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
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
- 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”)
- 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
- 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)
- 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.

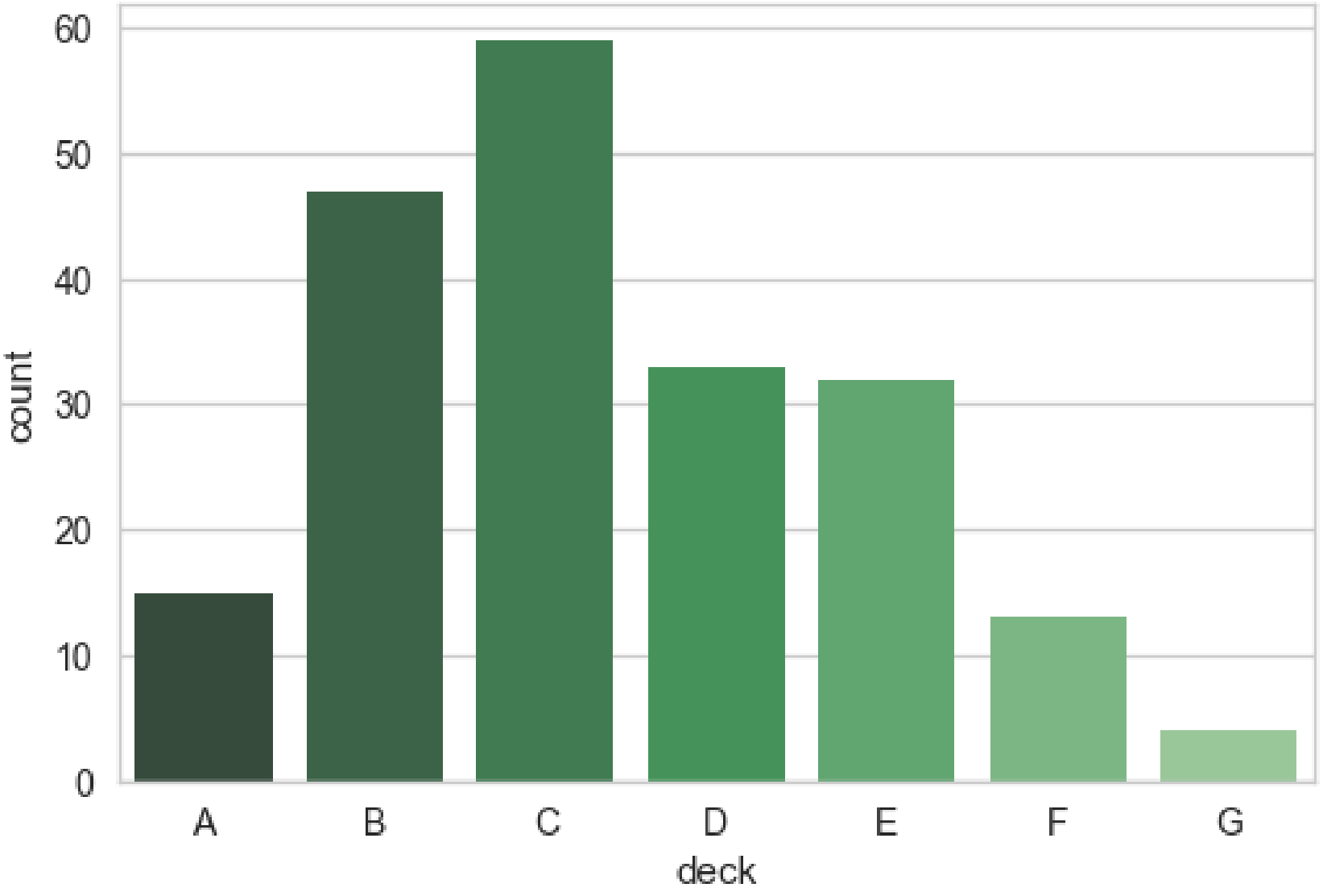
Outils

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

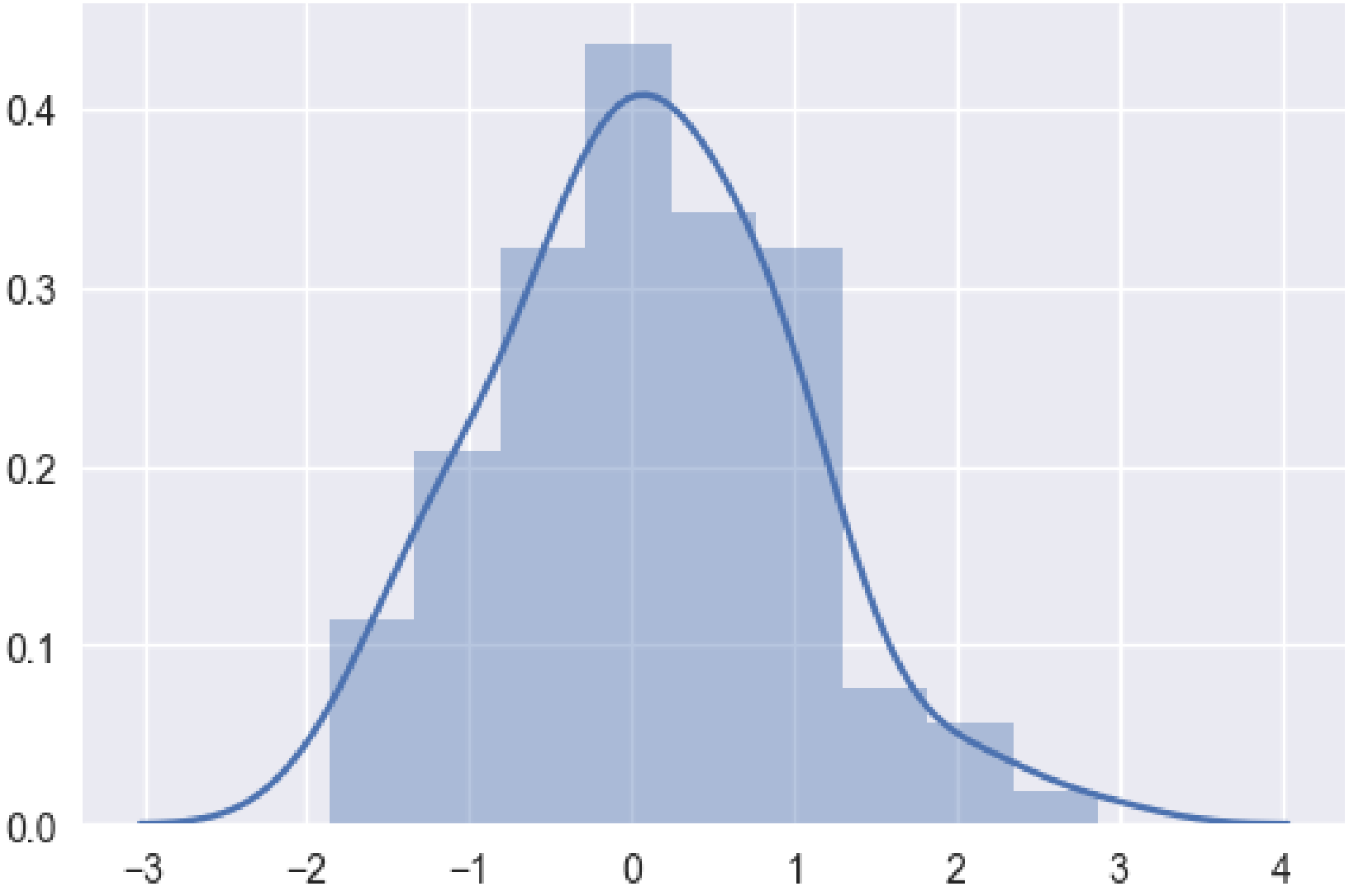
- se renseigner sur une distribution
- se renseigner sur la corrélation de deux distributions
- visualiser des corrélations linéaires

