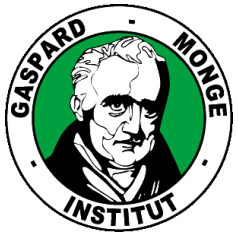
Master 1 Informatique – Université Marne-la-Vallée (IGM)
14/02/2014 – Cours 3
Ingéniérie Linguistique

Classification supervisée et non supervisée



UNIVERSITÉ
PARIS-EST
MARNE-LA-VALLÉE

Philippe Gambette

Sources du cours

- Cours de Matthieu Constant, *Ingéniérie Informatique 1*

http://igm.univ-mlv.fr/ens/Master/M1/2010-2011/IngenierieLinguistique1/cours.php

- Cours d'Alexandre Allauzen et Jérôme Azé, Université Paris-Sud

http://www.bioinfo-biostats-etudiants.u-psud.fr/Ressources/Cours/Master%202/ECT/

- Cours de Guillaume Wisniewski, Université Paris-Sud

http://perso.limsi.fr/Individu/wisniews/enseignement/old/10-11/10-11_rdf_m1/

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Introduction

Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

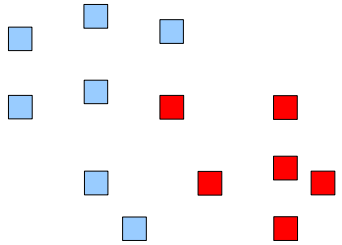
• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

Classification non supervisée

• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

Introduction



Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

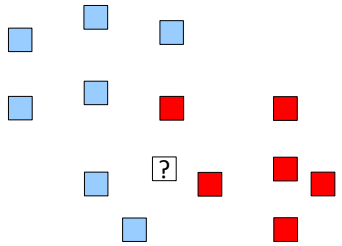
• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

Classification non supervisée

• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

Introduction



Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

Classification non supervisée

• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

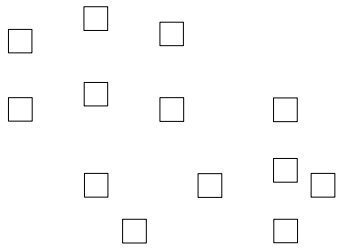
Introduction

Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

Classification non supervisée



• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

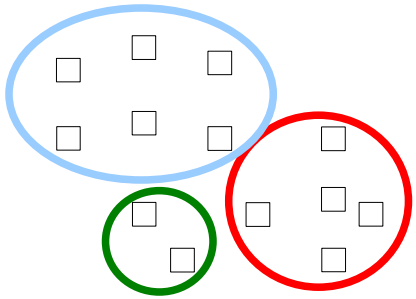
Introduction

Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

Classification non supervisée



• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

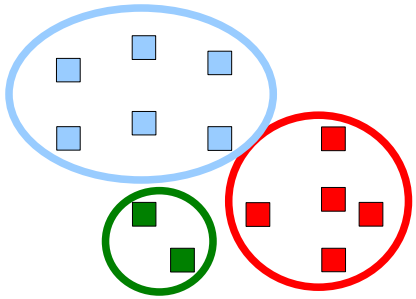
Introduction

Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...

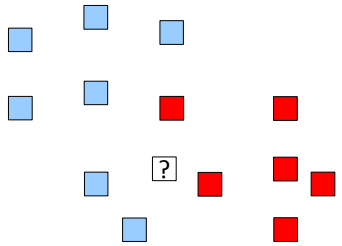
Classification non supervisée



• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

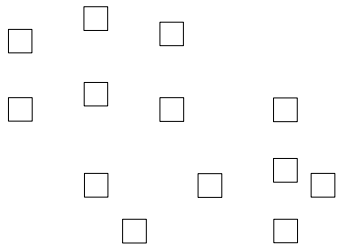
Introduction



Classification supervisée :

• On dispose d'**éléments déjà classés**
Exemple : articles en rubrique économie, politique, sport, culture...

• On veut **classer un nouvel élément**
Exemple : lui attribuer une étiquette parmi économie, politique, sport, culture...



Classification non supervisée

• On dispose d'**éléments non classés**
Exemple : mots d'un texte

• On veut les **regrouper en classes**
Exemple : si deux mots ont la même étiquette, ils sont en rapport avec une même thématique...

