et les limites de ces transformations simples.

Transformation linéaire

- algèbre linéaire = on se limite aux sommes pondérées des inputs.
- bonne nouvelle : énorme partie des opérations en machine learning

Description des données — échantillon

Python :

```
data = (1, 3)
```

Algèbre linéaire :

$$\mathbf{d} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Description des données — dataset

Python :

```
data = [(1, 3),
        (2, 2),
        (4, 2)]
```

Algèbre linéaire :

$$\mathbf{D} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

Description des transformations linéaires

Python :

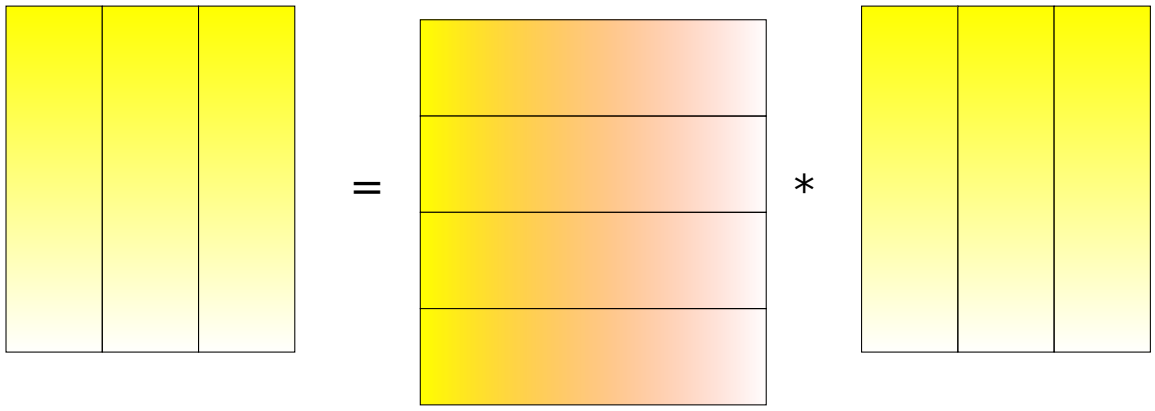
```
def weights(x, y):
    return x * 2 + y / 2
```

Algèbre linéaire :

$$\mathbf{w} = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix}$$

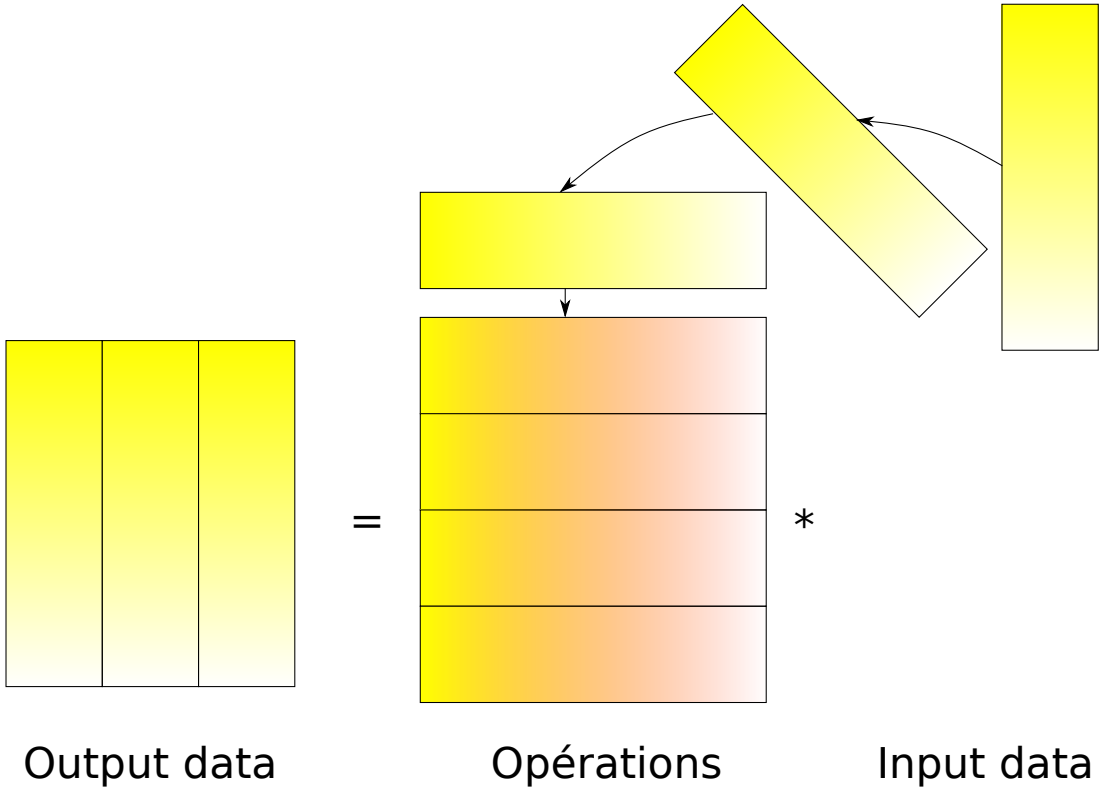
Transformation linéaire = somme pondérée.

Application d'une transformation linéaire à un exemple



Bonne intuition à garder : Verser les colonnes (les exemples du dataset) dans les lignes (les opérations).

Bonne intuition à garder



Bonne intuition à garder : Verser les colonnes (les exemples du dataset) dans les lignes (les opérations).

Application d'une transformation linéaire à un exemple

Python :

```
data = (1, 3)

def weights(x, y):
    return 2 * x + y / 2

res = weights(*data)
```

Algèbre linéaire :

$$f = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$
$$= 2 \times 1 + \frac{1}{2} \times 3$$

Application d'une transformation linéaire à un dataset

Python :

```
data = [(1, 3),
        (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

res = [f(x, y)
       for x, y
       in data]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

Application d'une transformation linéaire à un dataset

Python :

```
data = [(1, 3),
        (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

res = [f(x, y)
       for x, y
       in data]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & & \end{bmatrix}$$

Application d'une transformation linéaire à un dataset

Python :

```
data = [(1, 3),
        (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

res = [f(x, y)
       for x, y
       in data]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & \end{bmatrix}$$

Application d'une transformation linéaire à un dataset

Python :

```
data = [(1, 3),
        (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

res = [f(x, y)
       for x, y
       in data]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & 9 \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & & \\ & & \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & \\ & & \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & 9 \\ & & \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & 9 \\ 6,5 & & \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & 9 \\ 6,5 & 5 & \end{bmatrix}$$

Application de plusieurs transformations linéaires à un dataset

Python :

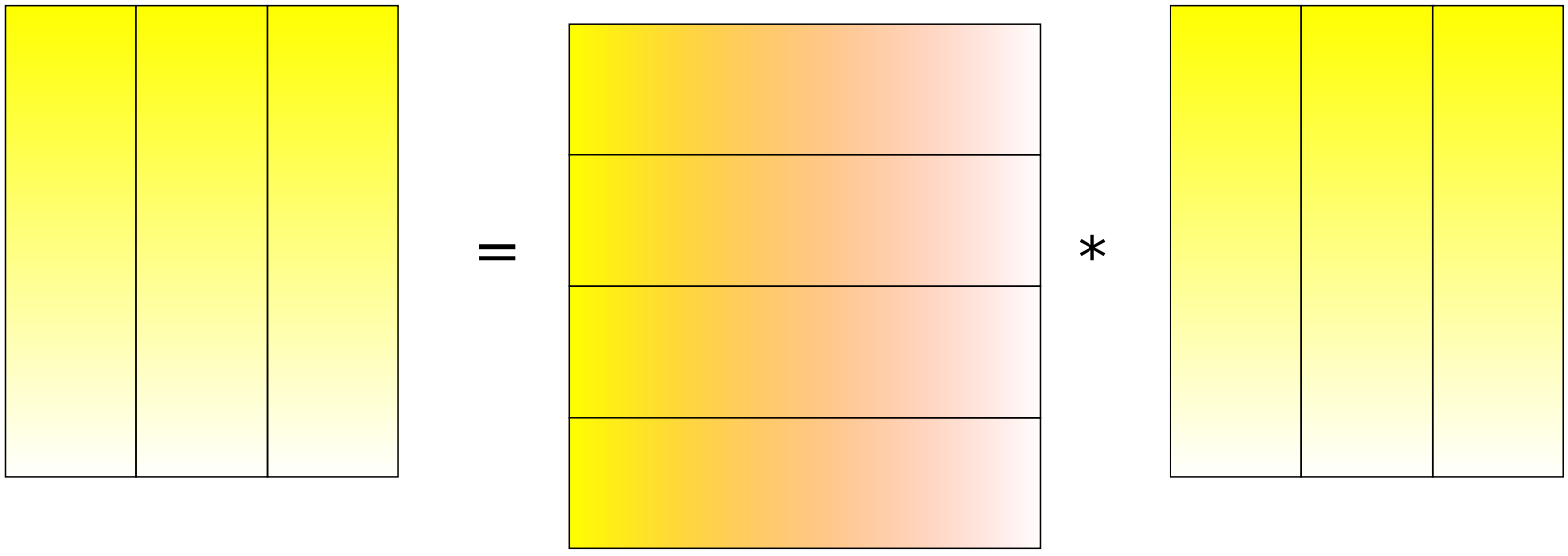
```
data = [(1, 3), (2, 2),
        (4, 2)]

def f(x, y):
    return x * 2 + y / 2

def g(x, y):
    return x / 2 + y * 2

res = [[t(x, y) for x, y
                in data]
       for t in [f, g]]
```

Algèbre linéaire :

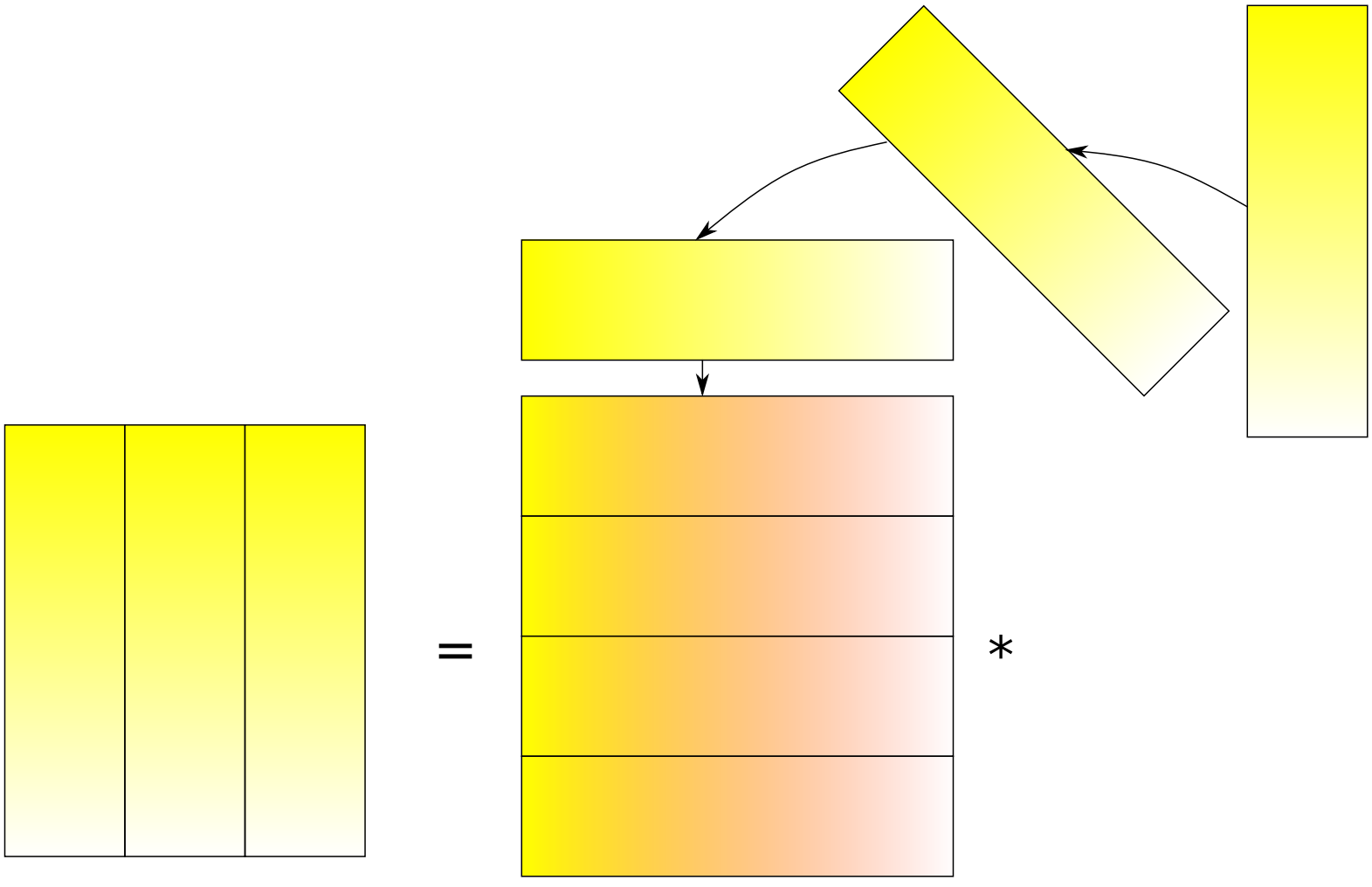
$$\text{res} = \begin{bmatrix} 2 & \frac{1}{2} \\ \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 3 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 3,5 & 5 & 9 \\ 6,5 & 5 & 6 \end{bmatrix}$$