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

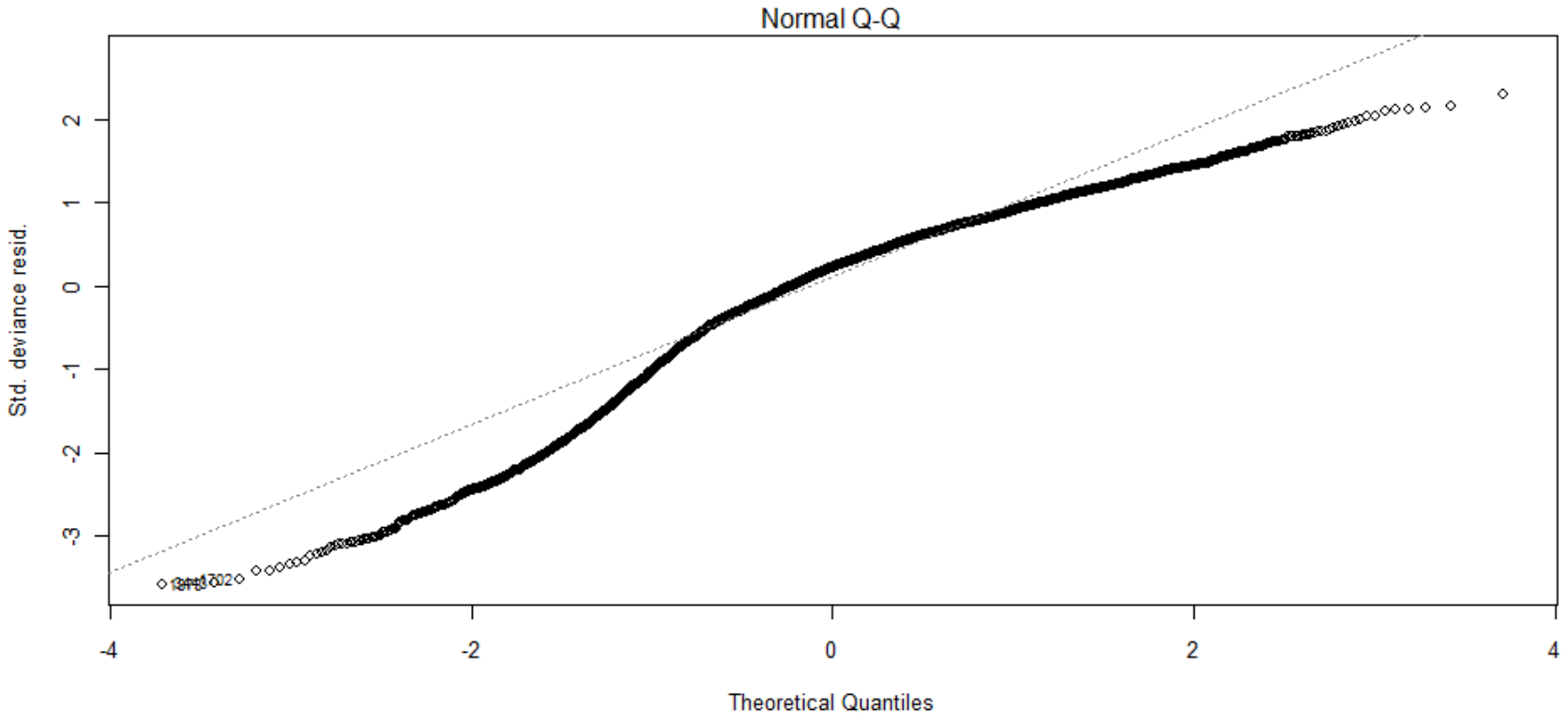
Outils — count plot



Outils — dist plot

