import sklearn

track 3 - intro au Machine Learning - 13 mars 2019

Plan

1 Découvrir le Machine Learning

2 Types de problèmes

3 Préparation de la donnée

4 Modélisation

5 Validation

1 Découvrir le machine learning

- Machine Learning ou Apprentissage Automatique
- Périmètre de l’intelligence artificielle
- L’utilisation de méthodes automatiques permettant aux machines d’agir de manière systématique

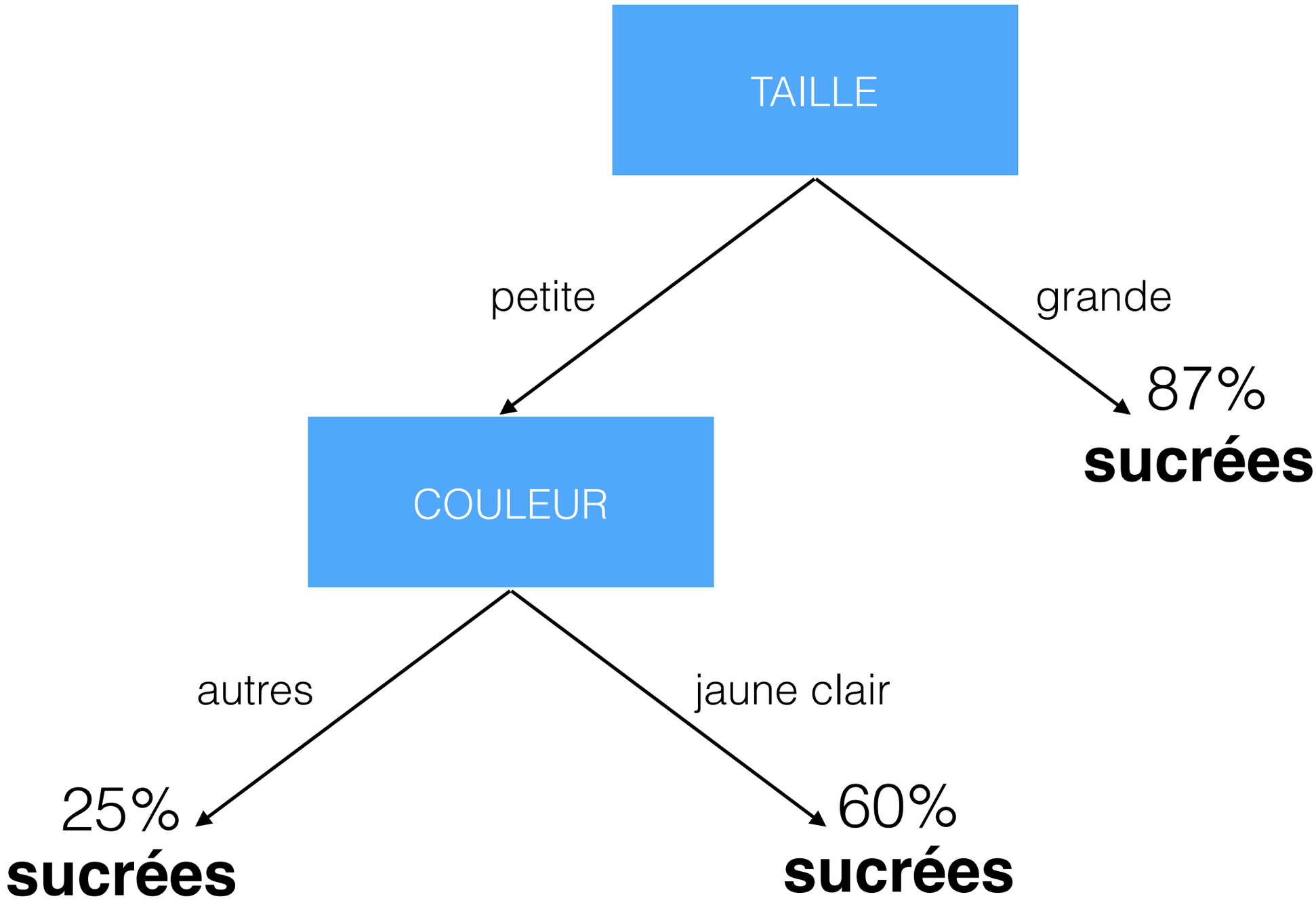
1 Découvrir le machine learning

La science des mangues

- On veut acheter des mangues sucrées (**objectif**)
- Votre grand-mère vous a dit que les mangues sucrées étaient les jaunes clair (**règles business**)
- On réalise que seulement 60% des mangues jaunes achetées sont sucrées (**performance**)
- Vous apprenez que les mangues ont des tailles différentes, donc vous en prenez des petites et des grandes, mais aussi de différentes couleurs (**échantillonnage**)
- Vous observez que dans les grosses mangues 87% sont sucrées. Vous créez une règle (**création d'une règle**)