Bonne intuition à garder



Bonne intuition à garder

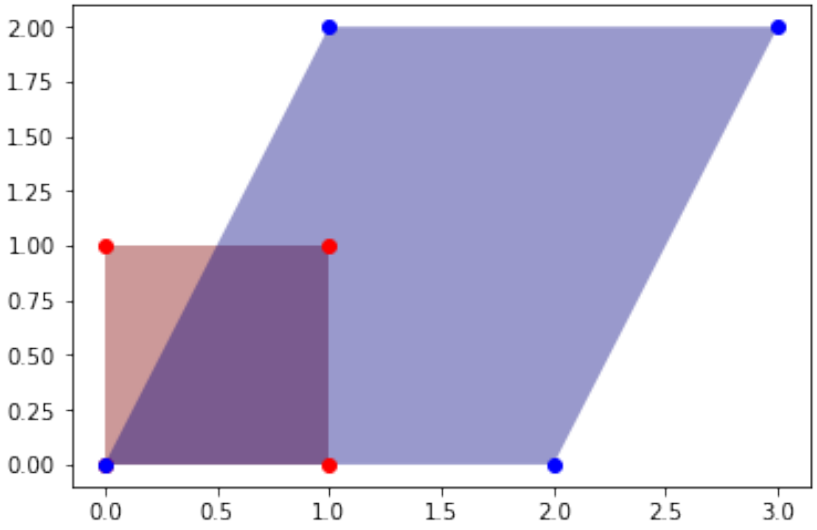


Output data Opérations Input data

Exercice

$$\begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} = ?$$

Exemple de transformation



$$\text{Bleu} = \text{Transformation} \times \text{Rouge}$$

$$= \begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

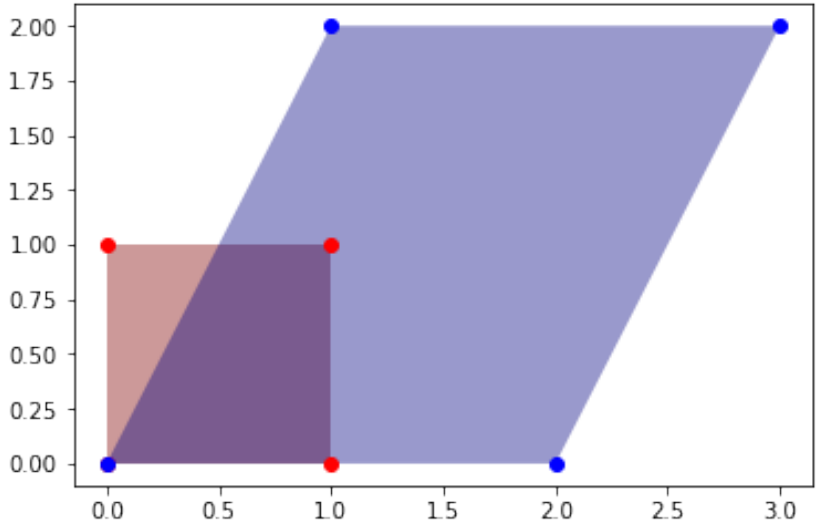
$$= \begin{bmatrix} 0 & 2 & 3 & 1 \\ 0 & 0 & 2 & 2 \end{bmatrix}$$

Formes

	Taille
D	(m, n)
W	(o, m)
WD	(o, n)

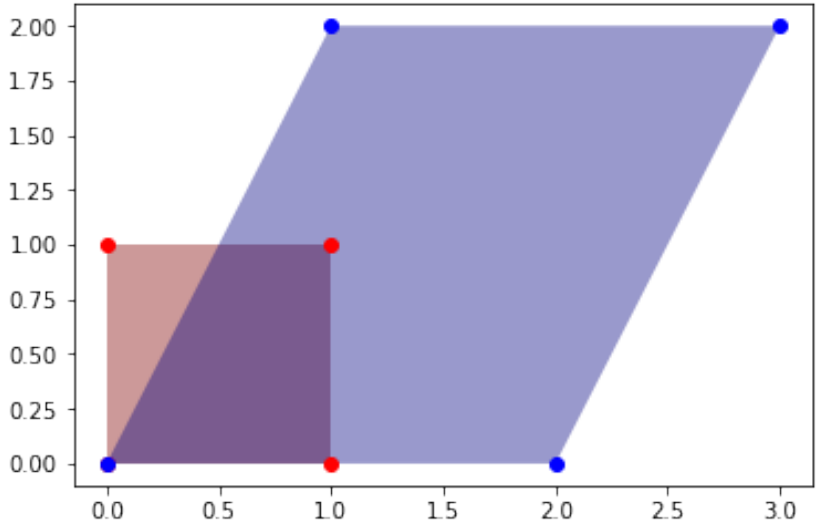
$\rightarrow$ dataset avec m features et n lignes transformé par o opérations donne dataset de o features et n lignes.

Déterminant



Facteur de dilatation de la transformation (ratio aire bleue sur aire rouge).

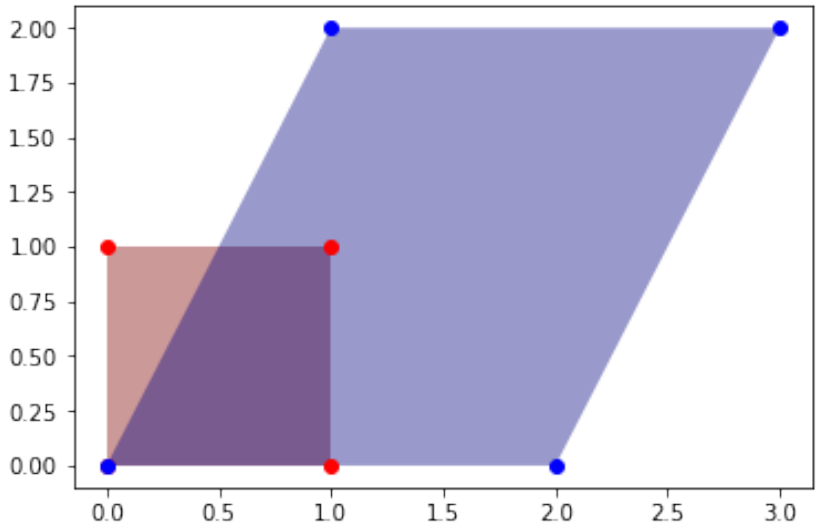
Déterminant



Facteur de dilatation de la transformation (ratio aire bleue sur aire rouge).

Quel est le déterminant de cette transformation ?

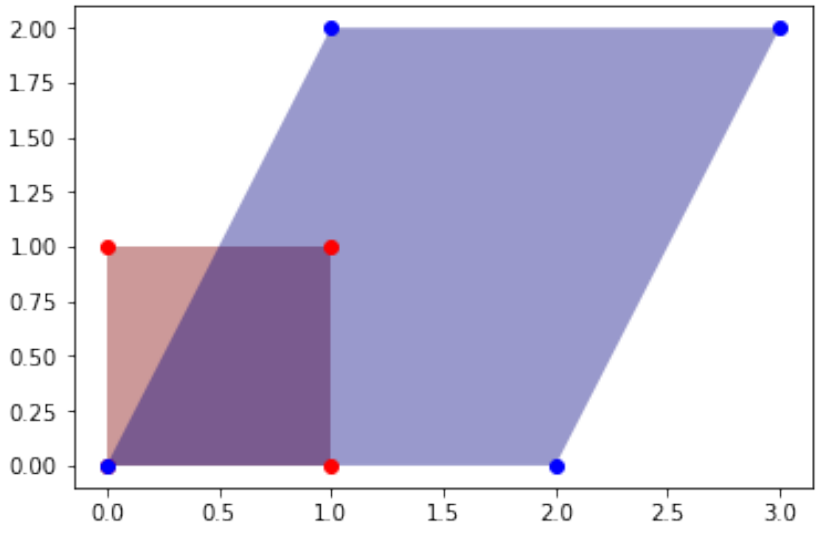
Déterminant



Facteur de dilatation de la transformation (ratio aire bleue sur aire rouge).

Quel est le déterminant de cette transformation ? 4.

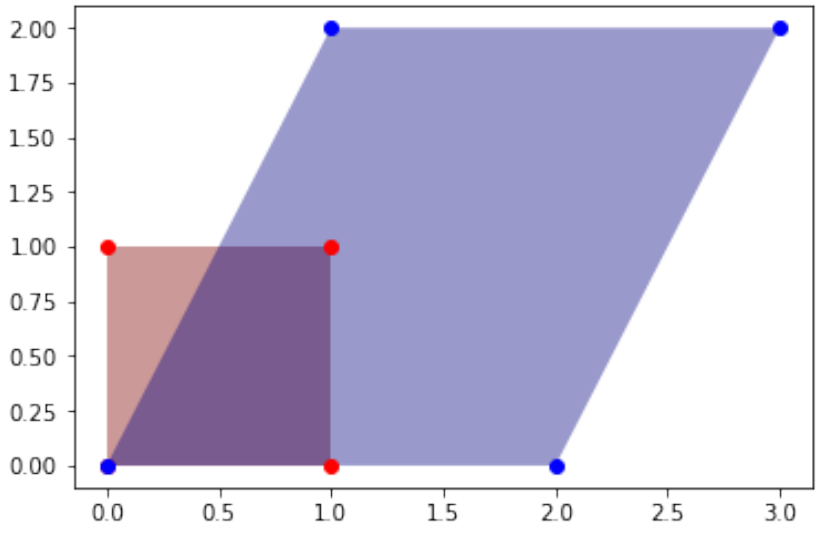
Vecteur propre



Vecteur partant de l'origine qui conserve sa direction malgré la transformation.

Pouvez-vous en trouver un ?

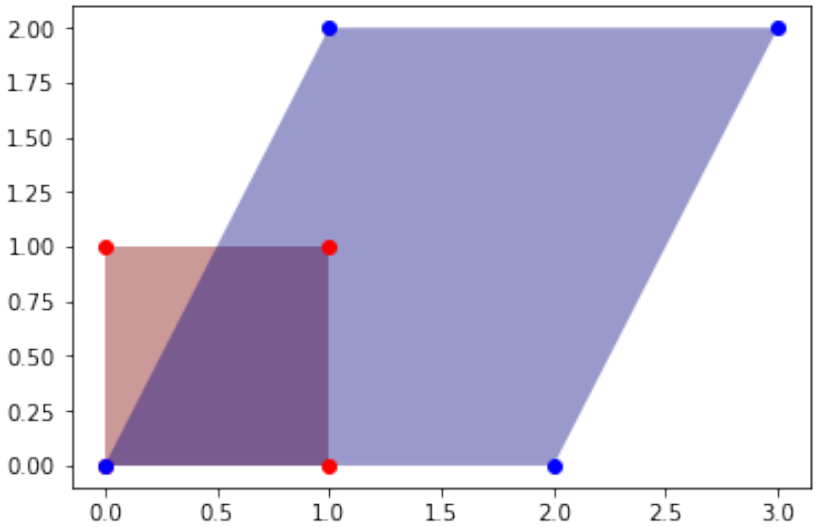
Vecteur propre



Vecteur partant de l'origine qui conserve sa direction malgré la transformation.

Pouvez-vous en trouver un ? $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ par exemple.

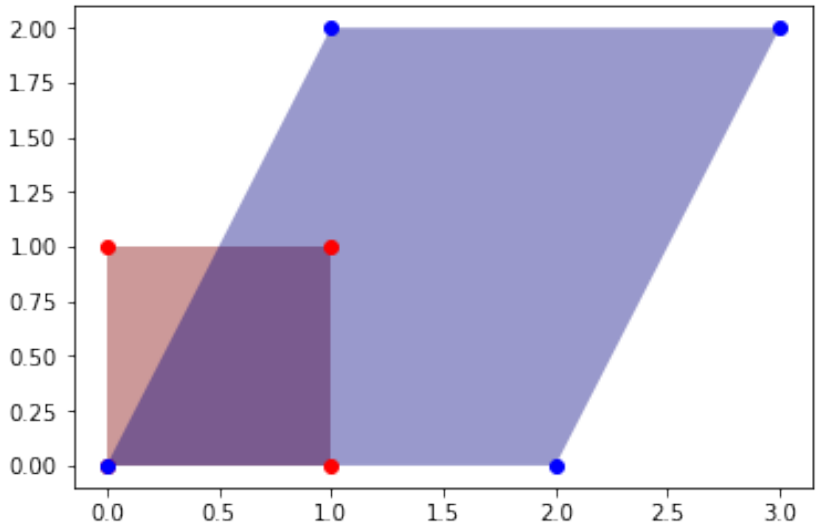
Valeur propre



Facteur par lequel un vecteur propre est redimensionné.

Quelle est la valeur propre de $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$?

Valeur propre



Facteur par lequel un vecteur propre est redimensionné.

Quelle est la valeur propre de $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$? 2.

Analyse

Utilité

Souvent besoin de minimiser une fonction en machine learning.

Utilité

Souvent besoin de minimiser une fonction en machine learning.

(trouver le x pour lequel f est minimale)

Utilité

Souvent besoin de minimiser une fonction en machine learning.

(trouver le x pour lequel f est minimale)

Vous souvenez-vous de comment l'on procède ?

Idée clef

Décider d'un x de départ puis suivre la pente jusqu'au minimum.

Idée clef

Décider d'un x de départ puis suivre la pente jusqu'au minimum.

Pente = dérivée

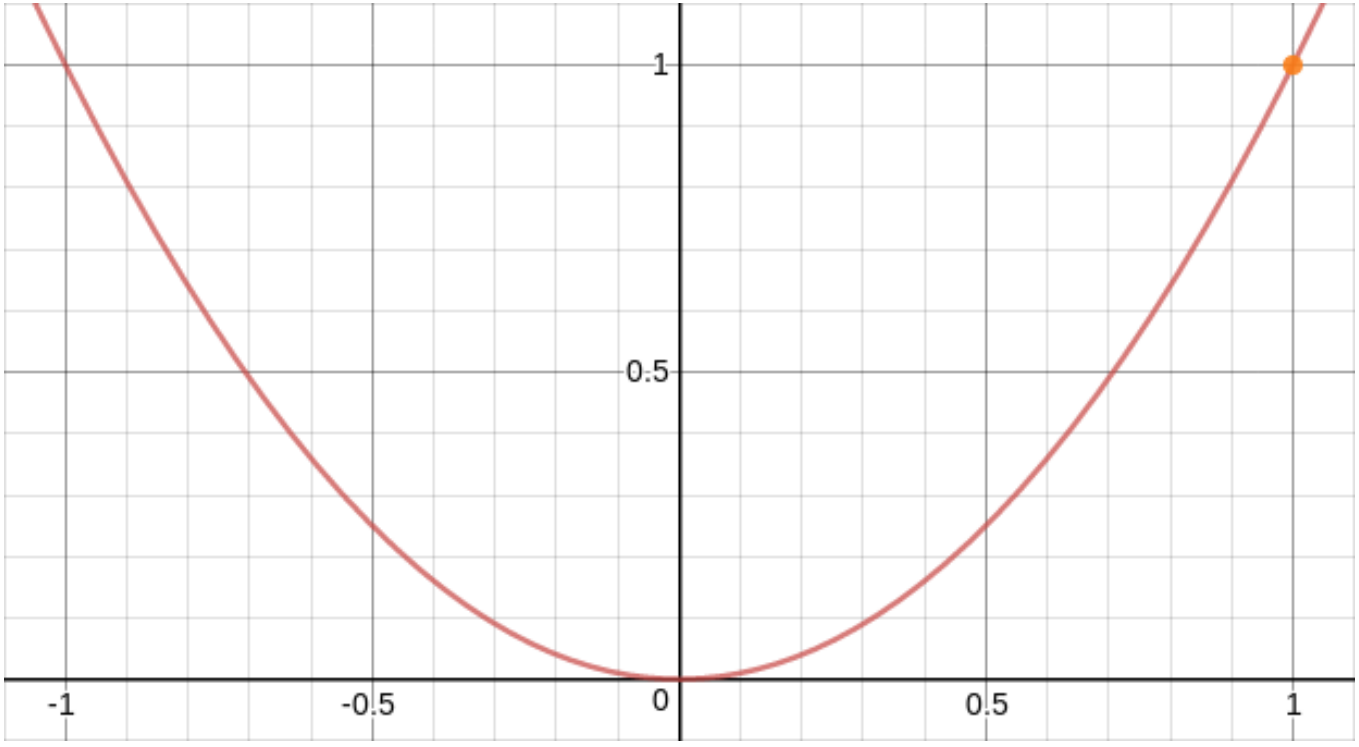
Idée clef

Décider d'un x de départ puis suivre la pente jusqu'au minimum.