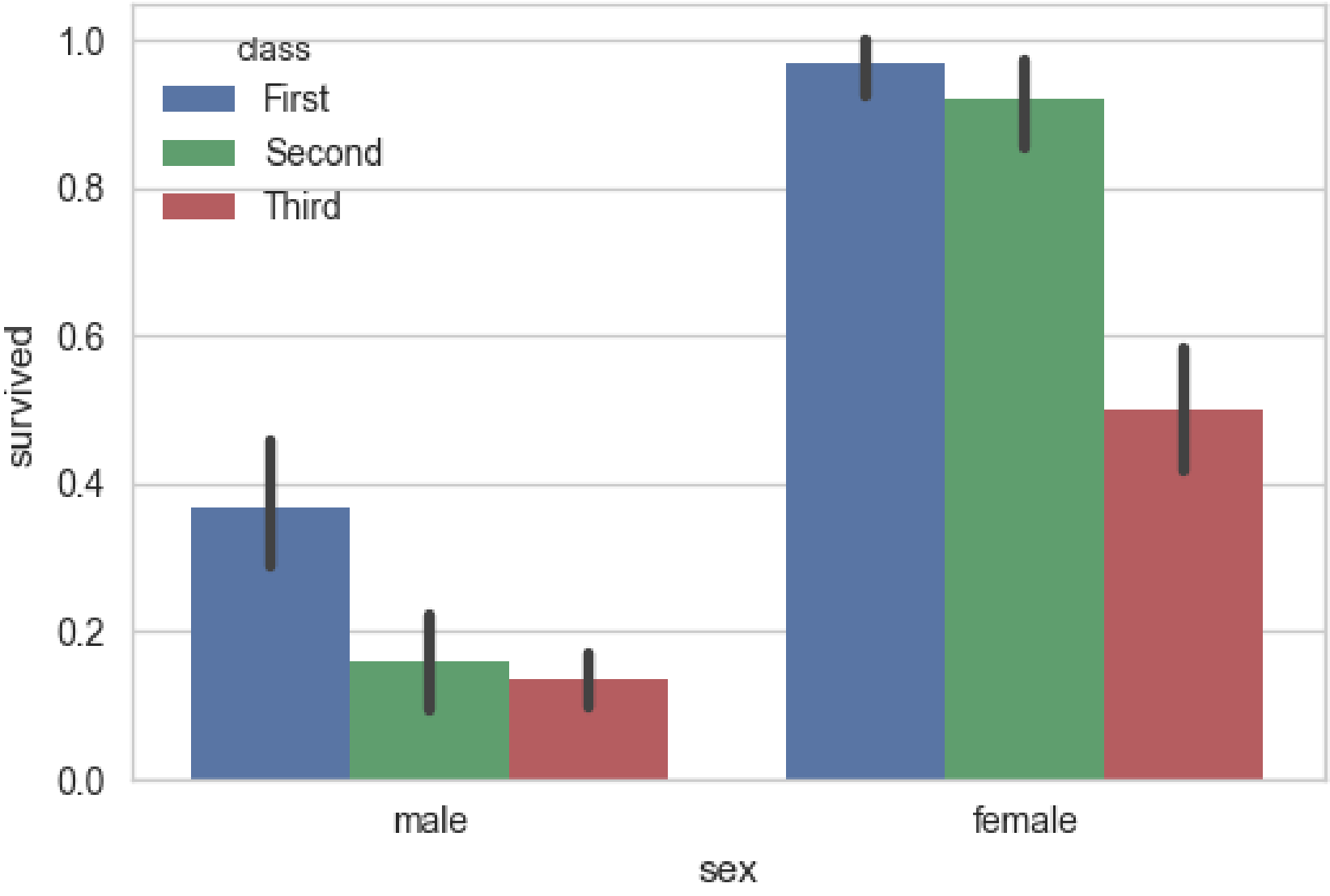


Outils — qq plot

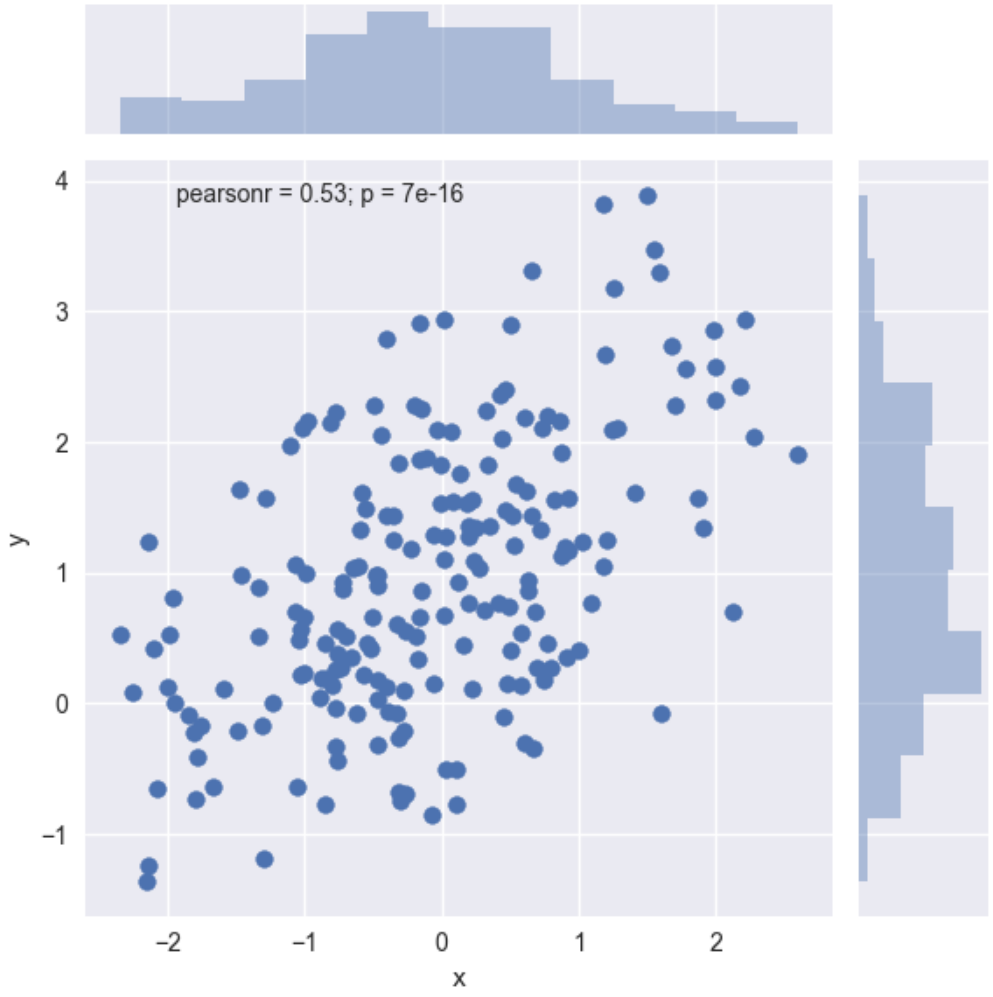


Attention, pas `seaborn` mais `statsmodel` ou `scipy.stats`.