1 Découvrir le machine learning

La science des mangues



1 Découvrir le machine learning

La science des mangues

- Votre vendeur part à la retraite et vous allez chez un autre vendeur, les mangues jaune sont décevantes (**sur-apprentissage**)

- Vous répétez donc votre expérience précédente et concluez que les petites rouges sont les meilleures (**ré-apprentissage**)

- Votre meilleur ami n'aime pas les mangues sucrées il préfère les juteuses (**nouveaux objectifs**)

- Vous vous mariez, et votre femme n'aime pas les mangues, elle préfère les pommes mais elle veut utiliser toutes l'experience sur les mangues pour trouver les meilleurs pommes (**périmètre projet**)

1 Découvrir le machine learning

La science des mangues

- Vous faites donc un Doctorat en science de la mangue
- Vous apprenez qu'il y a 400 sortes de mangues mais en France vous pouvez en acheter seulement 40 (**généralisation**)
- Vous achetez des mangues un peu partout (**données d'apprentissage**)
- Vous créez une table représentant les mangues : couleur, pays, taille, forme, magasin (**caractéristique - feature**)
- Vous remarquez que certaines variables peuvent être intéressantes comme : la date (jour d'achat, saison), l'emballage, les conditions météo, le type de magasin (**feature engineering**)
- Vous notez chaque mangue par douceur, jus, acidité, maturité (**objectif**)

1 Découvrir le machine learning

La science des mangues

- Vous utilisez un langage comme R ou Python pour créer un modèle de classification trouvant la corrélation entre les features et les objectifs (**modélisation**)

- Chaque fois que vous retournez au magasin vous testez votre prédiction (**test**)

- Vous entrainez un arbre de décision sur scikit-learn et réalisez que vous avez trop de règles et que le modèle ne fonctionne pas sur les nouvelles mangues (**sur-apprentissage**)

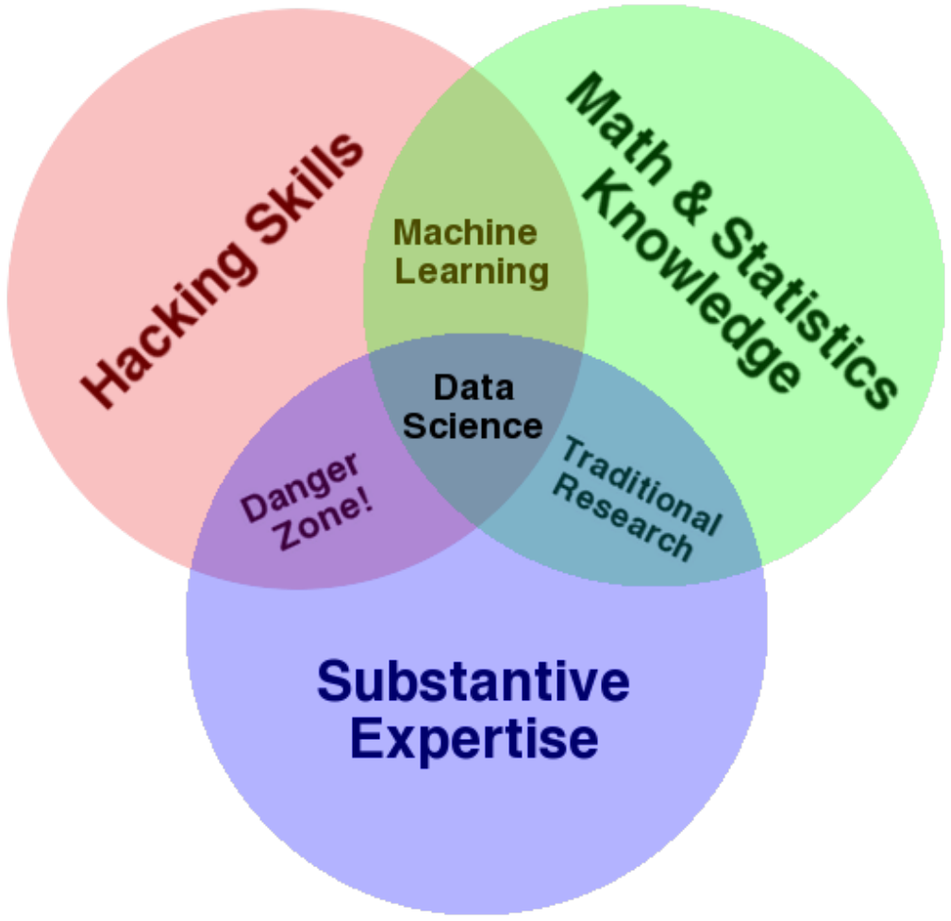
➡ Vous ne pourrez jamais tout tester, vous devez **généraliser** pour les variétés que vous n'avez pas essayé