Pente = dérivée

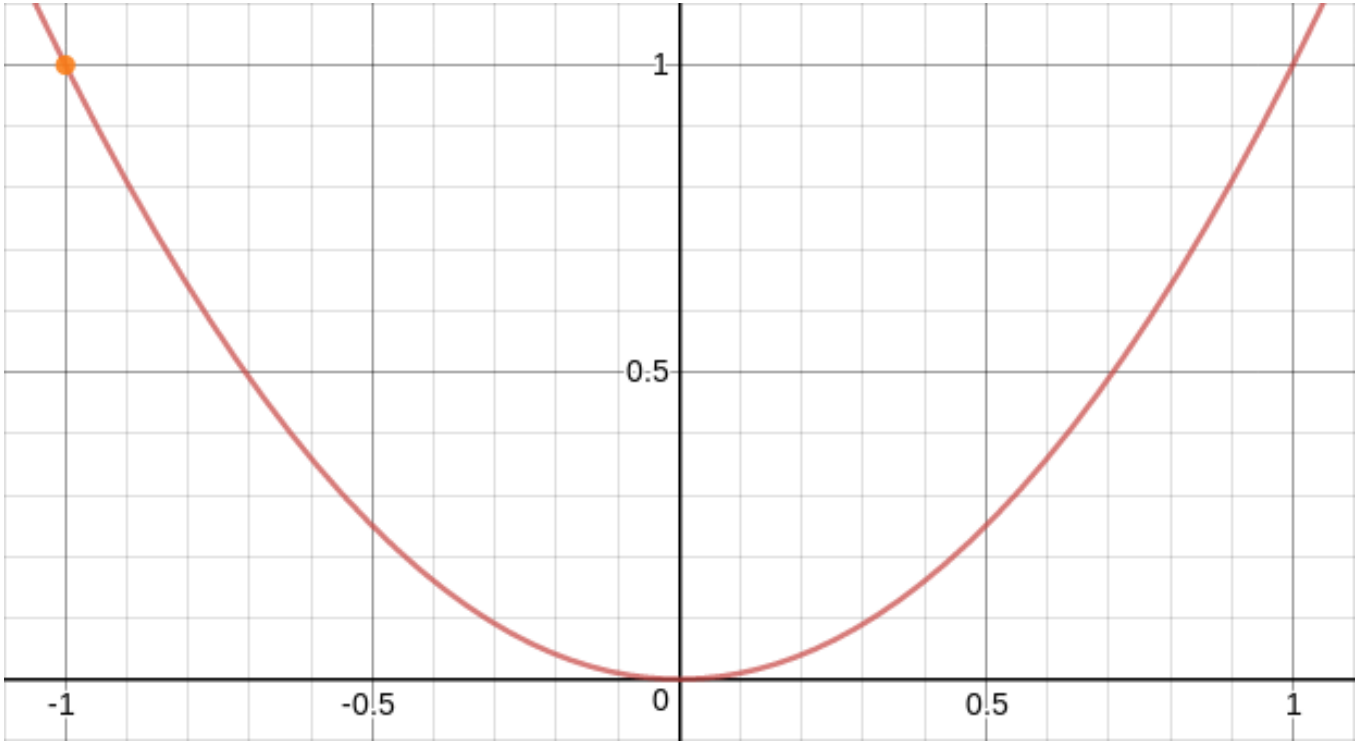
$\rightarrow$ Modifier itérativement x par un pas vers l'opposé de la dérivée.

Pente positive



Opposé de la pente $= -2$. Avec un pas de $0,1$, on passe de 1 à $0,8$.

Pente négative

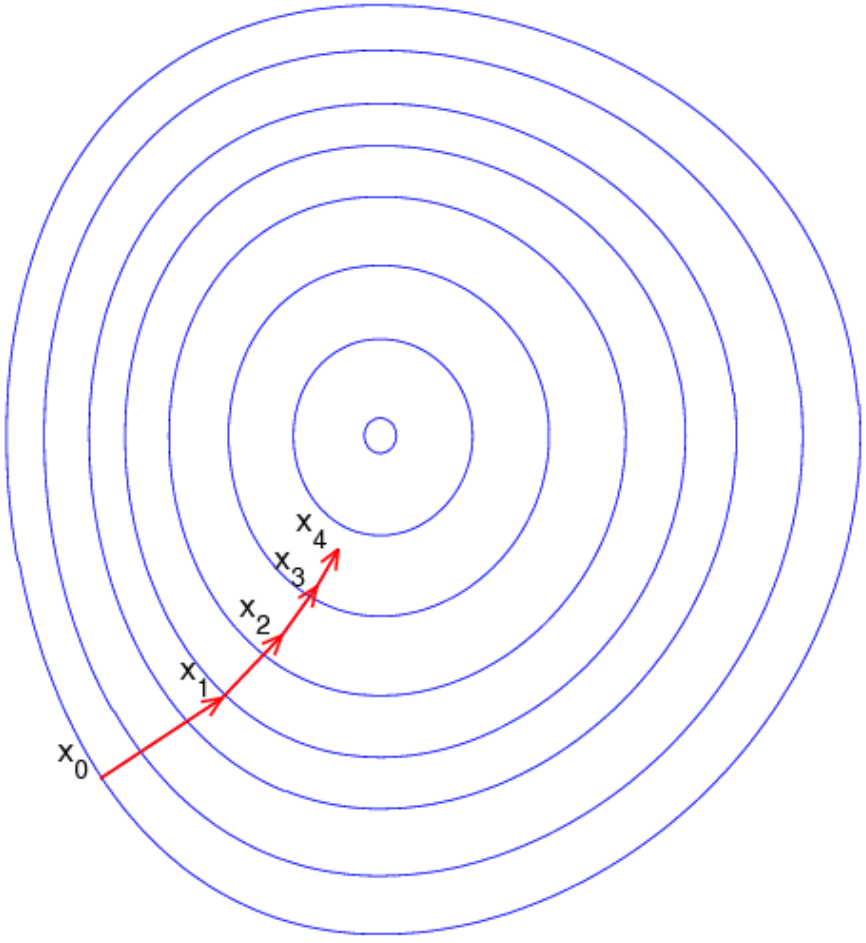


Opposé de la pente $= 2$. Avec un pas de $0,1$, on passe de -1 à $-0,8$.

Extension à plusieurs dimensions

- dérivée $\rightarrow$ gradient
- identique sinon !

Exemple en 2 dimensions



Probabilités

Utilité

- quantifier l'incertain
- support pour les statistiques

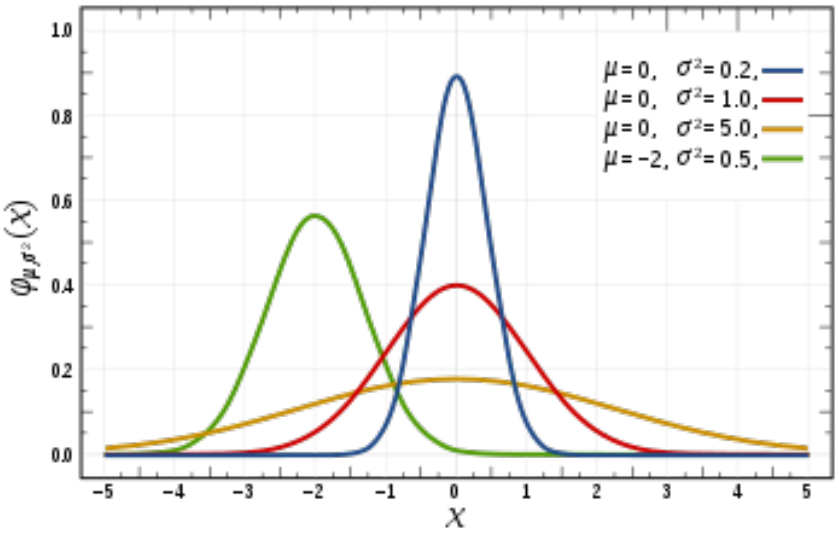
Probabilité

- la probabilité de l'événement X est notée $P(X)$
- $P(X) \in [0, 1]$
- $P(X) = 0 \iff$ X est impossible
- $P(X) = 1 \iff$ X est certain
- $P(\neg X) = 1 - P(X)$

Loi de probabilité

Décrit le comportement aléatoire d'un phénomène dépendant du hasard.

- $\sum_u P(X = u) = 1$ en discret
- $\int P(X)dX = 1$ en continu
- loi uniforme
- loi normale/gaussienne

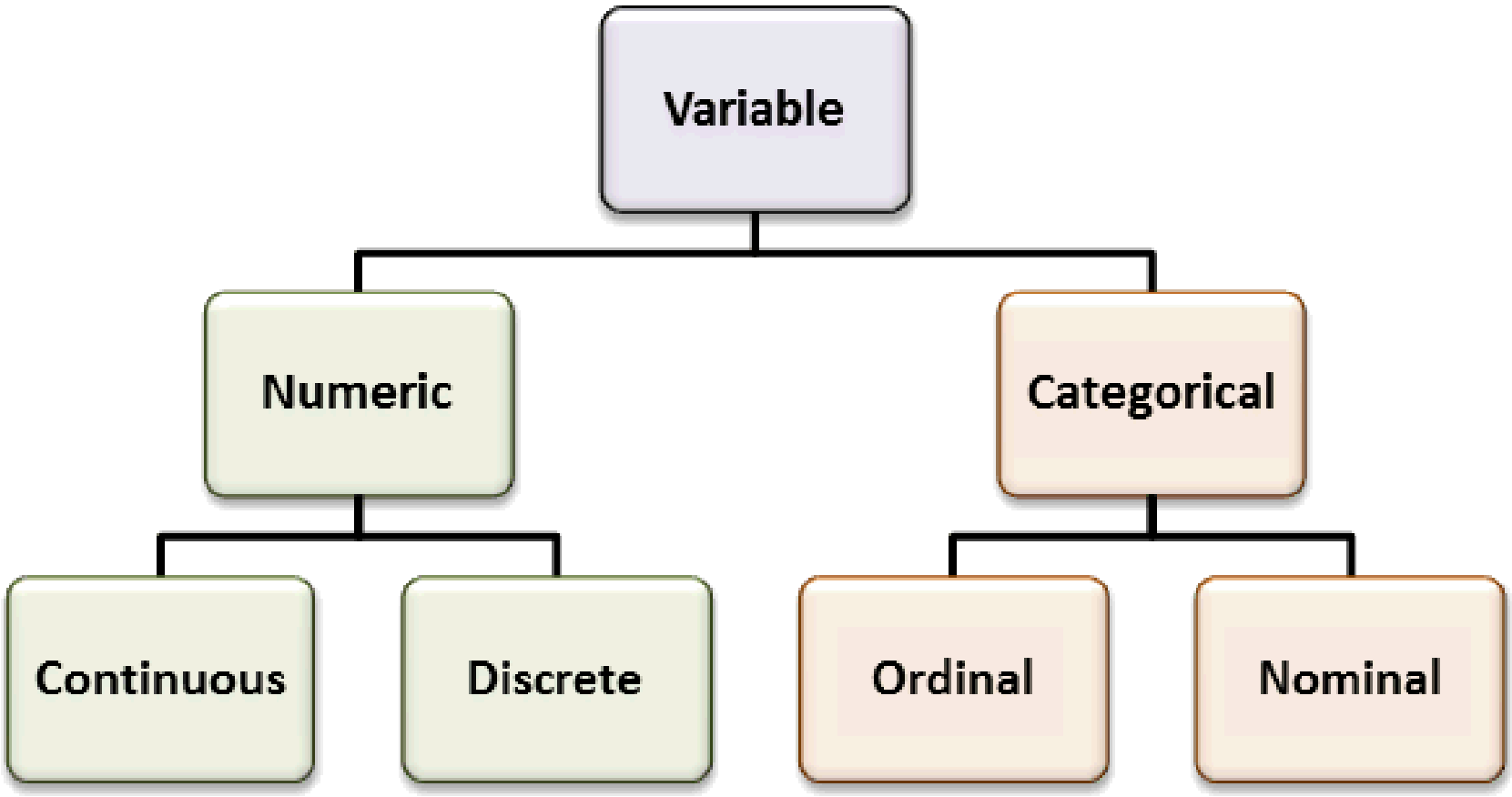


Statistiques

Utilité

- description et compréhension des données
- correction pour faciliter les traitements

Types de variables



Hypothèse

Pré-requis pour les mesures statistiques qui suivent (et la plupart du machine learning) :

- les données **doivent être issues d'une même loi**
- chaque échantillon doit être **indépendant** des autres
- pas évident en pratique !

Hypothèse

Pré-requis pour les mesures statistiques qui suivent (et la plupart du machine learning) :

- les données **doivent être issues d'une même loi**
- chaque échantillon doit être **indépendant** des autres
- pas évident en pratique ! Pourquoi ?