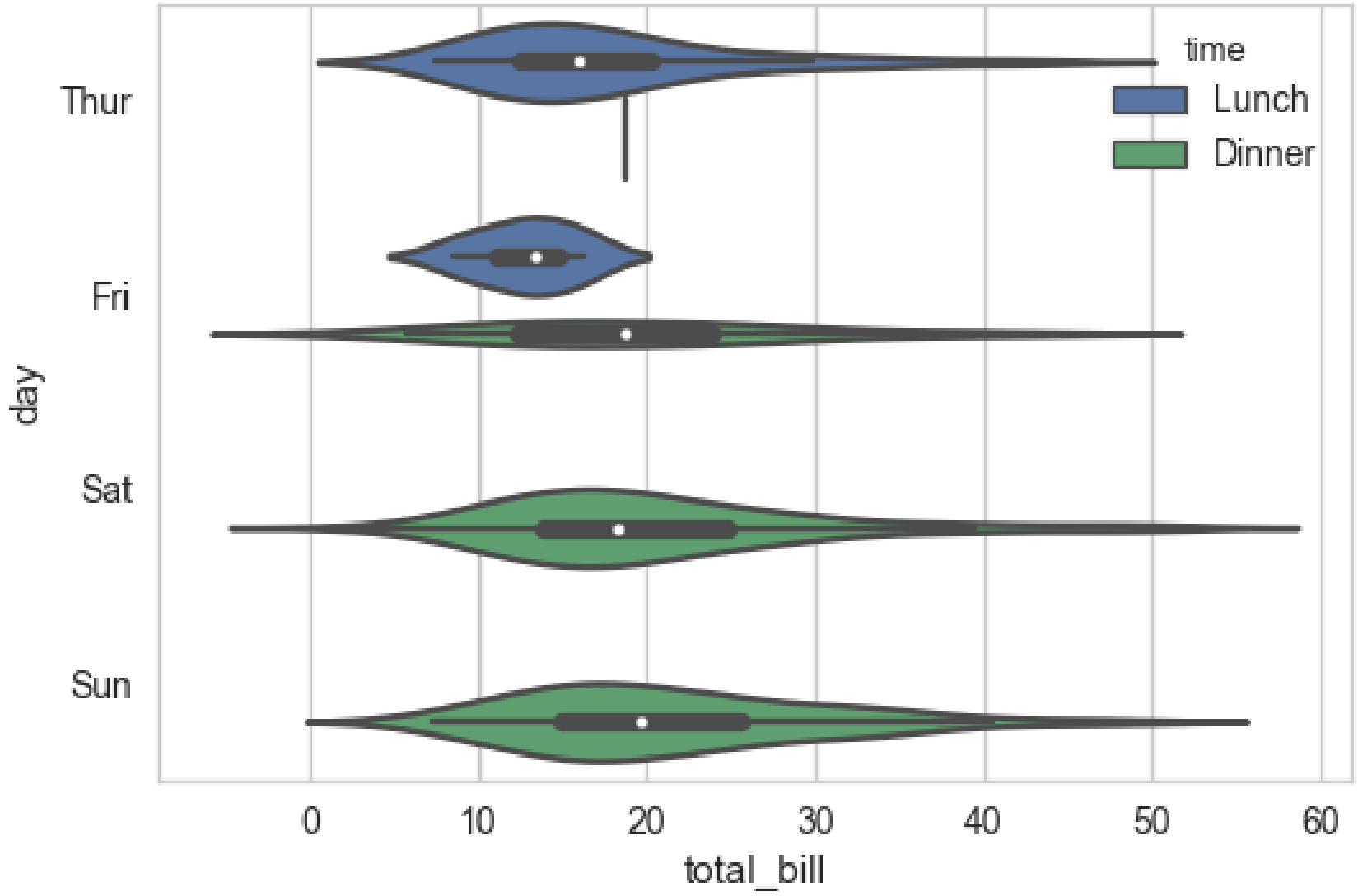
Outils — bar plot



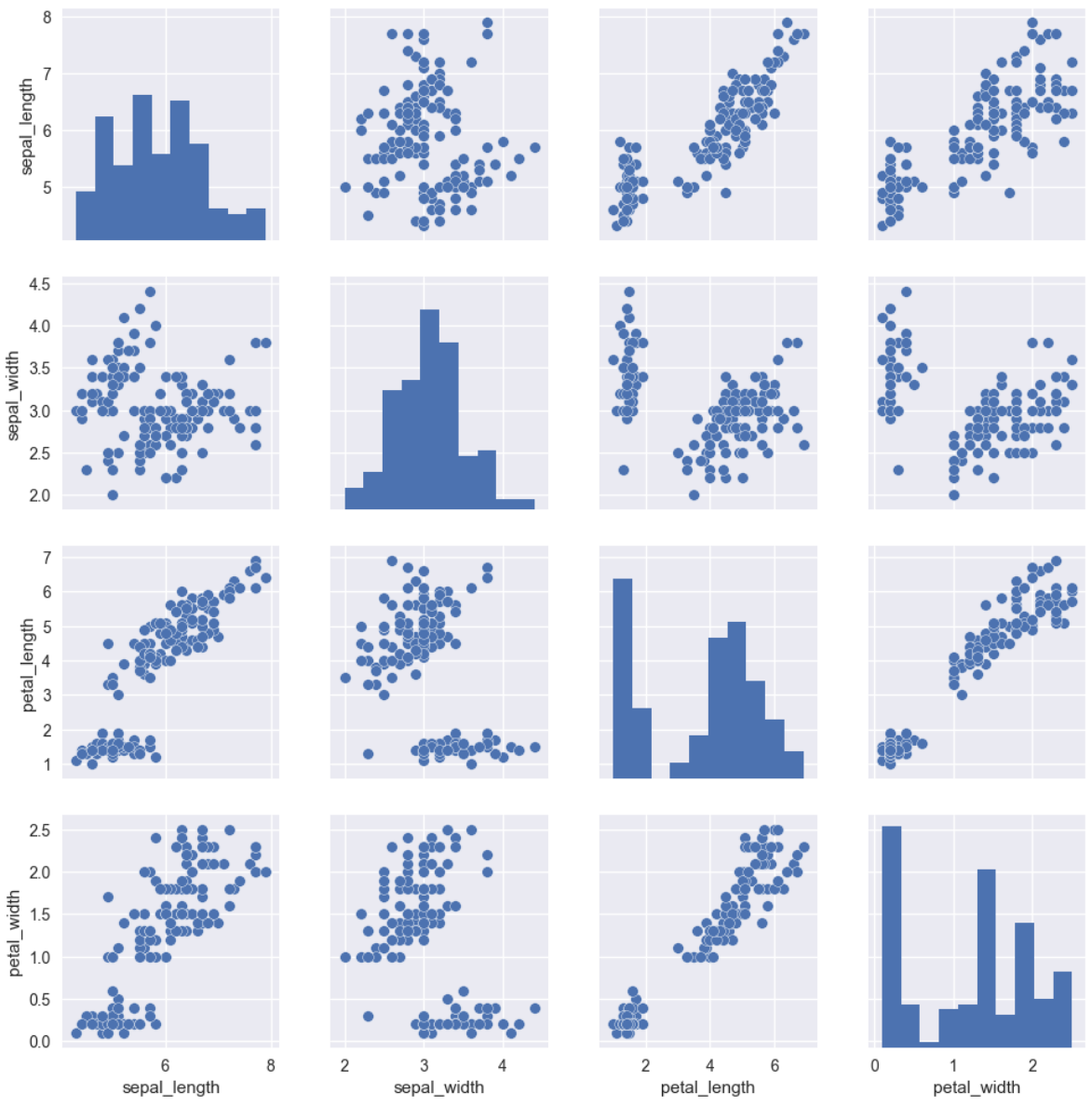
Outils — scatter plot



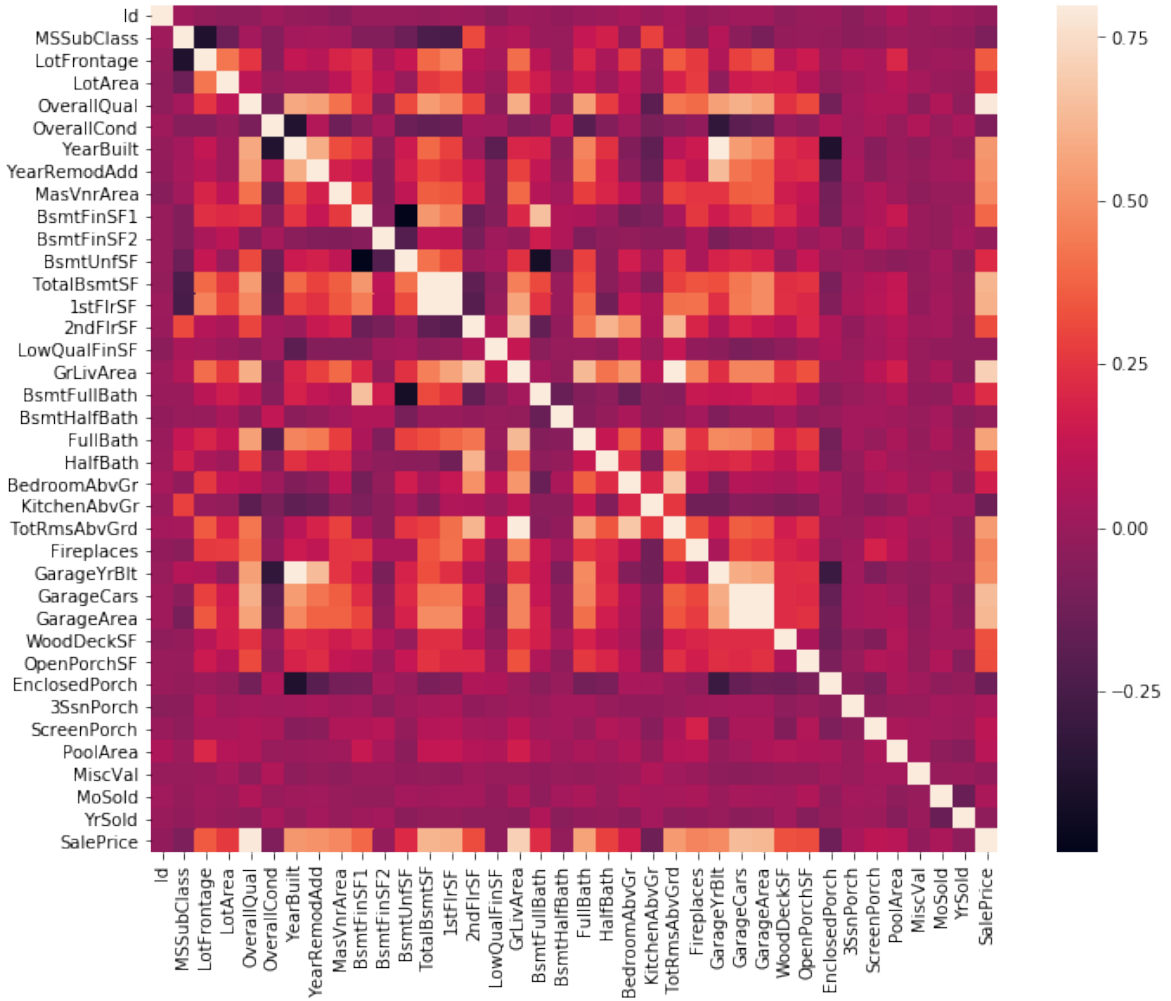
Outils — violin plot



Outils — pair plot



Outils — correlation matrix



Mode opératoire

Bonne baseline pour explorer un dataset :