* exemple traduit de Koby Karp

1 Découvrir le machine learning

La Data Science

- La datascience est le métier qui est à la croisée de différents chemins



2 Types de problèmes

- Il existe plusieurs types de problèmes en machine learning :
    - l'apprentissage supervisé
    - l'apprentissage non-supervisé
    - l'apprentissage par renforcement
- Pour aller plus loin, le deep learning

2 Types de problèmes

Apprentissage supervisé

- Dans l'apprentissage supervisé les données sont étiquetées, et donc on essaye de produire des règles automatiques à partir de ce que l'on connait

REGRESSION

Dans le cas de la regression nous allons **prédire** des **données continues**

CLASSIFICATION

Dans le cas de la classification nous allons **prédire des classes**, les classes peuvent être **binaires**, ou **multiples**

2 Types de problèmes

Apprentissage non-supervisé

- Dans l'apprentissage non-supervisé les données ne sont pas étiquetées, nous essayons de trouver des sous-groupes homogène dans la donnée

CLUSTERING

Méthode visant à diviser la donnée
en sous-groupe homogènes

2 Types de problèmes

Exemples

- Prédire le prix de vente d'une voiture — REGRESSION
- Prédire les défauts d'hypothèque d'une personne — CLASSIFICATION
- Grouper des chansons par genre — CLUSTERING
- Prédire le total de neige pour la semaine prochaine — REGRESSION
- Détecter les fraudeurs financiers — CLUSTERING, CLASSIFICATION
- Segmenter les visiteurs d'un site — CLUSTERING
- Prédire le moment d'achat d'un utilisateur — REGRESSION

REGRESSION CLUSTERING CLASSIFICATION

2 Types de problèmes

- Pour traiter notre données et faire la modélisation nous allons devoir la préparer

- La préparation couvre les sujets suivants :

  - quelles features dois-je utiliser ?

  - sous quel format ?

  - comment traiter mes données ?

  - quelle méthode utiliser pour valider la performance de mon modèle

3 Préparation de la donnée

Type de donnée

Catégorique

La donnée fait partie d'une liste de valeur

Exemples :

- Sexe
- Nationalité

Numérique

La donnée est une valeur numérique

Exemples :

- Age
- Prix
- Taille

Text

C'est de la donnée brute en format texte

Exemples :

- Description d'article
- Message
- Titre

3 Préparation de la donnée

Rappel

- Catégorique ⟶ Classification
- Numérique ⟶ Régression

3 Préparation de la donnée

“Binariser” la donnée

- Les modèles de régression fonctionnent bien avec des valeurs numériques pour l’entrainement

➡ On doit convertir les données catégoriques en numérique

- Pour cela on **dummy** ou **rendre binaire** l’information

- `pd.get_dummies()`

3 Préparation de la donnée

“Binariser” la donnée

```
data.head()
```

	date	channel	visits
0	2015-12-01	Affiliates	0.0
1	2015-05-01	Application	0.0
2	2015-06-01	Application	0.0
3	2016-03-01	Application	0.0
4	2015-04-01	Direct Load	2546.0

```
pd.get_dummies(data['channel']).head()
```

	Affiliates	Application	Direct Load	Display	Emailing	Internal	Natural Search	Other Websites	Paid Search	Partnerships	QR code	Social Media
0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

3 Préparation de la donnée

Binning

- Parfois nous avons des données numériques que l'on veut transformer en données catégoriques