Variance

Mesure la dispersion d'une série statistique (ou d'une variable) :

$$V(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Pour la calculer :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

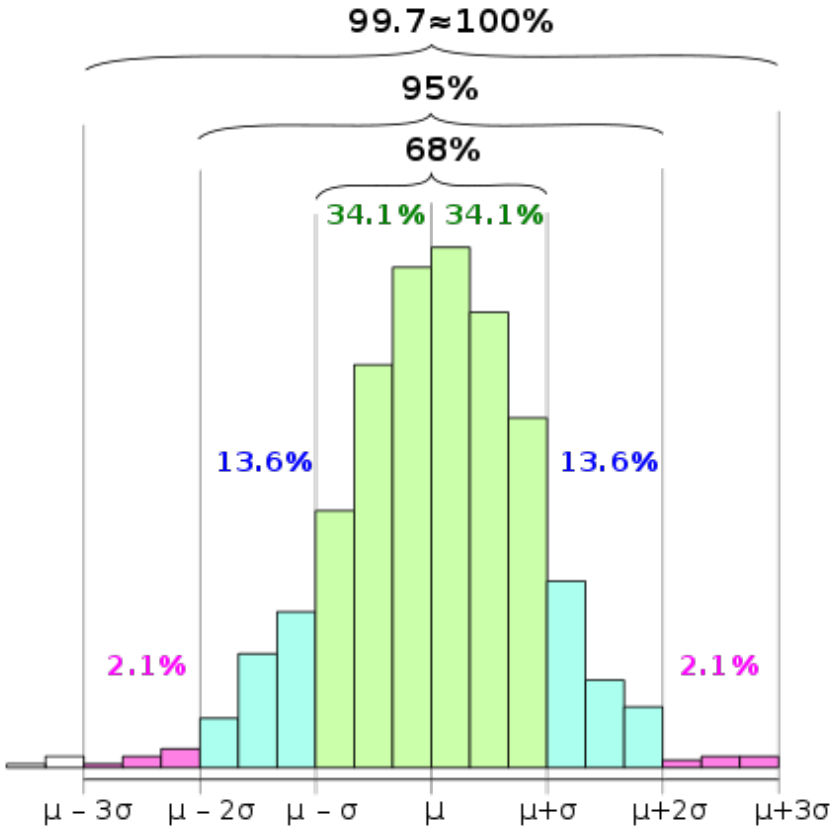
Écart-type

Racine carrée de la variance

$$\sigma(X) = \sqrt{V(X)}$$

Écart-type — règle des 68, 95 et 99,7

Pour les lois normales :



Quartile

Les quartiles (Q_1, Q_2 et Q_3) divisent les données en 4 intervalles contenant le même nombre d'observations.

Déclinable en quantile de taille arbitraire (décile, percentile).

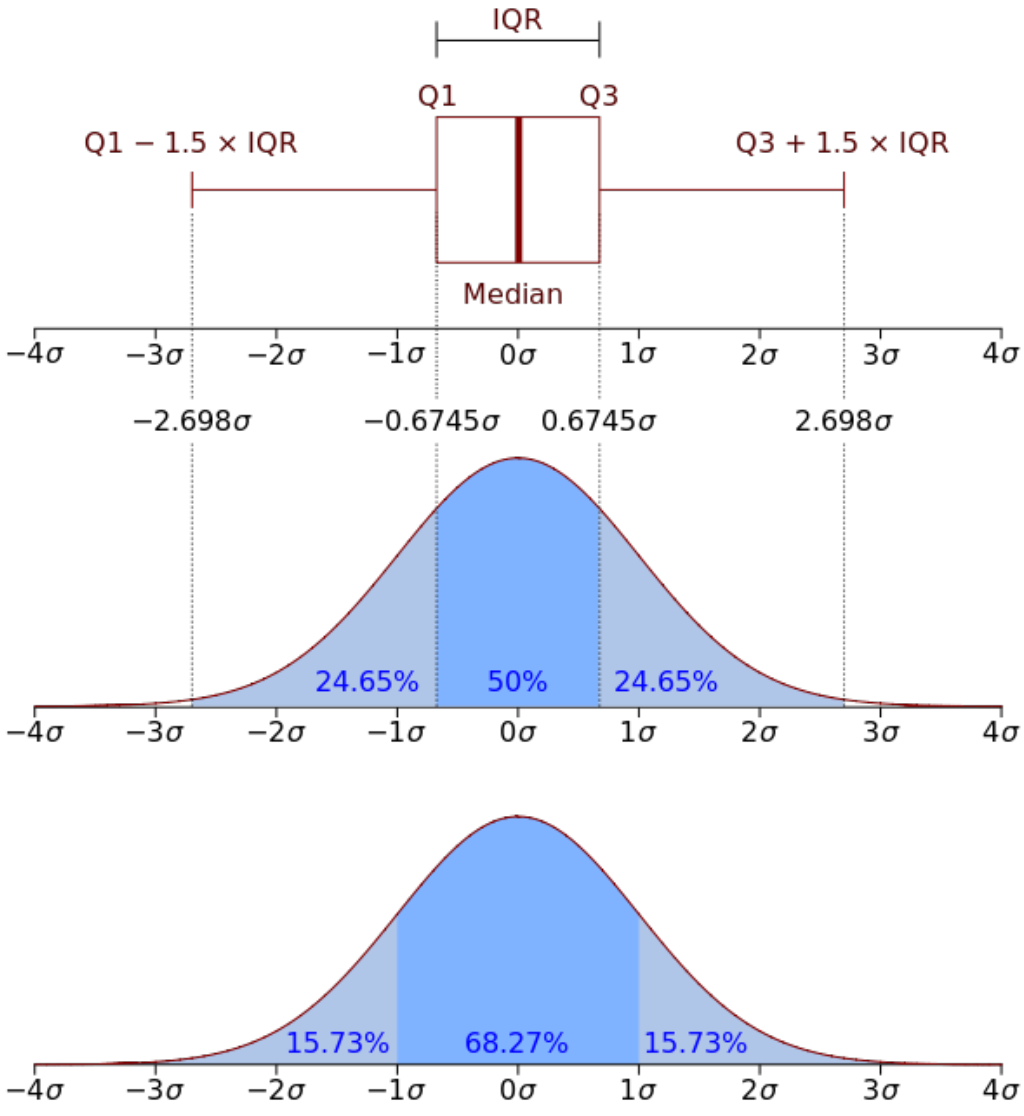
Quartile

Les quartiles (Q_1, Q_2 et Q_3) divisent les données en 4 intervalles contenant le même nombre d'observations.

Déclinable en quantile de taille arbitraire (décile, percentile).

Que veut dire être dans le 95^{e} percentile ?

Boxplot


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

Covariance

Mesure la variabilité jointe de deux variables aléatoires :

$$V(X) = \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])\right]$$

$$\text{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

Pour la calculer :

$$\text{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Corrélation

Covariance divisée par le produit des écart-types :

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Intérêt ?

Corrélation

Covariance divisée par le produit des écart-types :

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

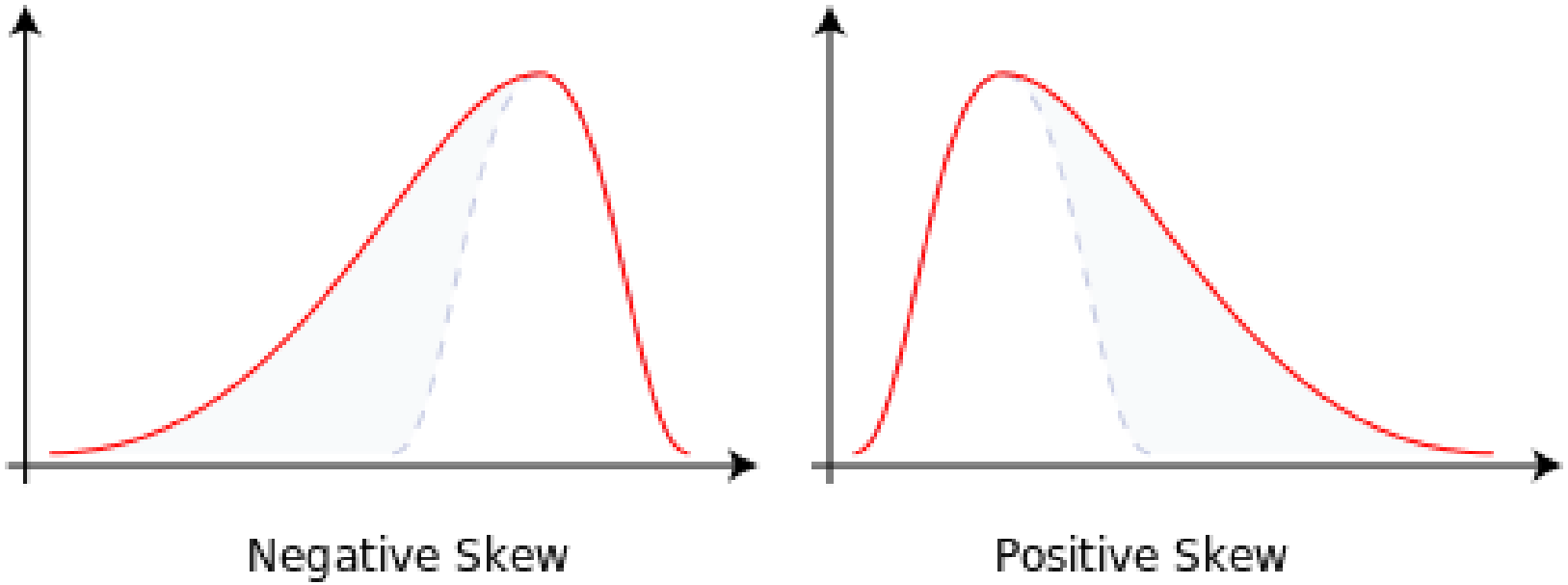
Intérêt ? Pas d'unité.

Test de normalité

Pour tester (et corriger) la normalité d'une distribution, on utilise deux mesures :

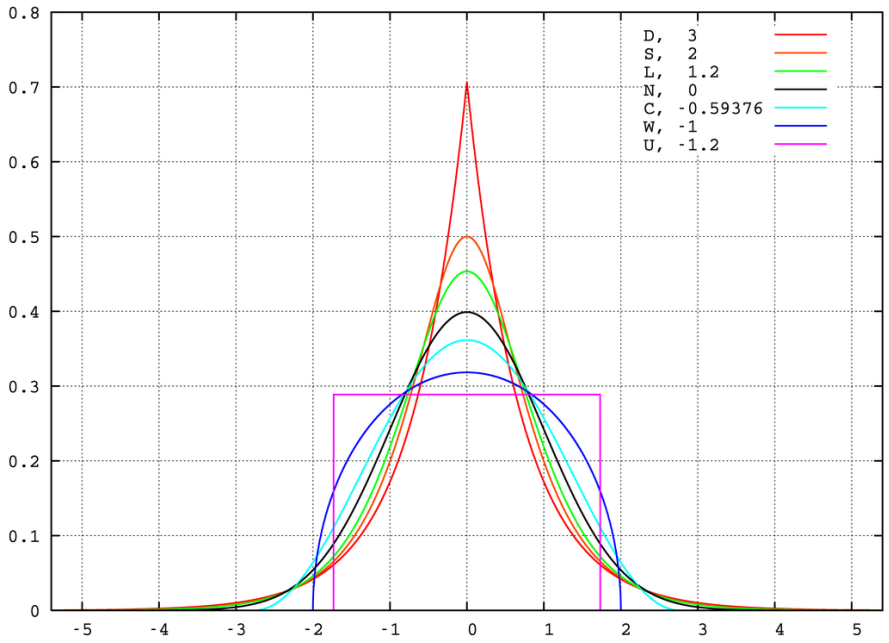
- l'asymétrie (*skew*)
- le kurtosis

Asymétrie



$$\text{asym}(X) = \mathbb{E}\left[\left(\frac{X - \bar{X}}{\sigma}\right)^3\right]$$

Kurtosis



$$\text{kurt}(X) = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Transformation de Box-Cox

Asymétrie et kurtosis peuvent se corriger avec la transformation de Box-Cox ou des transformations log.

Quizz

Algèbre linéaire

- comment décrit-on on enregistrement ?

- comment décrit-on on enregistrement ?
- comment décrit-on une transformation linéaire ?

Algèbre linéaire

- comment décrit-on on enregistrement ?
- comment décrit-on une transformation linéaire ?
- quel est le sens de la multiplication de matrices dans le contexte dataset/opérations ?

Analyse

- intuitivement, qu'est-ce que nous apprend une dérivée ?

Analyse

- intuitivement, qu'est-ce que nous apprend une dérivée ?
- à quelle valeur de la dérivée d'une fonction atteint-on un minimum ?

Analyse

- intuitivement, qu'est-ce que nous apprend une dérivée ?
- à quelle valeur de la dérivée d'une fonction atteint-on un minimum ?
- si la dérivée est négative, dans quel sens faut-il faire évoluer x ?

Analyse

- intuitivement, qu'est-ce que nous apprend une dérivée ?
- à quelle valeur de la dérivée d'une fonction atteint-on un minimum ?
- si la dérivée est négative, dans quel sens faut-il faire évoluer x ?
- est-ce que la pas d'apprentissage impacte seulement les performances en temps de calcul ?

Probabilités

- quelle est la forme d'une gaussienne ?

- quelle est la forme d'une gaussienne ?
- à combien somme une loi discrète ?

Statistiques

- quelle est l'hypothèse que l'on fait dans la plupart des approches de statistiques / machine learning ?

Statistiques

- quelle est l'hypothèse que l'on fait dans la plupart des approches de statistiques / machine learning ?
- que nous apprennent la variance et l'écart-type ?

Statistiques

- quelle est l'hypothèse que l'on fait dans la plupart des approches de statistiques / machine learning ?
- que nous apprennent la variance et l'écart-type ?
- que nous apprennent la covariance et la corrélation ?

Statistiques

- quelle est l'hypothèse que l'on fait dans la plupart des approches de statistiques / machine learning ?
- que nous apprennent la variance et l'écart-type ?
- que nous apprennent la covariance et la corrélation ?
- comment peut-on savoir si une distribution est normale ?