- analyser la(es) variable(s) de sortie (countplot/distplot)
- 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
- non-supervisé
- par renforcement

Entrainement supervisé d'un modèle — procédé

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

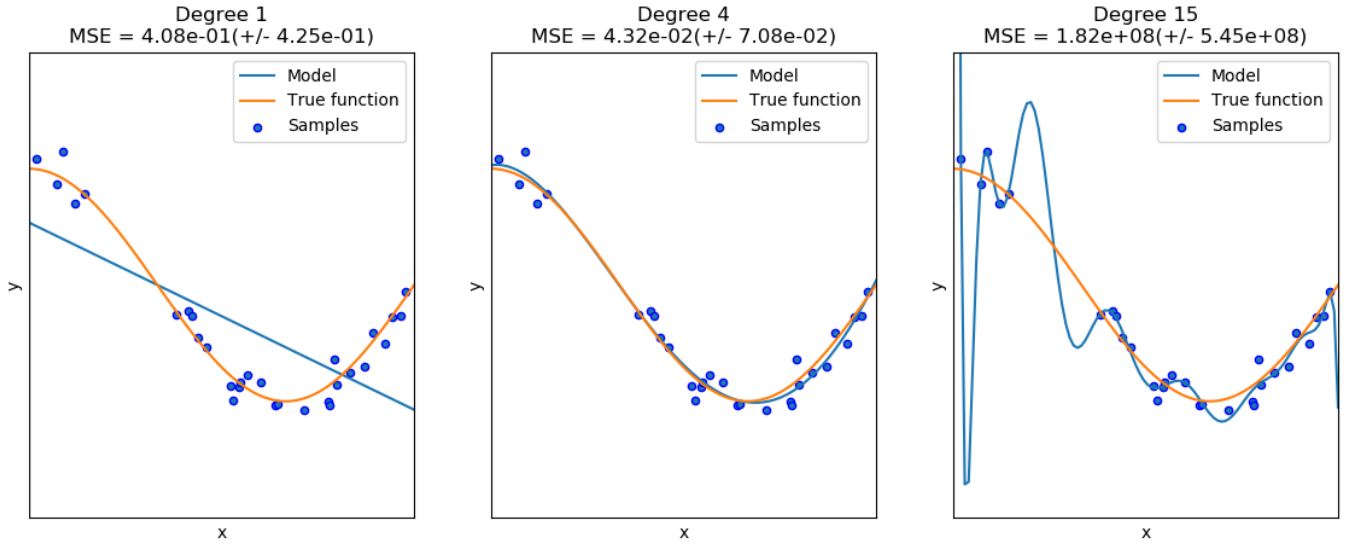
- but : étant donné x_i, output $h(x_i) = \hat{y}_i$ proche de y_i
- moyen : définition d'une perte (loss) $L(\hat{y}_i, y_i)$

  quelle fonction pourrait-on prendre en régression ?