- On va créer des ranges dans lequel on va associer des données

```
data
```

	name	age
0	Alfred	8
1	Maxime	17
2	Lucie	29
3	Caroline	31
4	Pierre	11
5	Marie	23
6	Marion	16
7	Christian	47
8	Giselle	70
9	Luc	81

```
bins = [0, 10, 18, 70, 150]
labels = ['baby', 'child', 'adult', 'senior']
```

```
data['categories'] = pd.cut(data['age'], bins, labels=labels)
```

```
data
```

	name	age	categories
0	Alfred	8	baby
1	Maxime	17	child
2	Lucie	29	adult
3	Caroline	31	adult
4	Pierre	11	child
5	Marie	23	adult
6	Marion	16	child
7	Christian	47	adult
8	Giselle	70	adult
9	Luc	81	senior

3 Préparation de la donnée

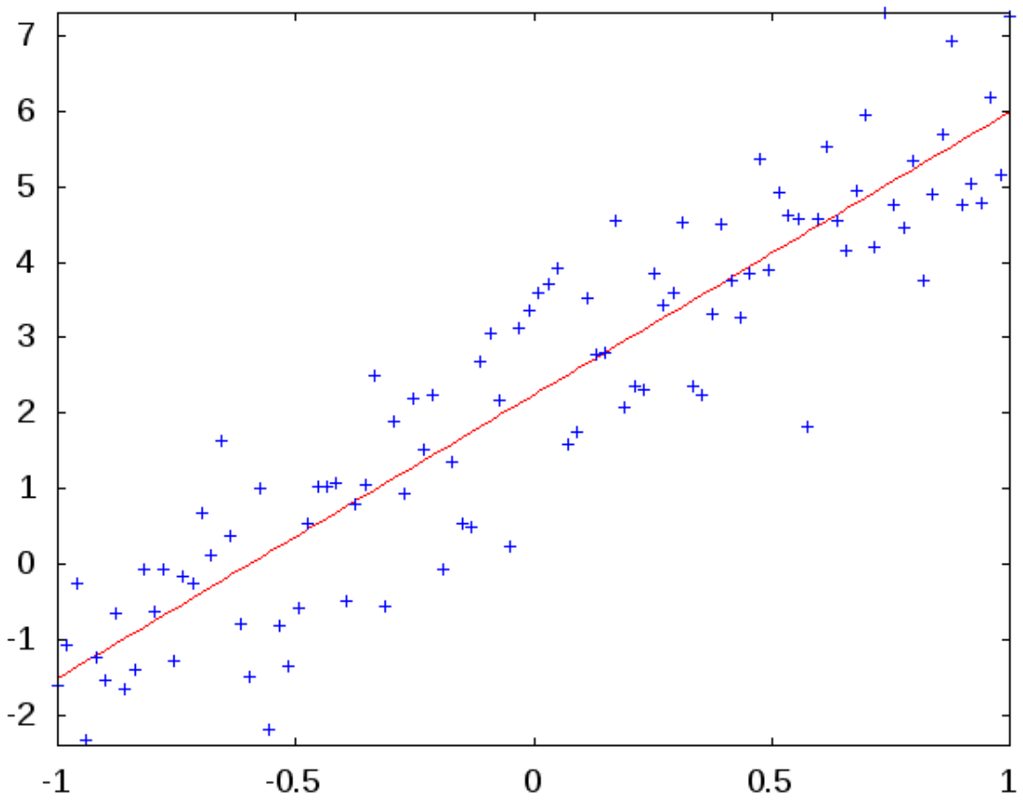
Retour au cours de pandas

- Il n'y a pas de technique magique pour préparer la donnée, il faut être inventif pour trouver ce que l'on veut et un peu de méthode
- La connaissance métier est aussi importante car elle aide à savoir l'importance de telle ou telle variable

4 Modélisation

Régression Linéaire

- La régression linéaire est le modèle le plus simple
- Il suppose que la sortie est corrélée à plusieurs variables d'entrées