Conclusion

Conclusion

- algèbre linéaire $\rightarrow$ raisonner sur des opérations simples et les décrire efficacement
- minimiser une fonction continue $\rightarrow$ dérivée
- décrire l'incertain $\rightarrow$ probabilités
- caractériser une série de données $\rightarrow$ statistiques

Fondamentaux du machine learning

Module 3

Objectifs

Objectifs

- adopter un workflow cohérent de data science
- comprendre les écueils à éviter (biais statistiques)
- acquérir les bonnes pratiques

Workflow

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. communiquer les résultats

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. communiquer les résultats
7. rendre son analyse reproductible

Workflow

Un projet de data science c'est :

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. communiquer les résultats
7. rendre son analyse reproductible

3 et 4 se font souvent en même temps. Pas linéaire, retours en arrière fréquents.

Définition du problème

Définition du problème

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. communiquer les résultats
7. rendre son analyse reproductible

Définition du problème

En partant d'un problème business ou scientifique réel :

- métrique pour quantifier le problème

Définition du problème

En partant d'un problème business ou scientifique réel :

- métrique pour quantifier le problème
- pas de métrique → problème mal posé. Pourquoi ?

Définition du problème

En partant d'un problème business ou scientifique réel :

- métrique pour quantifier le problème
- pas de métrique $\rightarrow$ problème mal posé. Pourquoi ?
- métriques intrinsèque et extrinsèque si possible

Obtention des données

Obtention des données

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Types d'étude

- observationnelle
- expérimentale

Différence ?

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...
- trouver de fausses variables explicatives

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...
- trouver de fausses variables explicatives

Datascience

Datascience $\rightarrow$ études souvent observationnelles.

Risques importants de :

- variables confondantes (Ex : “obésité” dans la corrélation entre “conso. viande” et “cancer colon”)
- biais statistiques
  - sélection, autosélection
  - mesure
  - attrition
  - ...
- trouver de fausses variables explicatives

$\rightarrow$ Le garder en tête pendant toute l'étude.

Qualité

Souvent, meilleures données $>$ meilleurs modèles

$\rightarrow$ À garder en tête pendant toute l'étude, en particulier durant l'entraînement de modèles

Préparation des données

Préparation des données

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. préparer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Préparation des données

- valeurs manquantes
- préprocessing (texte, image)
- standardisation
- transformation

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante
  - moyenne de la colonne

Préparation des données — valeurs manquantes

Gênant pour certains modèles. Plusieurs options :

- supprimer les enregistrements
- remplacer par une valeur (imputation) :
  - constante
  - moyenne de la colonne
  - prédiction d'un autre modèle

Préparation des données — préprocessing

- tokenizer, POS-tagger le texte (`https://spacy.io/`)
- utiliser un réseau de neurones préentraîné sur les images (`https://keras.io/applications/`)
- appliquer une transformée de fourier sur le son
- ...

Préparation des données — standardisation

Beaucoup de modèles travaillent mieux avec des données normales et sont plus efficaces autour de $[-5, 5]$:

- centrer sur la moyenne puis diviser par l'écart-type
- transformation de Box-Cox en cas d'asymétrie
- transformations spécifiques en fonction de la distribution

Préparation des données — transformation

Quand un modèle n'accepte pas de données catégorielles :

- label encoding si ordinal
- one-hot encoding sinon

Préparation des données — label encoding

Si les données sont ordinales :

Ordinal :

Température
Froid
Froid
Tiède
Chaud
Tiède

Label encoding :

Température
1
1
2
3
2

Préparation des données — one-hot encoding

Remplacer une feature par n features avec n le nombre de catégories.

Catégoriel :

Couleur
Rouge
Rouge
Jaune
Vert
Jaune

One-hot :

Rouge	Jaune	Vert
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Exploration des données

Exploration des données

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. nettoyer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Exploration des données

But :

- se rendre compte des prétraitements à effectuer (Box-Cox, imputations, etc)

Exploration des données

But :

- se rendre compte des prétraitements à effectuer (Box-Cox, imputations, etc)
- comprendre la variable de sortie : distribution, équilibre des classes, features les plus corrélées, ...

Exploration des données

But :

- se rendre compte des prétraitements à effectuer (Box-Cox, imputations, etc)
- comprendre la variable de sortie : distribution, équilibre des classes, features les plus corrélées, ...
- détecter les corrélations

Exploration des données

But :

- se rendre compte des prétraitements à effectuer (Box-Cox, imputations, etc)
- comprendre la variable de sortie : distribution, équilibre des classes, features les plus corrélées, ...
- détecter les corrélations
- appréhender la complexité nécessaire du modèle

Exploration des données

But :

- se rendre compte des prétraitements à effectuer (Box-Cox, imputations, etc)
- comprendre la variable de sortie : distribution, équilibre des classes, features les plus corrélées, ...
- détecter les corrélations
- appréhender la complexité nécessaire du modèle

Attention : garder des données de côté (test set) et ne pas les regarder.
Sinon biais statistique énorme.

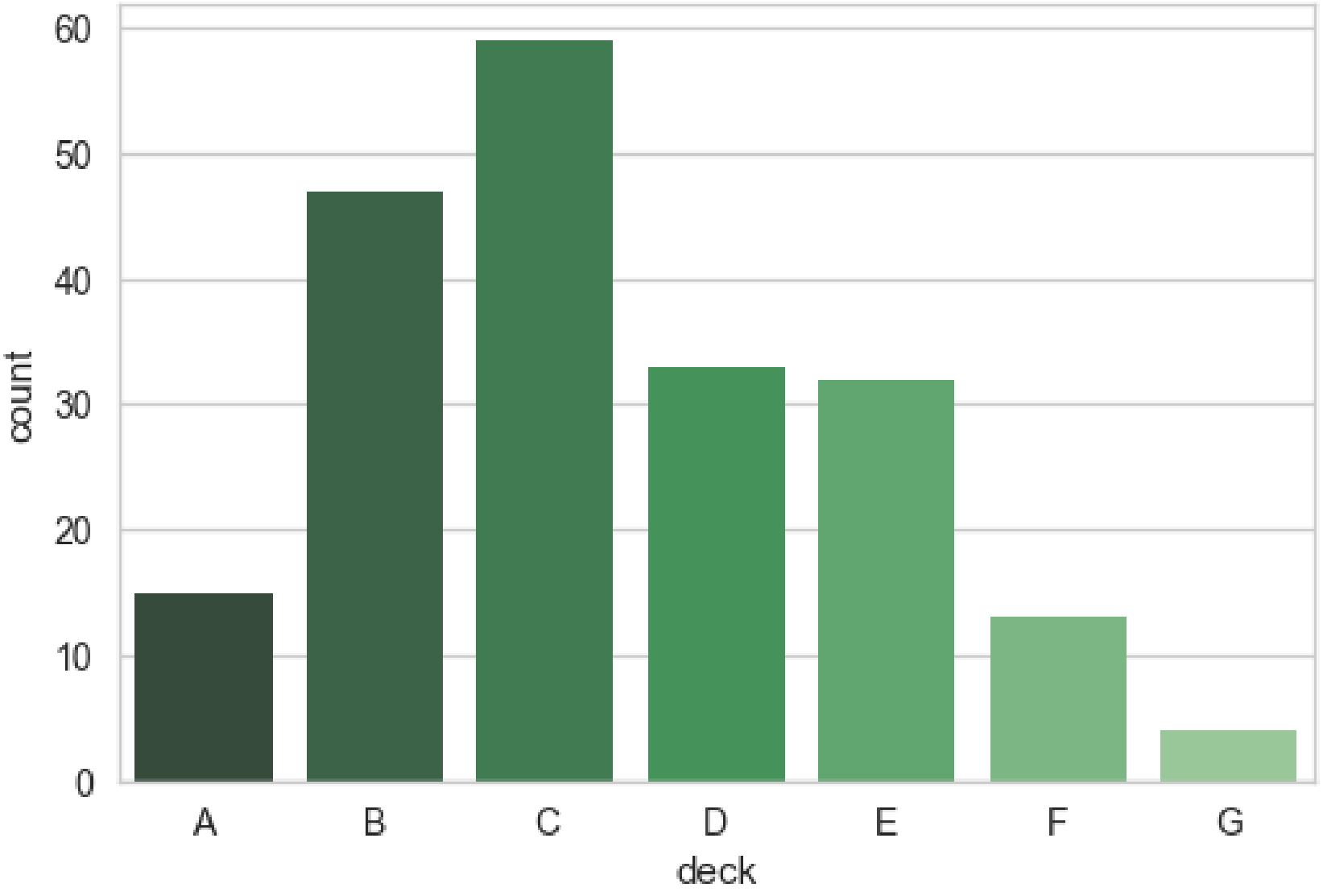
Outils

Plusieurs outils sont disponibles pour explorer des données. On utilise principalement des plots pour :

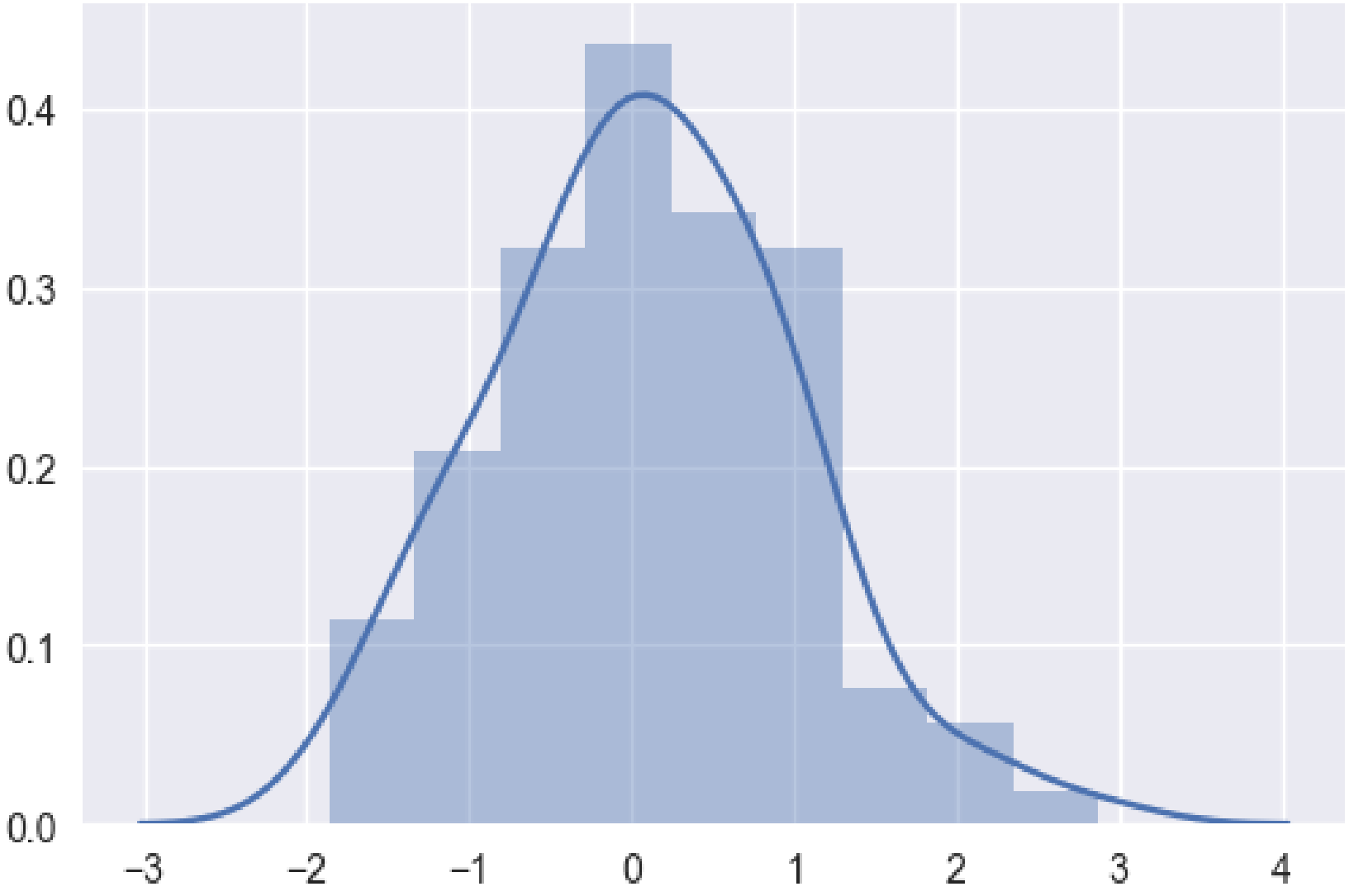
- se renseigner sur une distribution
- se renseigner sur la corrélation de deux distributions
- visualiser des corrélations linéaires

Les outils suivants sont sauf mention contraire présents dans `seaborn`.