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

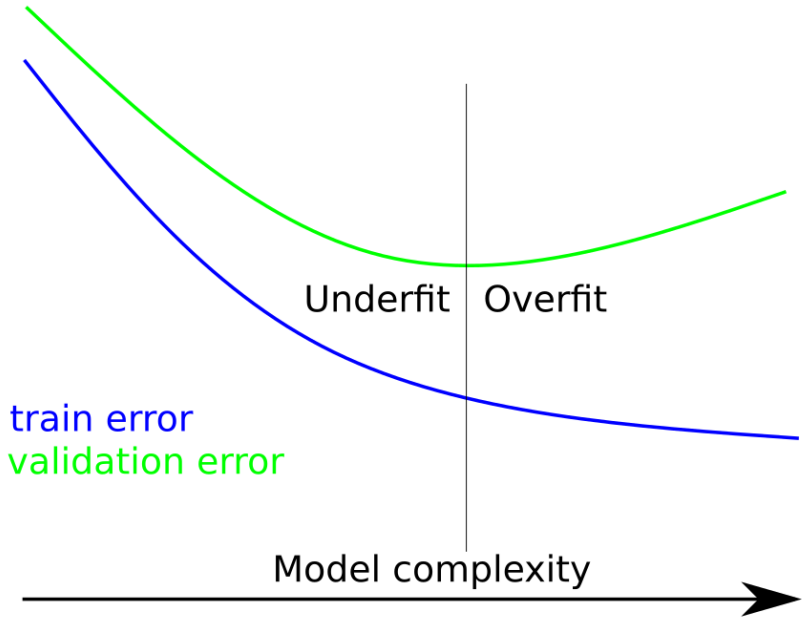
  puis minimisation

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

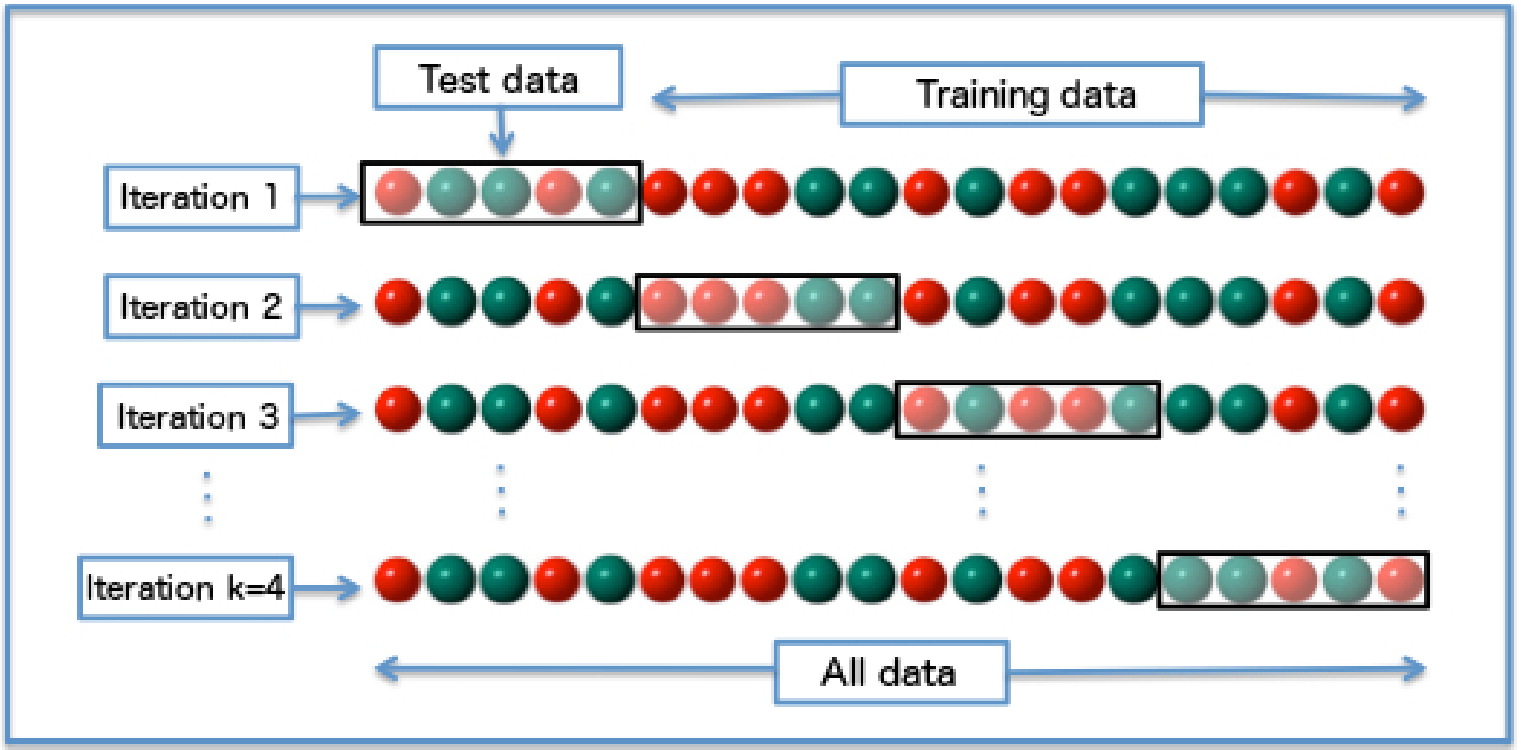
Il nous faut donc :

- 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 :
  - 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