$$Y = \beta 0 + \beta 1X + \varepsilon$$

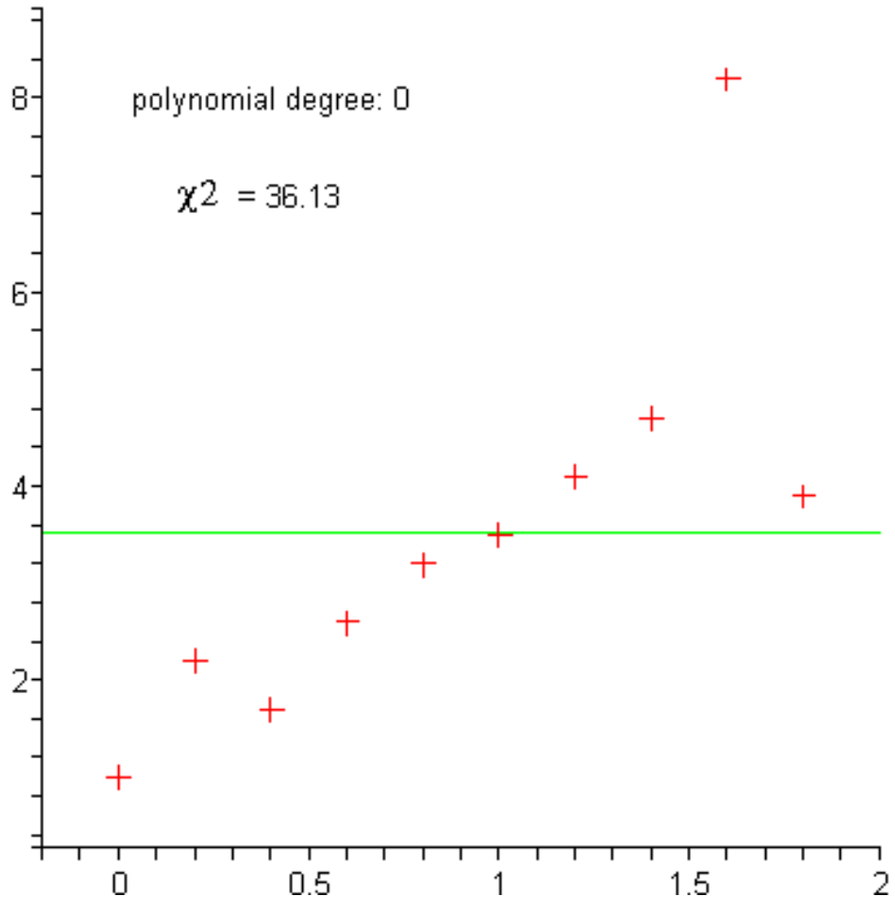
4 Modélisation

Régression Linéaire: exercice

4 Modélisation

Régression Polynomiale

- Ce modèle cherche par régression à lier les variables par un polynôme de degré n



http://scikit-learn.org/stable/modules/linear_model.html

4 Modélisation

Fonction de coût

$$J = \frac{1}{m} \sum (h(x) - y)^2$$

- J est la fonction de coût
- h(x) la prédiction
- y la valeur initial
- et m le nombre de valeur

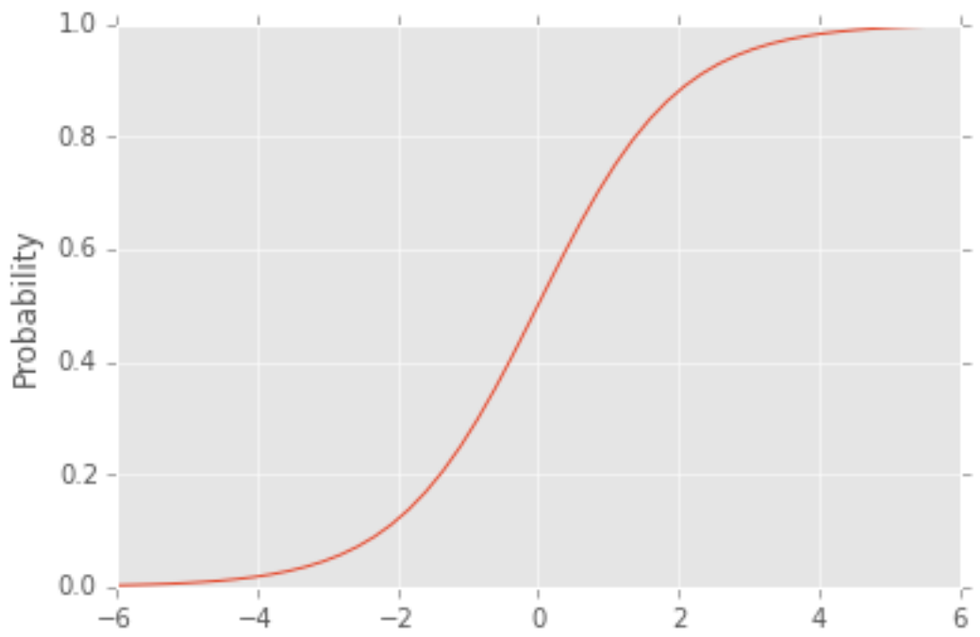
Le but est de **minimiser** le coût

4 Modélisation

Régression Logistique

- Pour résoudre un problème de classification
- On va utiliser le modèle logit qui ramène les valeurs entre 0 et 1 (une probabilité)