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Classification supervisée à deux classes

Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

Classification supervisée à deux classes

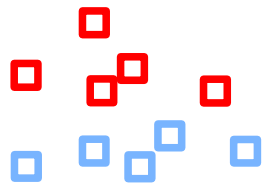
Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

Évaluation

- **Précision** : moyenne de la proportion de vrais documents + parmi les documents classés +, et de la proportion de vrais documents - parmi les documents classés -
- **Rappel** : moyenne de la proportion parmi les vrais documents +, des documents classés +, et de la proportion, parmi les vrais documents -, des documents classés -.



vrai document +
document classé +
vrai document +
document classé +

Classification supervisée à deux classes

Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

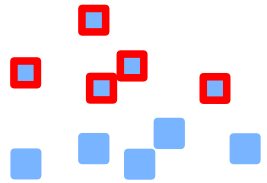
- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

Évaluation

précision = 0.75
rappel = 0.5

précision = 0.5
rappel = 1

précision = 1
rappel = 0



vrai document +
document classé +
vrai document -
document classé -

- **Précision** : moyenne de la proportion de vrais documents + parmi les documents classés +, et de la proportion de vrais documents - parmi les documents classés -
- **Rappel** : moyenne de la proportion parmi les vrais documents +, des documents classés +, et de la proportion, parmi les vrais documents -, des documents classés -.

Classification supervisée à deux classes

Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

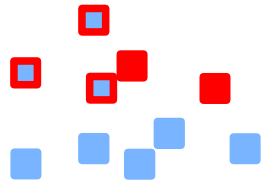
- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

Évaluation

précision = 0.8125
rappel = 0.7

précision = 0.625
rappel = 1

précision = 1
rappel = 0.4



vrai document +
document classé +
vrai document -
document classé -

- **Précision** : moyenne de la proportion de vrais documents + parmi les documents classés +, et de la proportion de vrais documents - parmi les documents classés -
- **Rappel** : moyenne de la proportion parmi les vrais documents +, des documents classés +, et de la proportion, parmi les vrais documents -, des documents classés -.

Classification supervisée à deux classes

Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

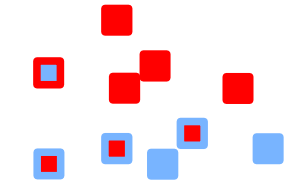
Évaluation

précision = 0.619
rappel = 0.65

précision = 0.667
rappel = 0.3

précision = 0.571
rappel = 0.8

- **Précision** : moyenne de la proportion de vrais documents + parmi les documents classés +, et de la proportion de vrais documents - parmi les documents classés -
- **Rappel** : moyenne de la proportion parmi les vrais documents +, des documents classés +, et de la proportion, parmi les vrais documents -, des documents classés -.



vrai document +
document classé +
vrai document -
document classé -

Classification supervisée à deux classes

Conception d'une méthode

Étiquetage manuel du corpus + partitionnement du corpus en deux :

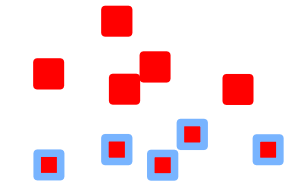
- un **corpus d'apprentissage** *APP* : éléments déjà classés + ou - (80%)
- un **corpus d'évaluation** *EVAL* : éléments à classer (20%)
- (si on organise un concours, un **corpus de test** *TEST* est fourni aux candidats)