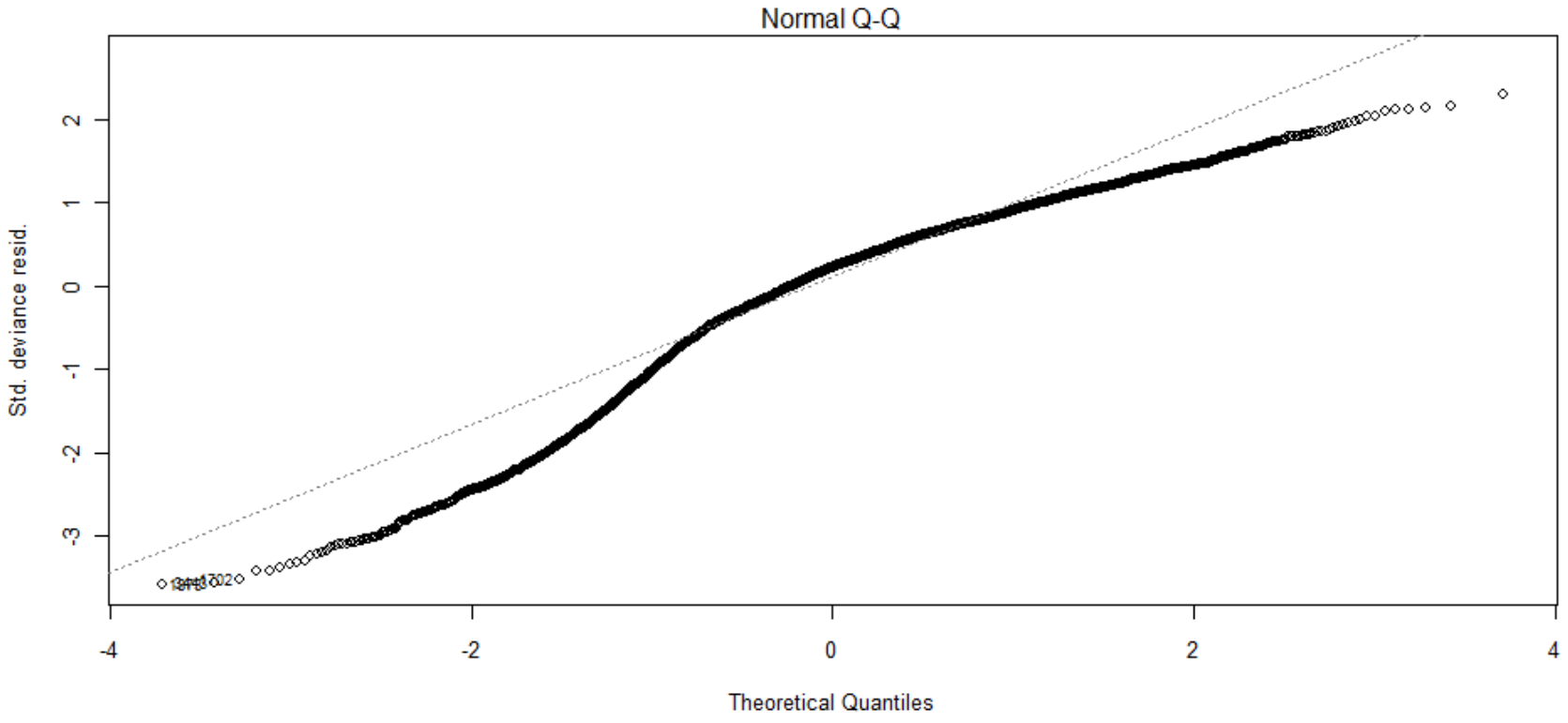
Outils — count plot



Outils — dist plot

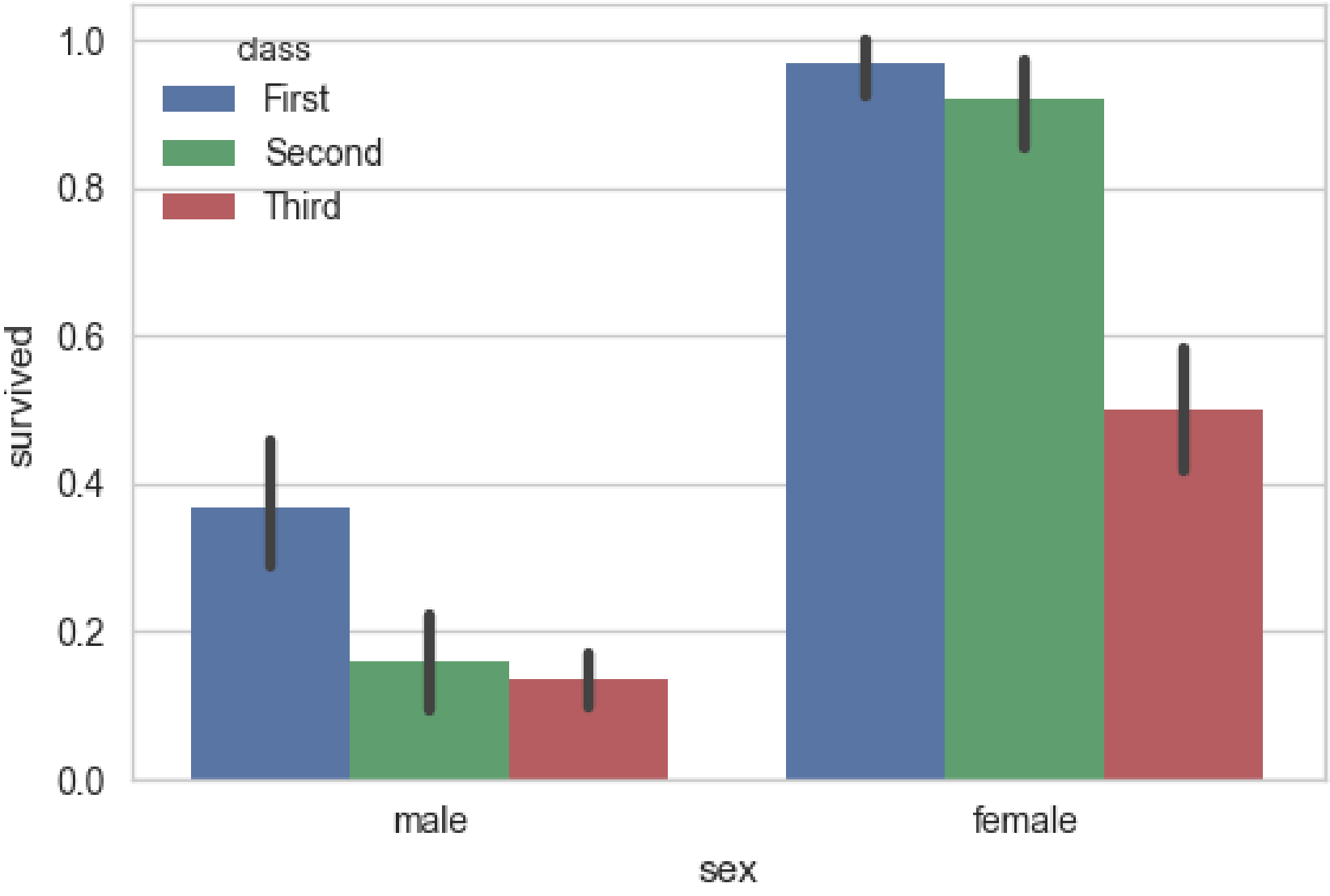


Outils — qq plot

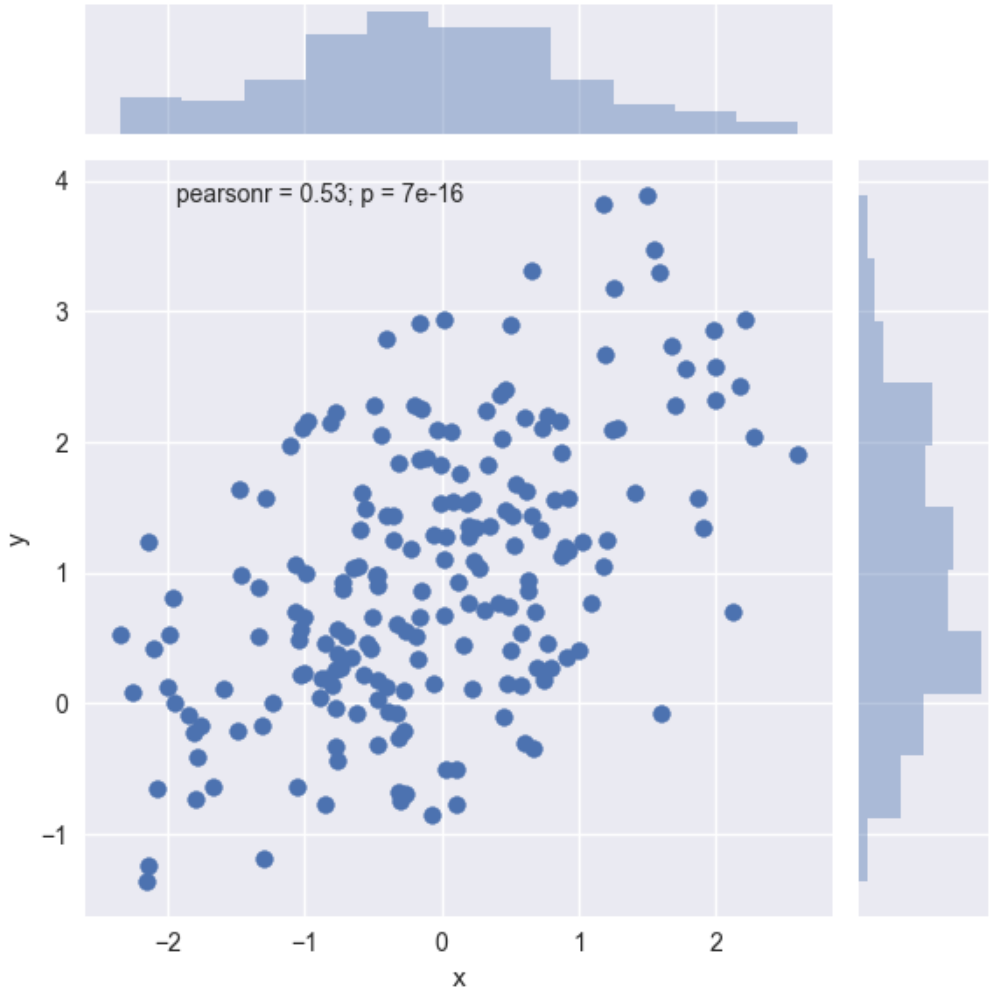


Attention, pas `seaborn` mais `statsmodel` ou `scipy.stats`.