$$y(t) = \frac{e^t}{1+e^t}$$



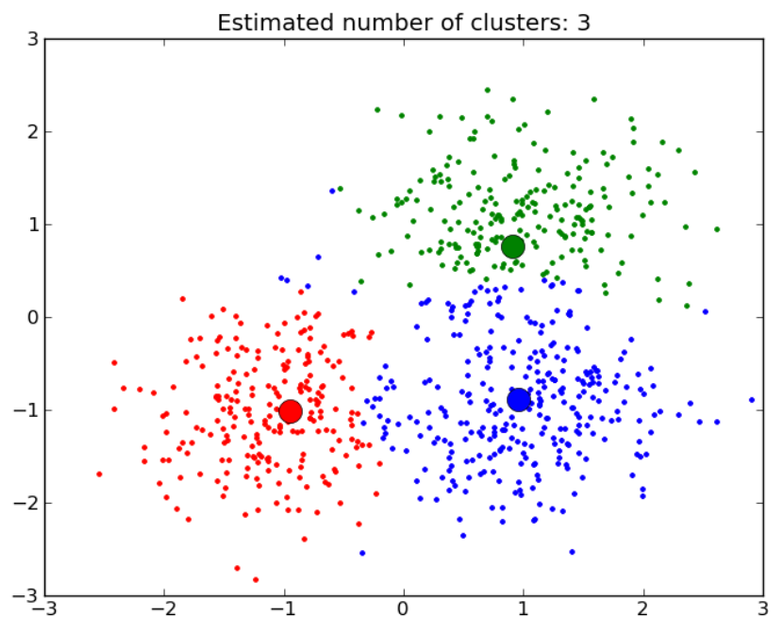
4 Modélisation

Régression Logistique: exercice

4 Modélisation

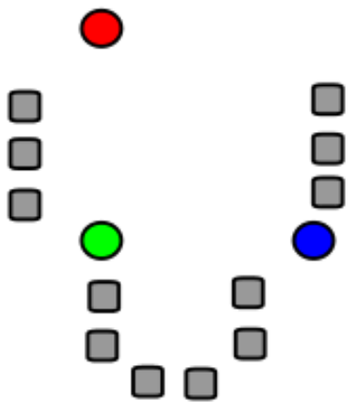
Clustering

- Le but du clustering est de trouver des groupes homogènes dans la donnée via des formules de distances
- Le plus classique est la distance euclidienne