Évaluation

précision = 0.75
rappel = 0.5

précision = 1
rappel = 0

précision = 0.5
rappel = 1

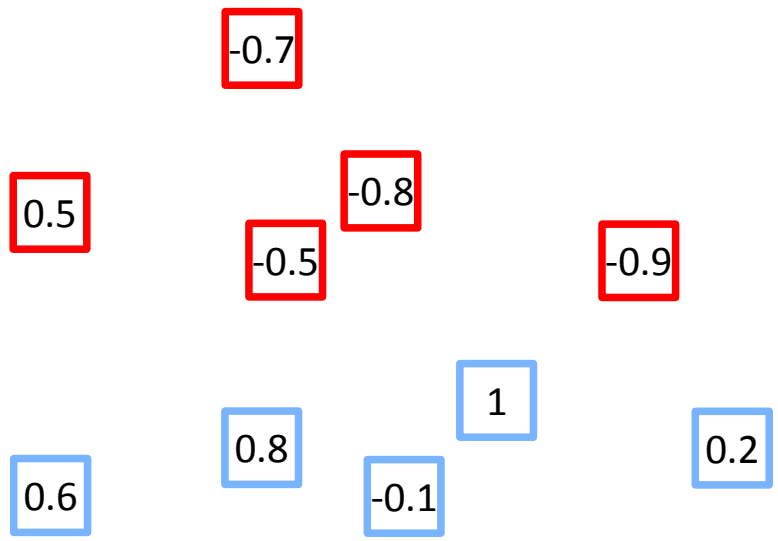


vrai document +
document classé +
vrai document -
document classé -

- **Précision** : moyenne de la proportion de vrais documents + parmi les documents classés +, et de la proportion de vrais documents - parmi les documents classés -
- **Rappel** : moyenne de la proportion parmi les vrais documents +, des documents classés +, et de la proportion, parmi les vrais documents -, des documents classés -.

Classification supervisée à deux classes

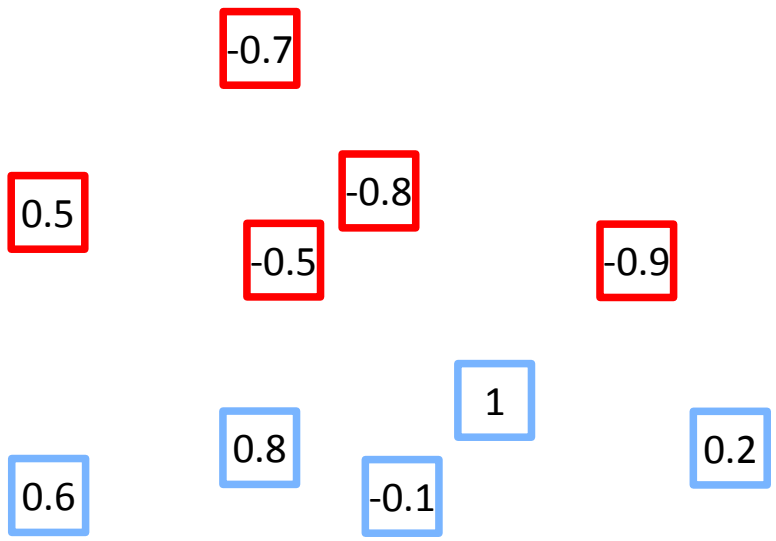
Étiquettes non binaires, scores entre -1 et 1 :



- vrai document +
- document classé +
- vrai document -
- document classé -
- vrai-positif
- faux-positif

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :

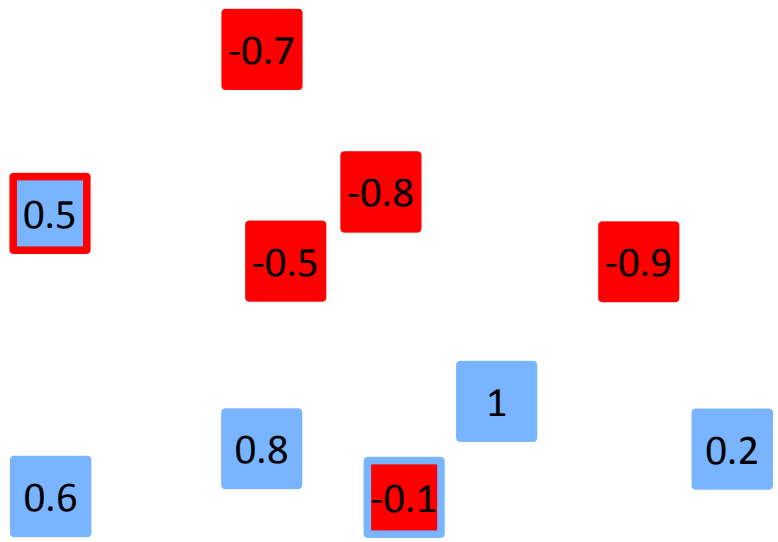


vrai document +
document classé +
vrai document -
document classé -
vrai-positif
faux-positif

→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :



vrai document +
document classé +
vrai document -
document classé -
vrai-positif
faux-positif