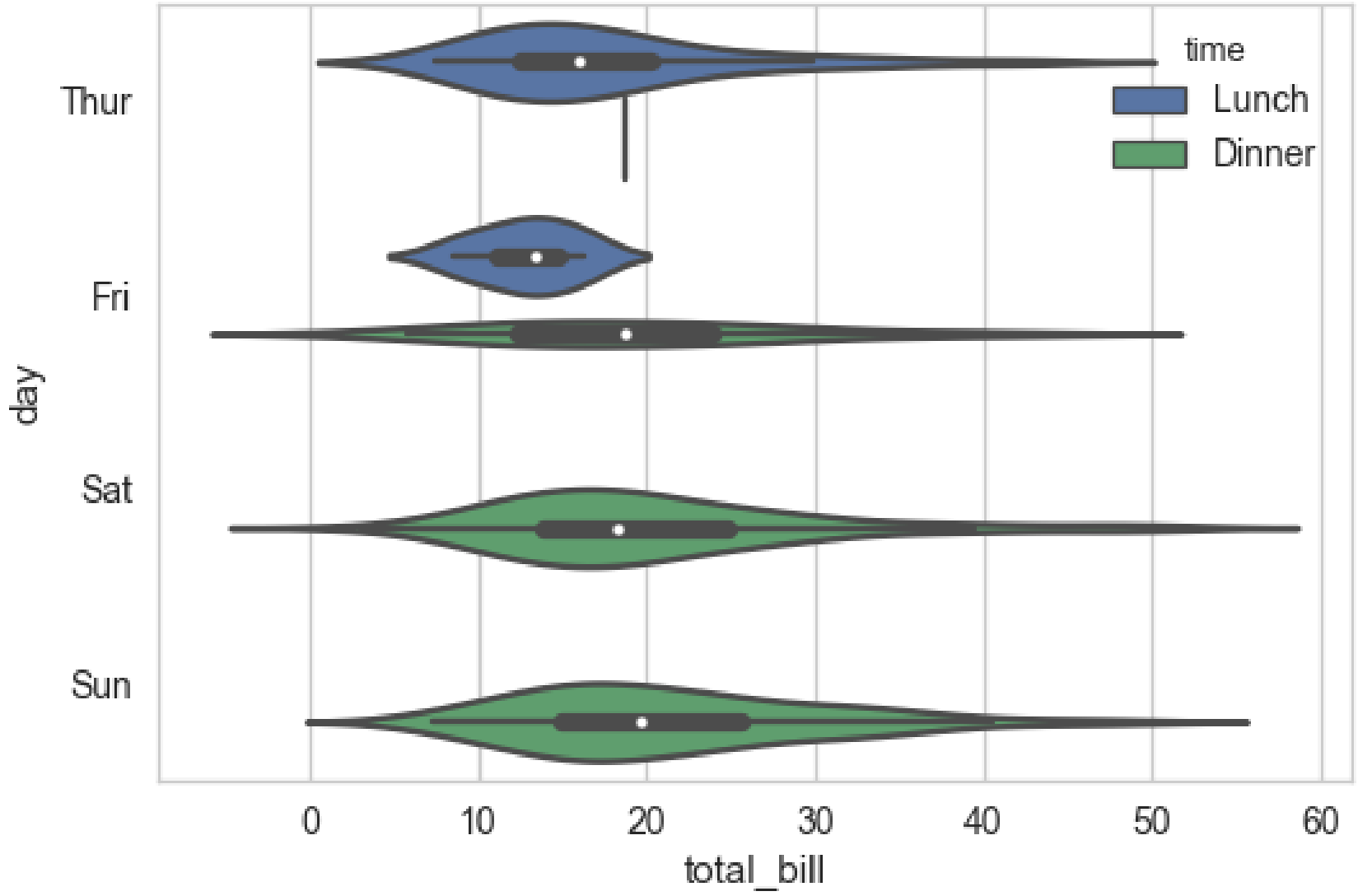
Outils — bar plot



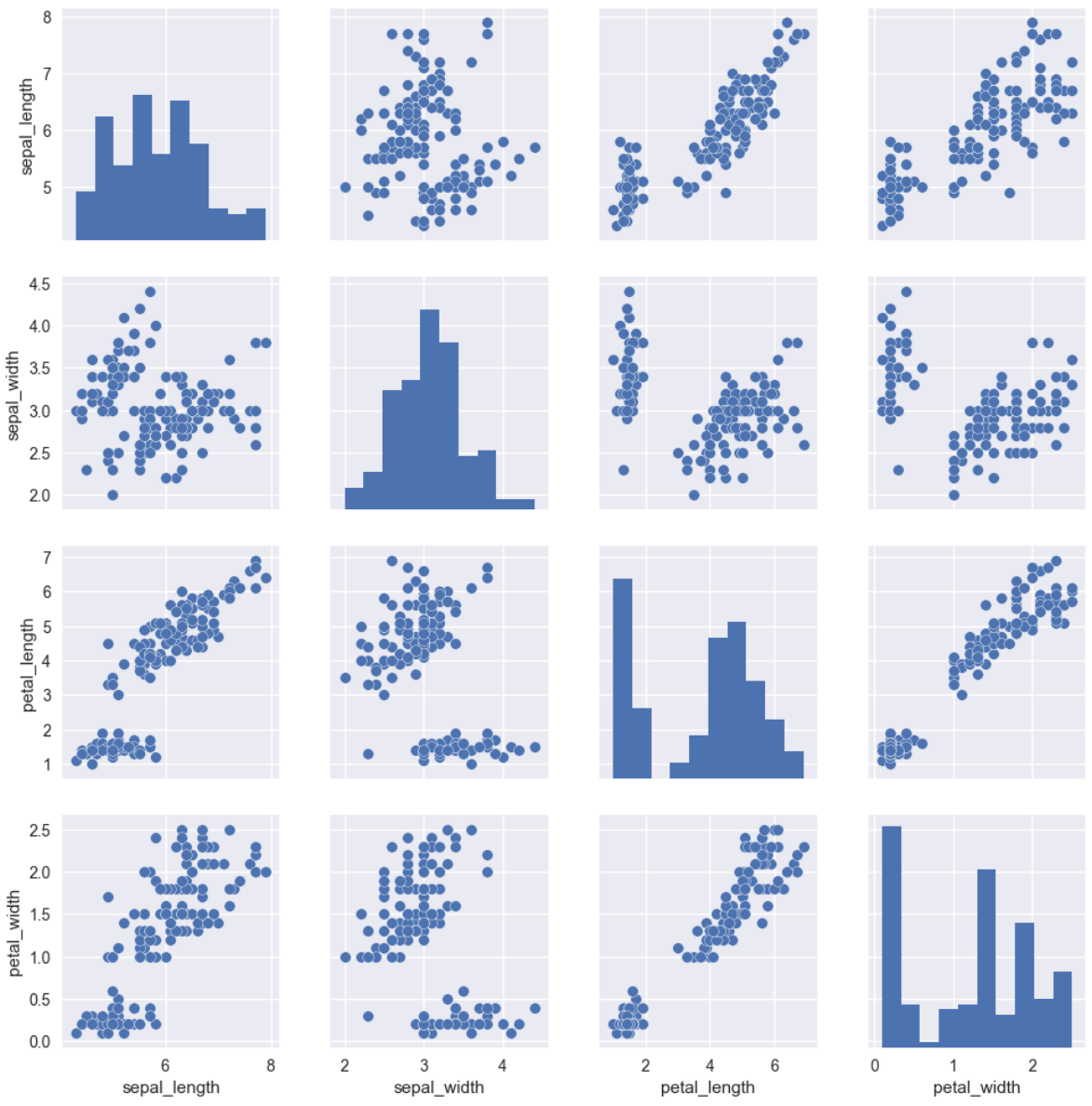
Outils — scatter plot



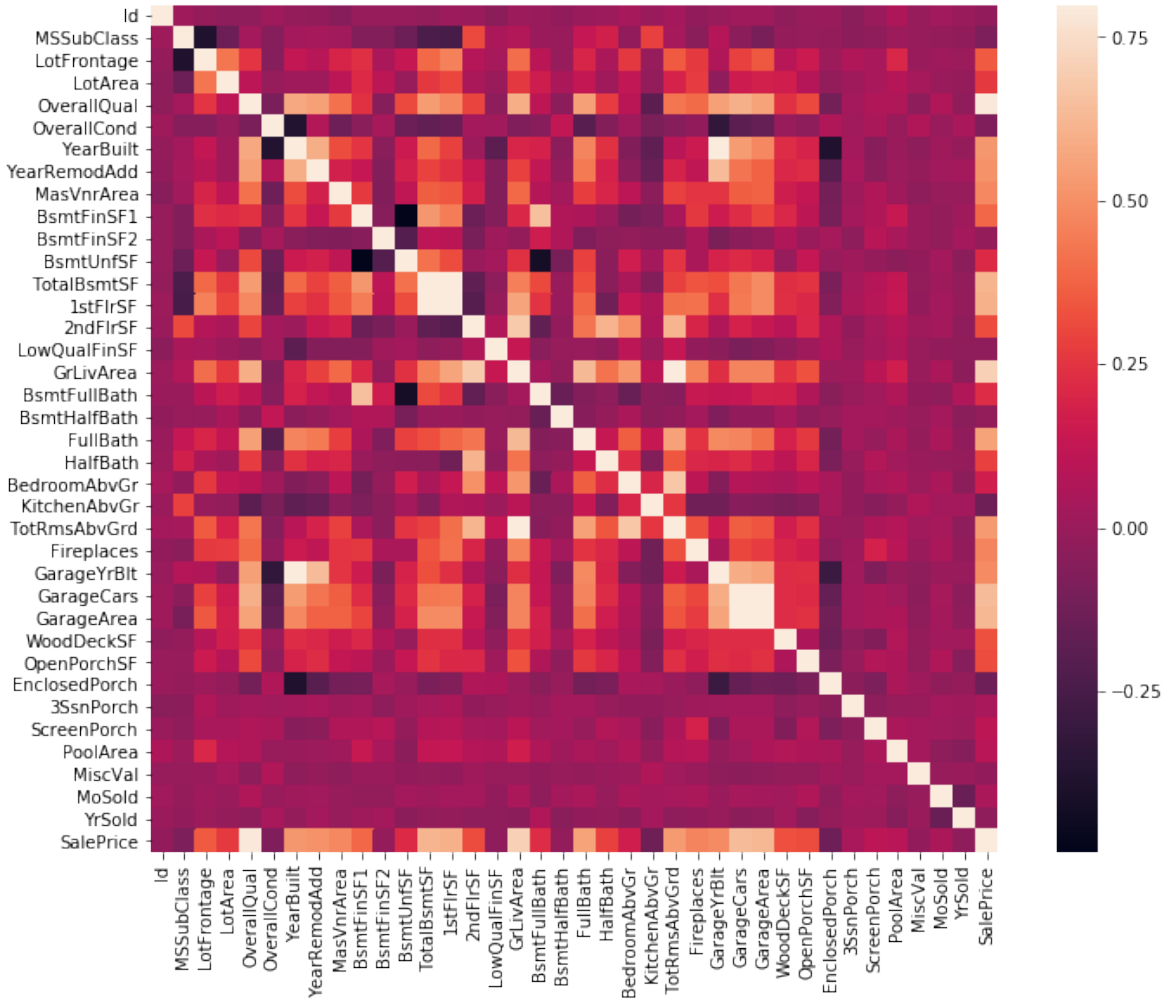
Outils — violin plot



Outils — pair plot



Outils — correlation matrix



Mode opératoire

Bonne baseline pour explorer un dataset :

- analyser la(es) variable(s) de sortie (countplot/distplot)

Mode opératoire

Bonne baseline pour explorer un dataset :

- analyser la(es) variable(s) de sortie (countplot/distplot)
- trouver les corrélations linéaires les plus fortes

Mode opératoire

Bonne baseline pour explorer un dataset :

- analyser la(es) variable(s) de sortie (countplot/distplot)
- trouver les corrélations linéaires les plus fortes
- analyser les variables correspondantes

Mode opératoire

Bonne baseline pour explorer un dataset :

- analyser la(es) variable(s) de sortie (countplot/distplot)
- trouver les corrélations linéaires les plus fortes
- analyser les variables correspondantes
- regarder s'il y a des outliers évidents dans ces variables

Entrainement d'un modèle

Entrainement d'un modèle

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. nettoyer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Types d'entrainement

- supervisé

Types d'entrainement

- supervisé
- non-supervisé

Types d'entrainement

- supervisé
- non-supervisé
- par renforcement

Entrainement supervisé d'un modèle — procédé

Étant donné des exemples d'entrainement (x_i, y_i), trouver un modèle h :

- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i

Entrainement supervisé d'un modèle — procédé

Étant donné des exemples d'entrainement (x_i, y_i), trouver un modèle h :

- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i
- moyen : définition d'une perte (loss) $L(\hat{y}_i, y_i)$

Entrainement supervisé d'un modèle — procédé

Étant donné des exemples d'entrainement (x_i, y_i), trouver un modèle h :

- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i
- moyen : définition d'une perte (loss) $L(\hat{y}_i, y_i)$

  quelle fonction pourrait-on prendre en régression ?

Entrainement supervisé d'un modèle — procédé

Étant donné des exemples d'entrainement (x_i, y_i), trouver un modèle h :

- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i
- moyen : définition d'une perte (loss) $L(\hat{y}_i, y_i)$

  quelle fonction pourrait-on prendre en régression ?

  par exemple, $L(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$

Entrainement supervisé d'un modèle — procédé

Étant donné des exemples d'entrainement (x_i, y_i), trouver un modèle h :

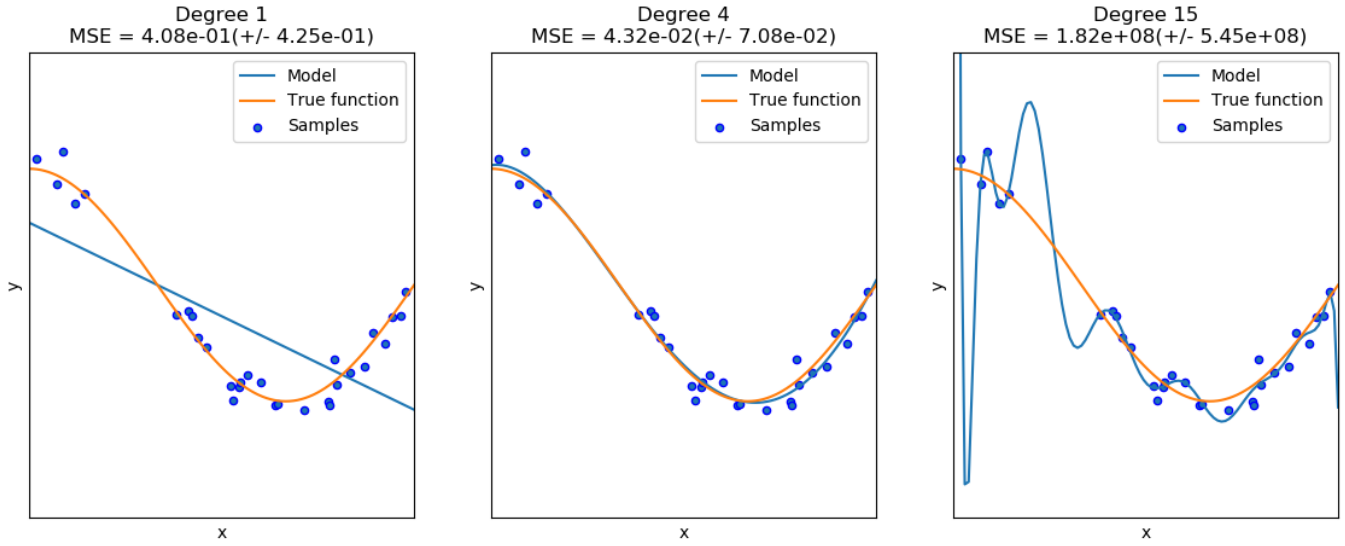
- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i
- moyen : définition d'une perte (loss) $L(\hat{y}_i, y_i)$

  quelle fonction pourrait-on prendre en régression ?

  par exemple, $L(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2$

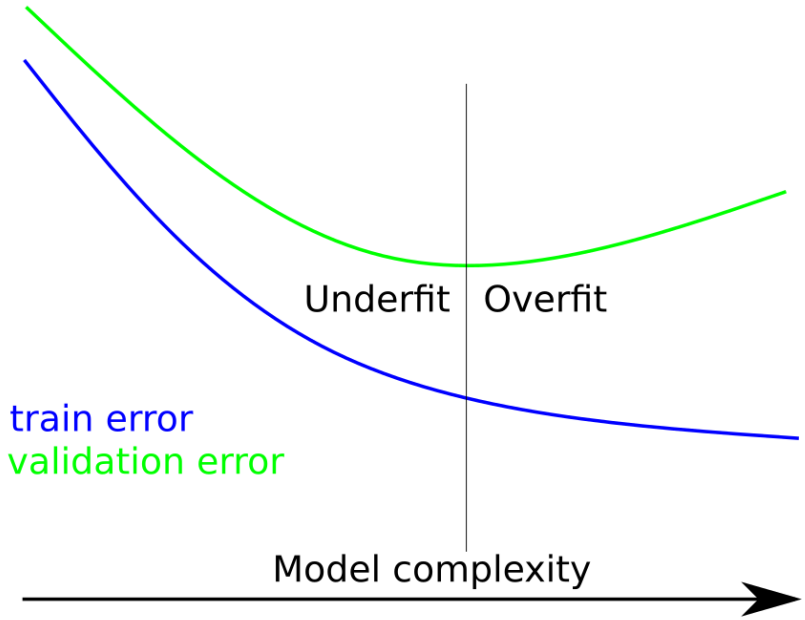
  puis minimisation

Entrainement supervisé d'un modèle — overfit



Problème : trop minimiser la perte n'est pas bon !