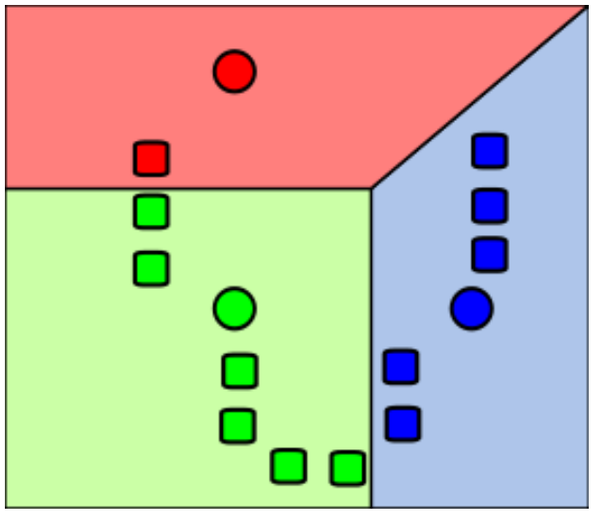
4 Modélisation

Clustering: algorithme k-means



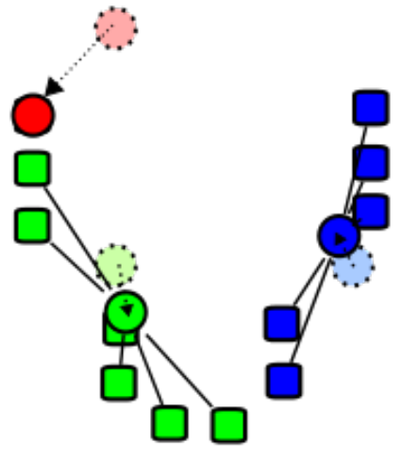
Initialisation

les k premiers centroid sont générés aléatoirement



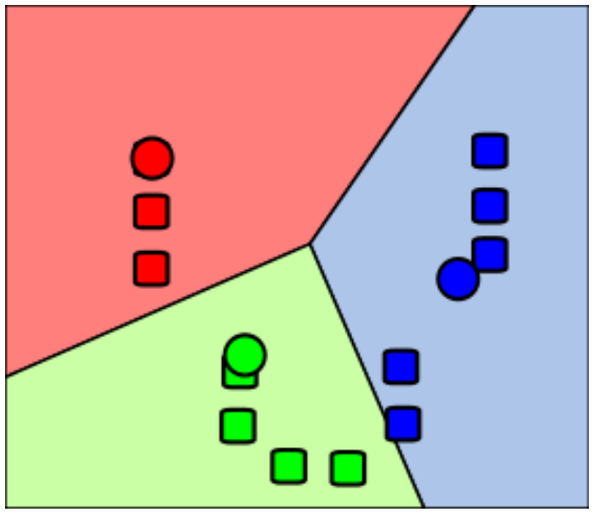
step n

les k clusters sont identifiées et chaque observation est associée à un cluster (voronoi)



correction

les centroids sont modifiés pour arriver au centre de chaque cluster



Répétition

on répète les étapes 2 et 3 pour arriver à une convergence et donc aux clusters finaux

4 Modélisation

Clustering: exercice

4 Modélisation

Évaluation de l'erreur (en regression)

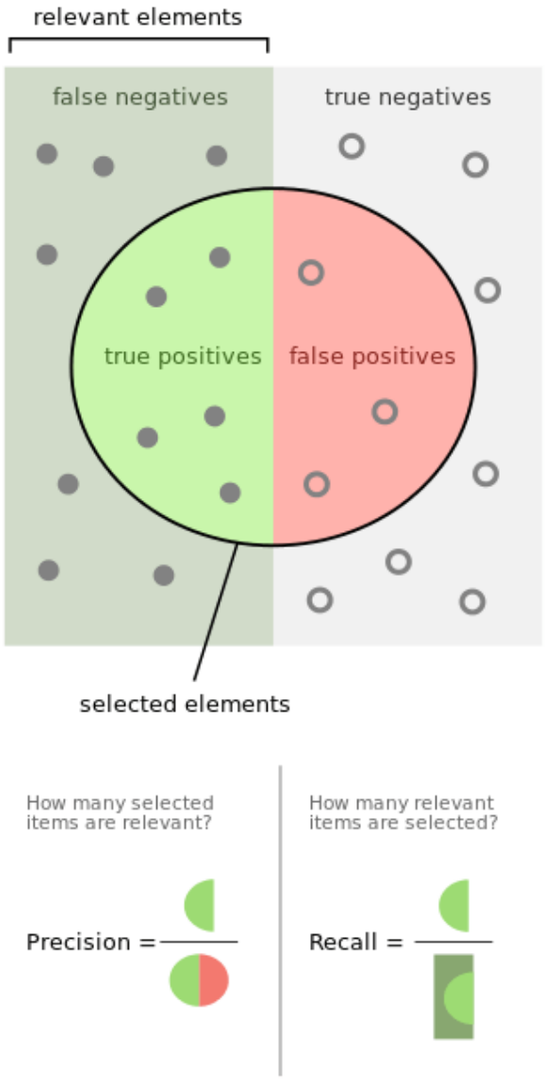
- MSE :
  - utilisé comme fonction de coût sur la regression
  - puni les grosses erreurs
- RMSE :
  - moins sensible aux grosses erreurs
- MAE :
  - facile à interpreter
- MAPE :
  - exprimé en pourcentage
  - utile quand l'objectif varie beaucoup

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

4 Modélisation

Évaluation de l'erreur (en classification)

$$précision = \frac{Vrai\ \ positif}{Vrai\ \ positif\ +\ Faux\ \ positif}$$



		Prédiction	
		Vrai	Faux
Réalité	Vrai	**Vrai positif**	Faux négatif
	Faux	Faux positif	**Vrai négatif**

5 Validation

Train, test…

C’est la donnée que l’on va utiliser pour l’apprentissage, dans cette donnée-là nous gardons les labels pour pouvoir entrainer notre modèle.

La donnée sur laquelle nous allons tester le modèle que l’on aura entrainé.

On connait les labels, mais ils ne sont pas passés pour la prédiction, ils sont utiles pour la validation.