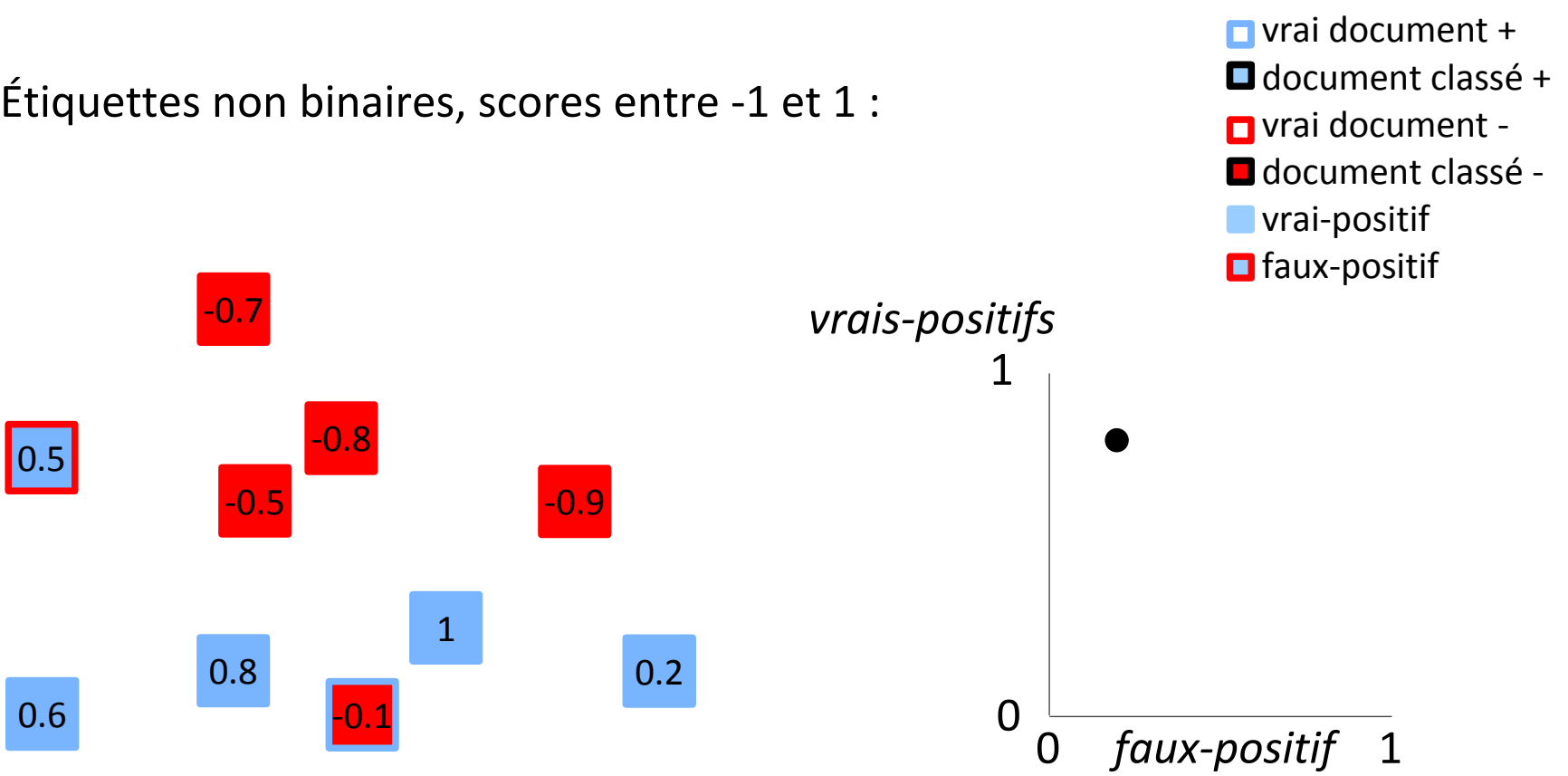
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

Exemple : *s*=0 (4 vrais positifs → fraction des vrais + classés + : 0.8)

(1 faux positif → fraction des vrais - classés + : 0.2)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :



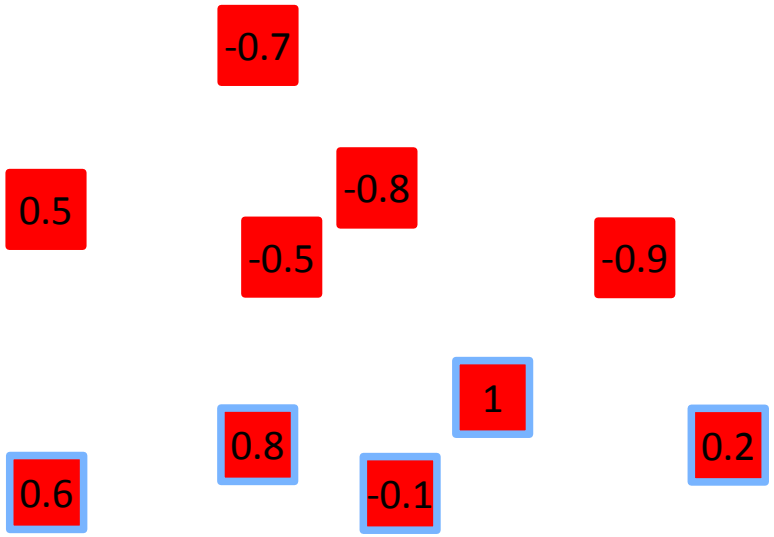
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

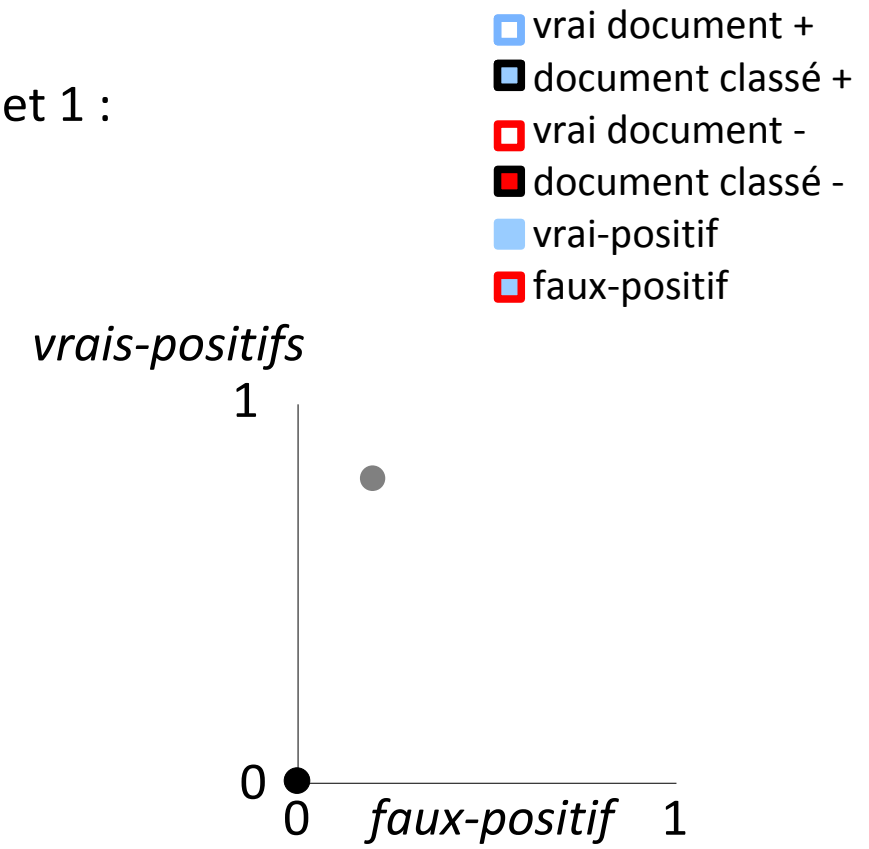
Exemple : *s*=0 (4 vrais-positifs → fraction des vrais + classés + : 0.8)

(1 faux-positif → fraction des vrais - classés + : 0.2)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :





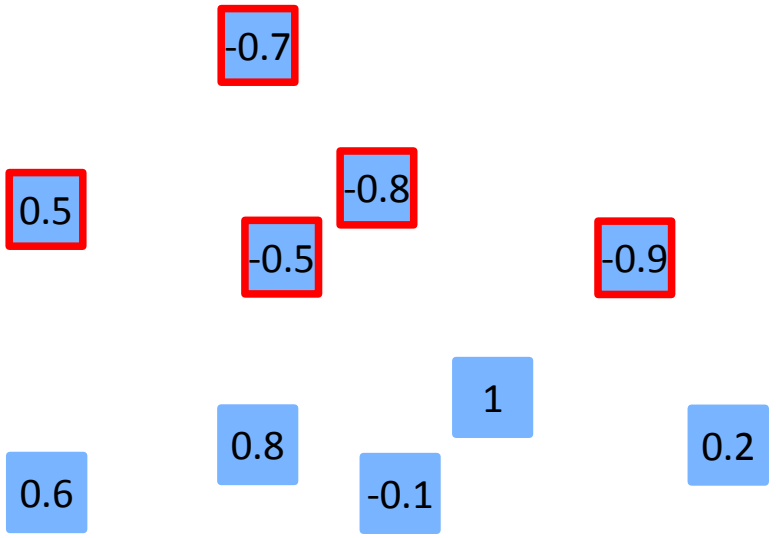
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