Entrainement supervisé d'un modèle — learning curve



→ Minimiser la perte sur un ensemble de validation

Entrainement supervisé d'un modèle — data split

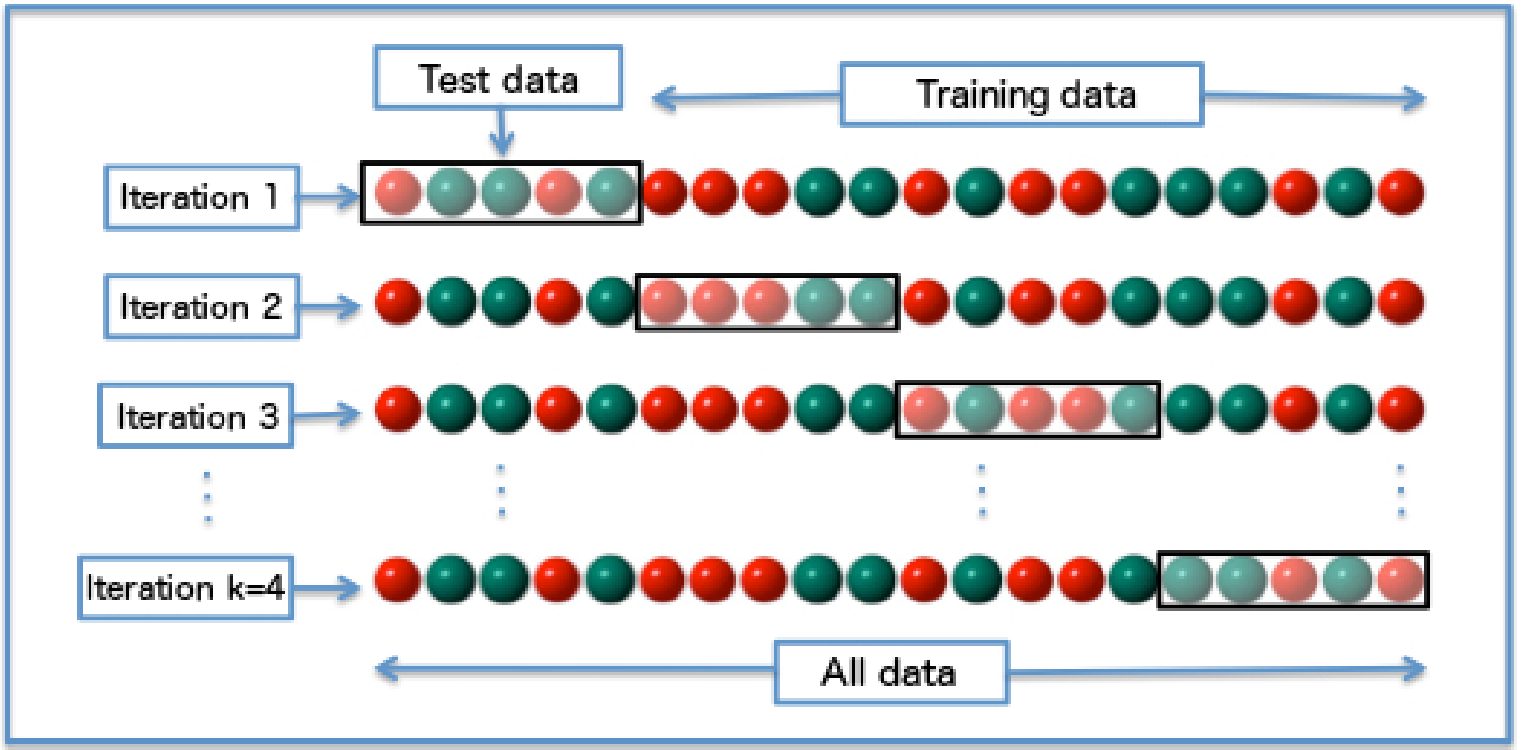
Il nous faut donc :

- ensemble d'entrainement
- ensemble de validation pour mesurer la généralisation
- ensemble de test (pour éviter le biais statistique)

→ Split 60/20/20 habituel.

Entrainement supervisé d'un modèle — cross-validation

Pour « perdre » moins de données et mieux tester la généralisation, cross-validation :



Ici, 4-fold cross-validation.

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon
  - recommandations : $y_i =$ liste d'items $x_{k \neq i}$

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon
  - recommandations : $y_i =$ liste d'items $x_{k \neq i}$
  - réduction de dimensionnalité : $y_i = x_i$ projeté dans moins de features

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon
  - recommandations : $y_i =$ liste d'items $x_{k \neq i}$
  - réduction de dimensionnalité : $y_i = x_i$ projeté dans moins de features
- on définit quand même une perte (loss)

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon
  - recommandations : $y_i =$ liste d'items $x_{k \neq i}$
  - réduction de dimensionnalité : $y_i = x_i$ projeté dans moins de features
- on définit quand même une perte (loss)

  par exemple, densité intra- et inter-clusters en clustering

Entrainement non-supervisé d'un modèle

Étant donné des exemples x_i, trouver un modèle h :

- but moins défini qu'en supervisé :
  - clustering : $h(x_i) = \hat{y}_i =$ cluster de x_i
  - détection d'anomalies : $y_i = 1$ si anomalie, 0 sinon
  - recommandations : $y_i =$ liste d'items $x_{k \neq i}$
  - réduction de dimensionnalité : $y_i = x_i$ projeté dans moins de features
- on définit quand même une perte (loss)

  par exemple, densité intra- et inter-clusters en clustering

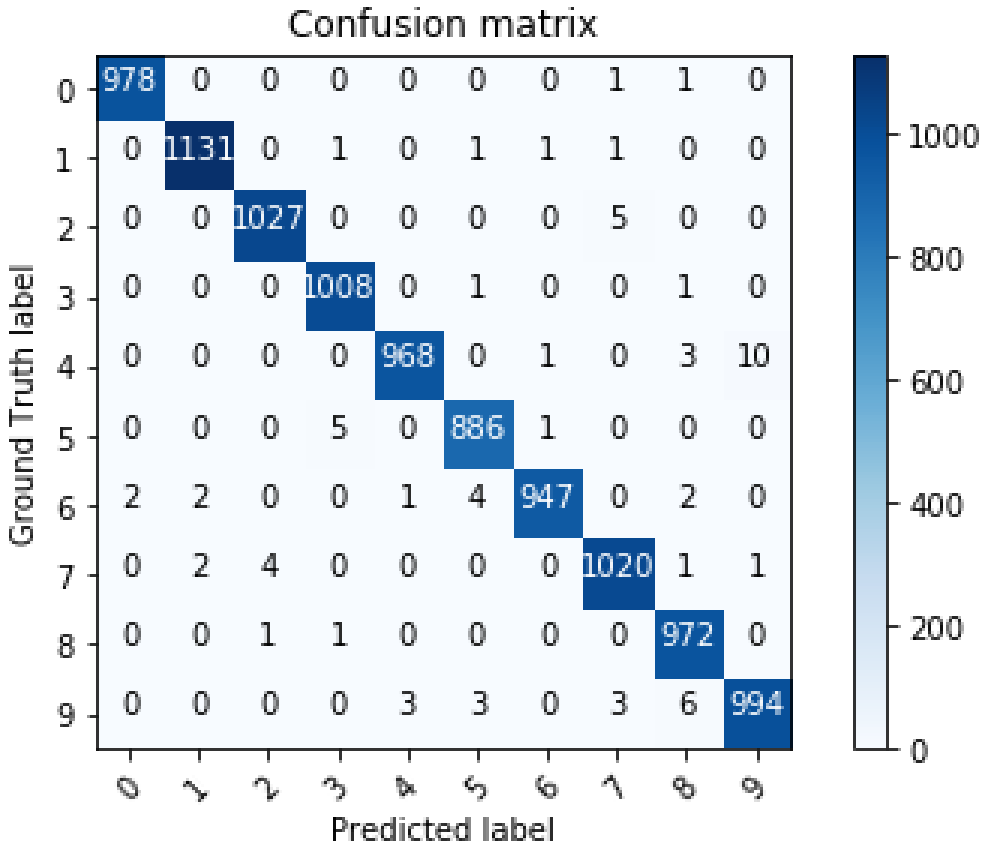
  puis minimisation

Évaluation des résultats

Évaluation des résultats

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. nettoyer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Évaluation — outils — matrice de confusion



Évaluation — outils — précision, rappel

En classification :

Précision

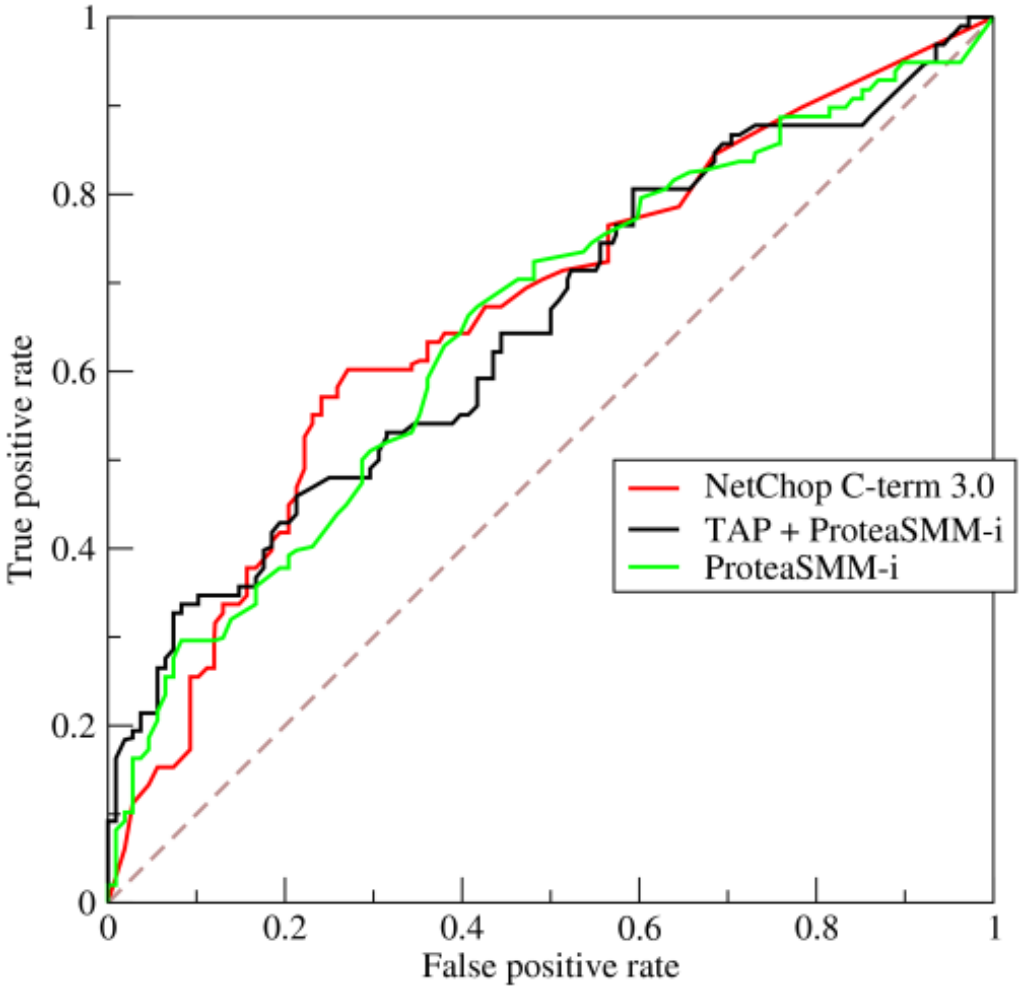
$$\frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

Rappel

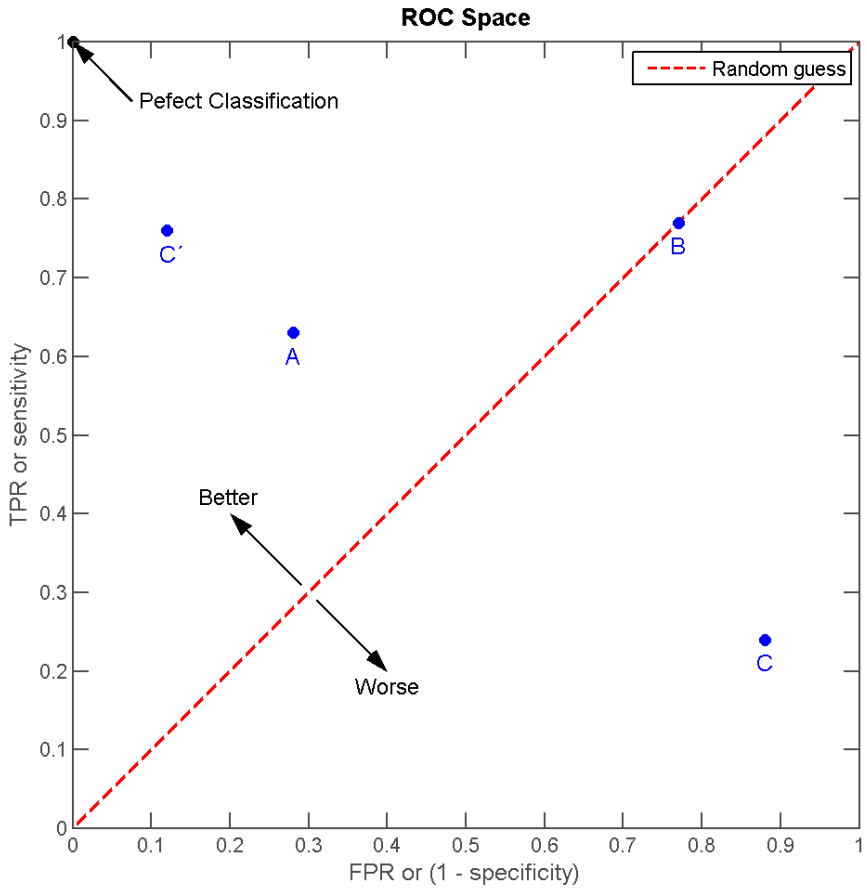
$$\frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

F-mesure moyenne harmonique entre précision et rappel (aussi appelée F1 score)

Outils — courbe ROC



Outils — courbe ROC



Reproductibilité

Reproductibilité

1. définir la question à laquelle on veut répondre
2. obtenir des données
3. nettoyer les données
4. explorer les données
5. entraîner un modèle
6. évaluer et communiquer les résultats
7. rendre son analyse reproductible

Reproductibilité

- extrêmement importante pour compléter les analyses après les retours business
- ensemble de bonnes pratiques

Reproductibilité — bonnes pratiques

- garder une trace exacte du préprocessing

Reproductibilité — bonnes pratiques

- garder une trace exacte du préprocessing
- de préférence utiliser des notebooks

Reproductibilité — bonnes pratiques

- garder une trace exacte du préprocessing
- de préférence utiliser des notebooks
- faire attention au random (utiliser des seeds)

Reproductibilité — bonnes pratiques

- garder une trace exacte du préprocessing
- de préférence utiliser des notebooks
- faire attention au random (utiliser des seeds)
- définir les datasets utilisés, dates comprises

Reproductibilité — bonnes pratiques

- garder une trace exacte du préprocessing
- de préférence utiliser des notebooks
- faire attention au random (utiliser des seeds)
- définir les datasets utilisés, dates comprises
- garder une trace de l'environnement

Conclusion

Conclusion

- attention au biais statistique
- poser une question sur laquelle on peut **mesurer** le progrès
- acquérir des données les moins biaisées possible
- explorer et nettoyer les données en tandem
- fit un modèle avec une perte adaptée
- construire des résultats significatifs
- rester reproductible

TP 1

Exploration de données

TP 1 : Exploration de Données

Cliquez sur le lien ci-dessous et ouvrez le dans colaboratory.
Une fois le notebook chargé :
Fichier > Enregistrer une copie dans Drive

www.exploration-données.ipynb