5 Validation

… Validate

- Après avoir entrainé notre modèle nous devons le valider avec les méthodes de validation pour savoir si celui-ci est acceptable

- Nous allons vouloir minimiser les erreurs

- Une variable comme la précision est intéressante mais pas suffisante

- L’étape de validation applique le modèle entrainé avec TRAIN sur TEST

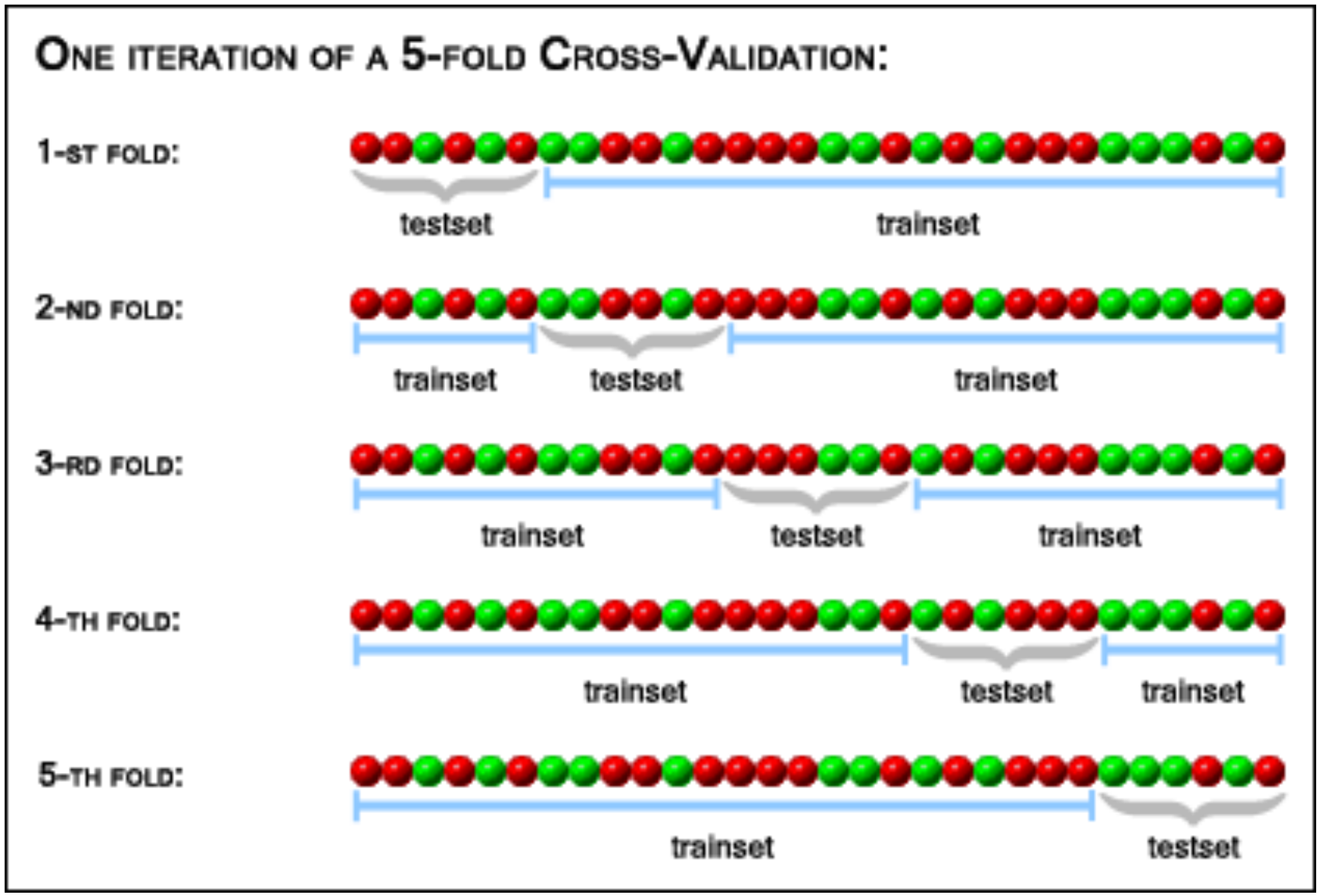
5 Validation

… Validate…

- La cross validation est une étape importante en Machine Learning, elle permet d'éviter de sur-apprendre sur un sous-ensemble de votre dataset
- Pour améliorer la cross-validation nous utilisons une méthode appelé K-Fold qui permet de valider plus complètement le modèle

5 Validation

… Validate…



5 Validation

…et enfin on prédit

- Une fois que notre modèle est OK (ou nous semble OK) nous pouvons lancer la prédiction sur les données complètes

- En conclusion : l'étape la plus importante du Machine Learning est de préparer les données pour les insérer dans le modèle que vous aurez choisi, la partie de prédiction est en général très rapide

➡ On pratique