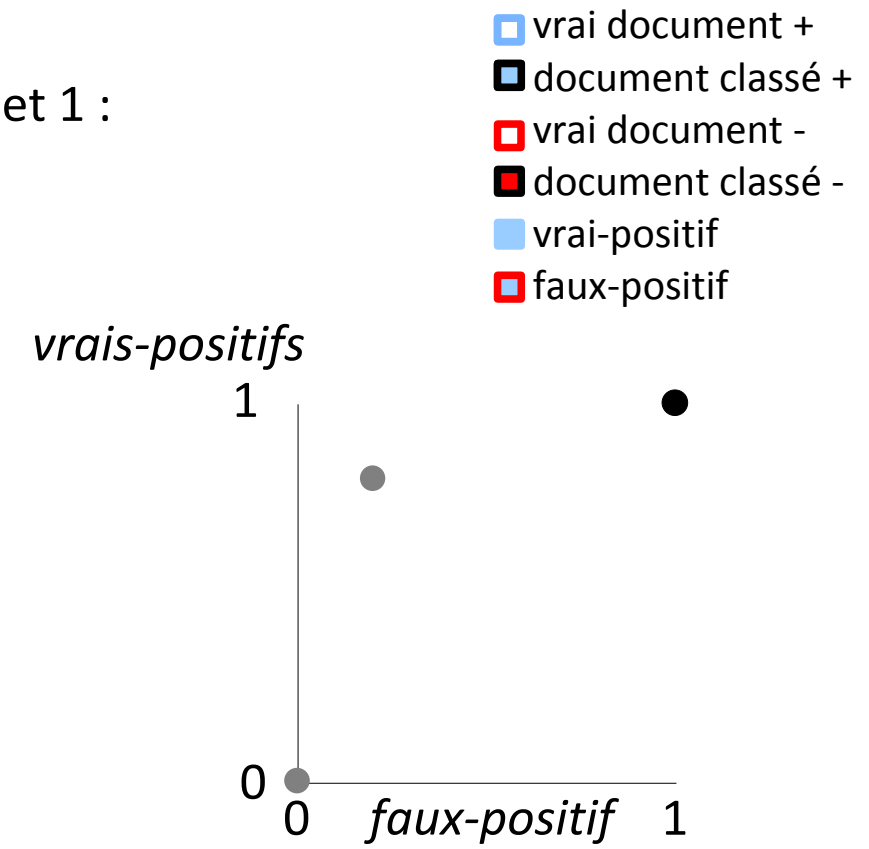
Exemple : *s*=1.1 (0 vrai-positif → fraction des vrais + classés + : 0)

(0 faux-positif → fraction des vrais - classés + : 0)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :





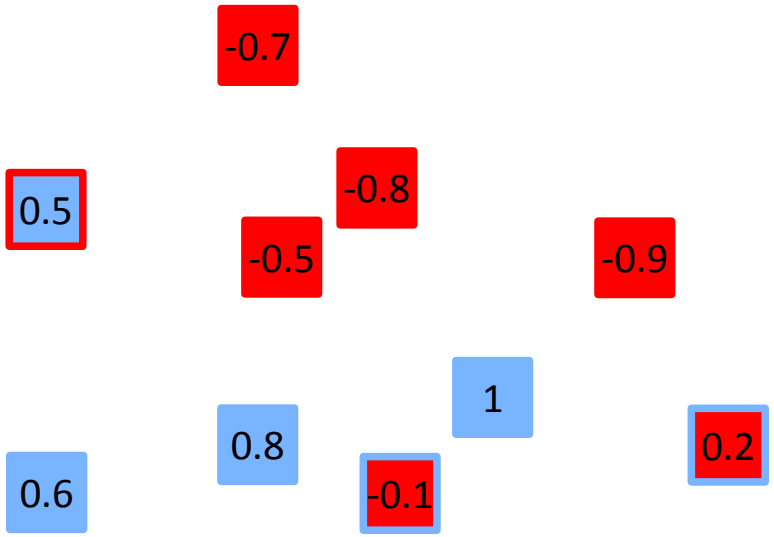
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