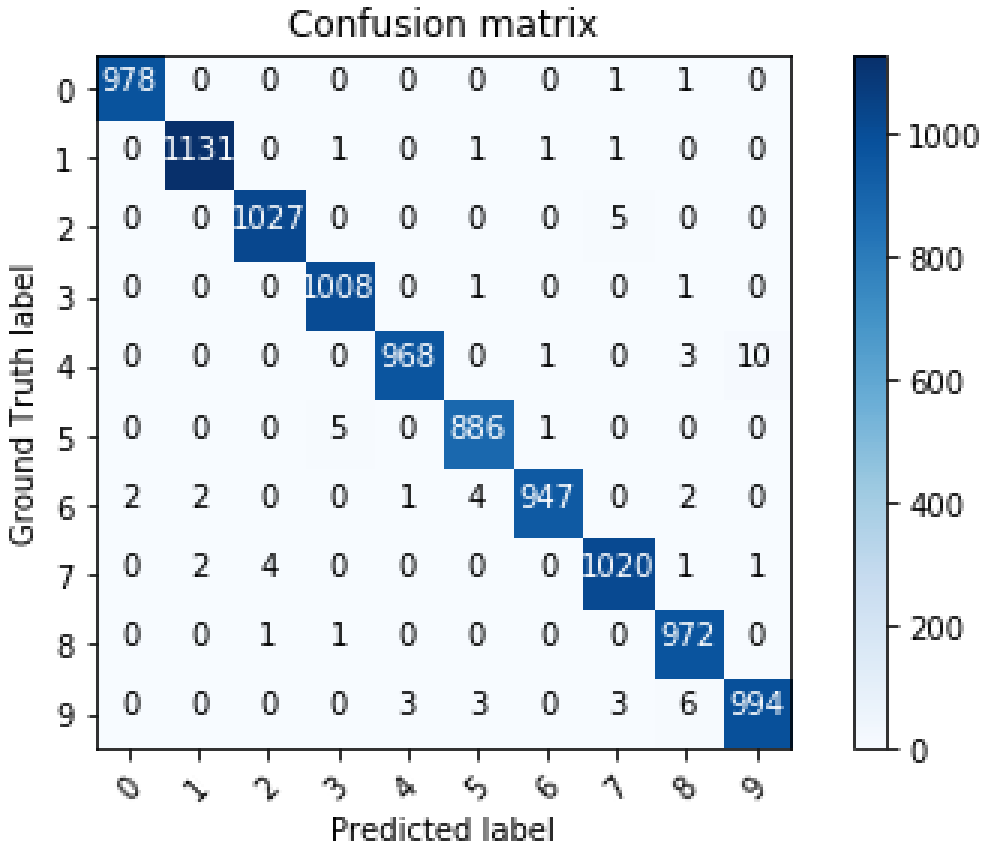
  puis minimisation

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

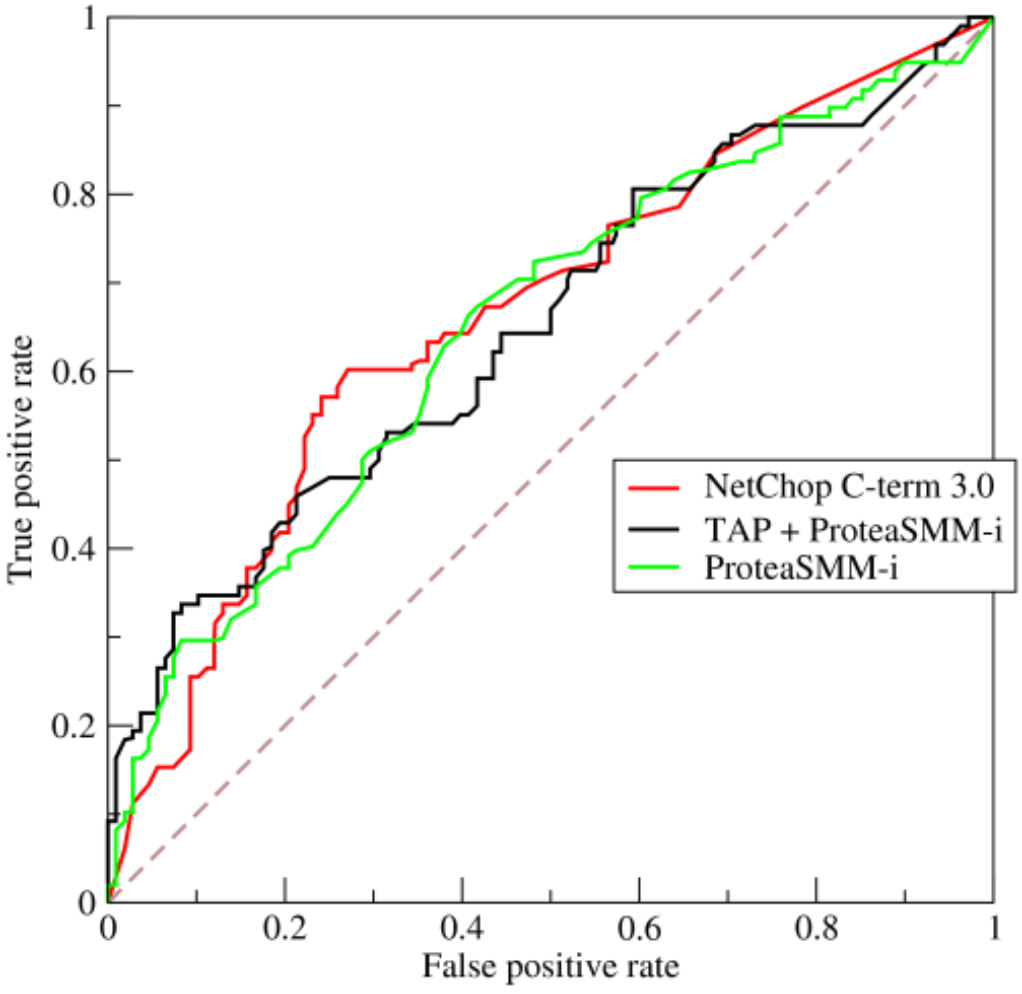
$$\frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux positifs}}$$

Rappel

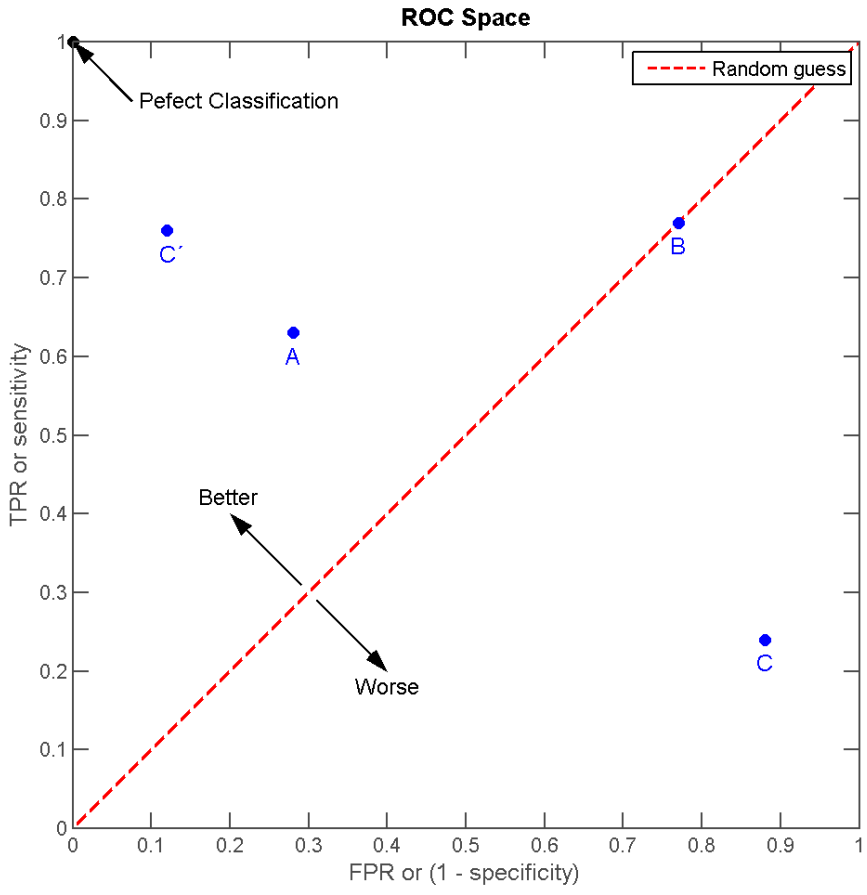
$$\frac{\text{vrais positifs}}{\text{vrais positifs} + \text{faux négatifs}}$$

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
- 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