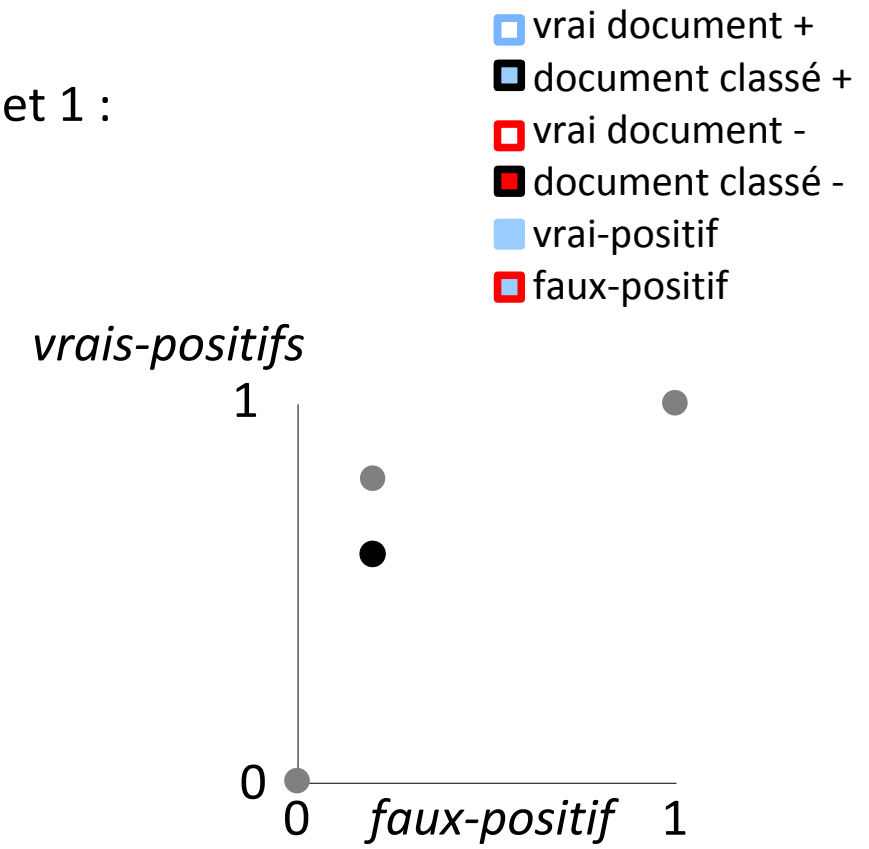
Exemple : *s*=-1.1 (5 vrais-positifs → fraction des vrais + classés + : 1)

(5 faux-positif → fraction des vrais - classés + : 1)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :





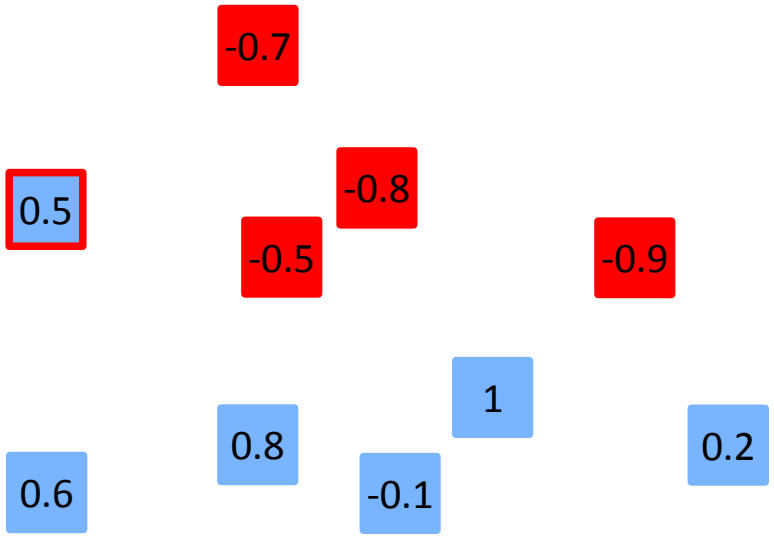
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

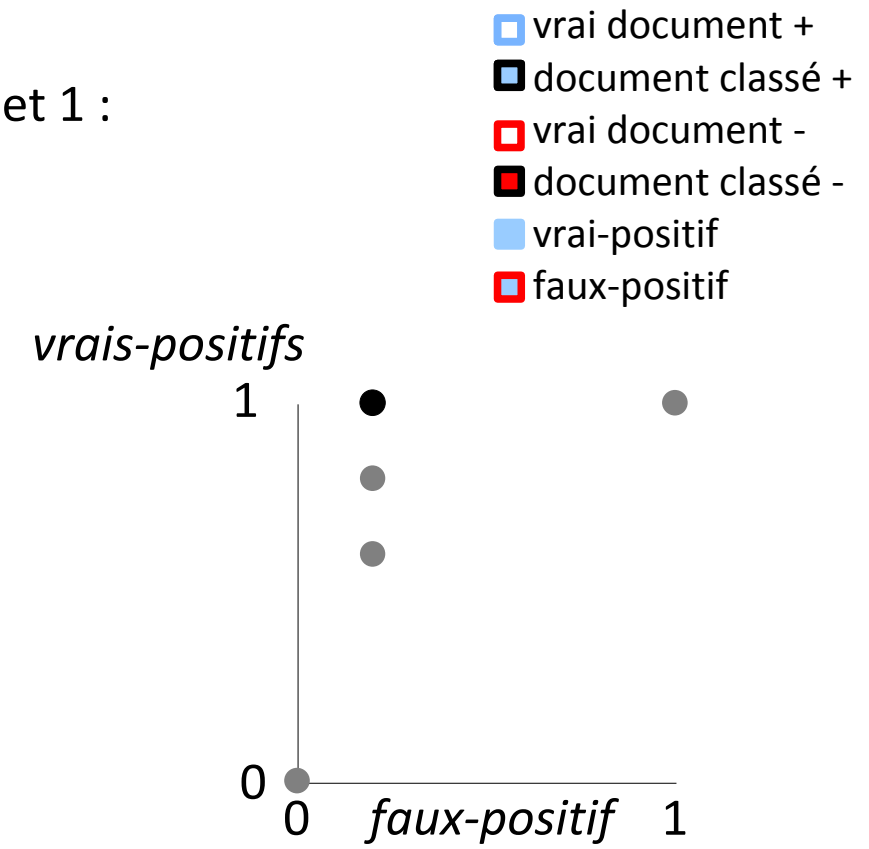
Exemple : *s*=0.4 (3 vrais-positifs → fraction des vrais + classés + : 0.6)

(1 faux-positif → fraction des vrais - classés + : 0.2)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :





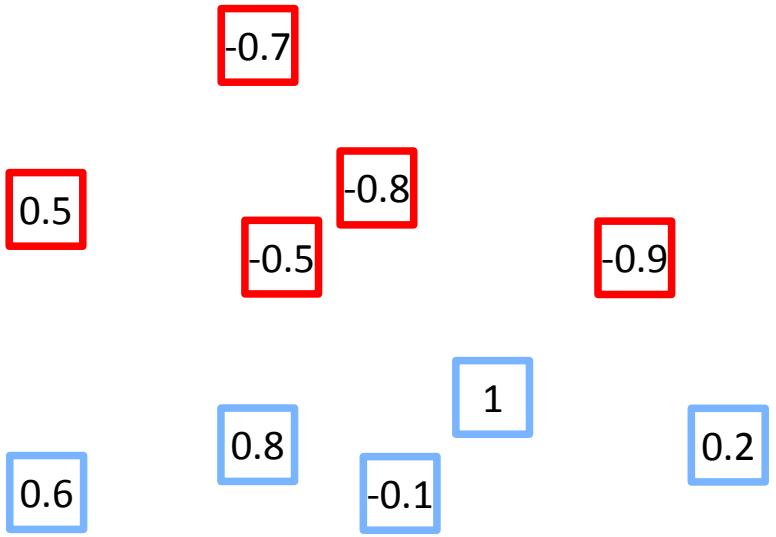
→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

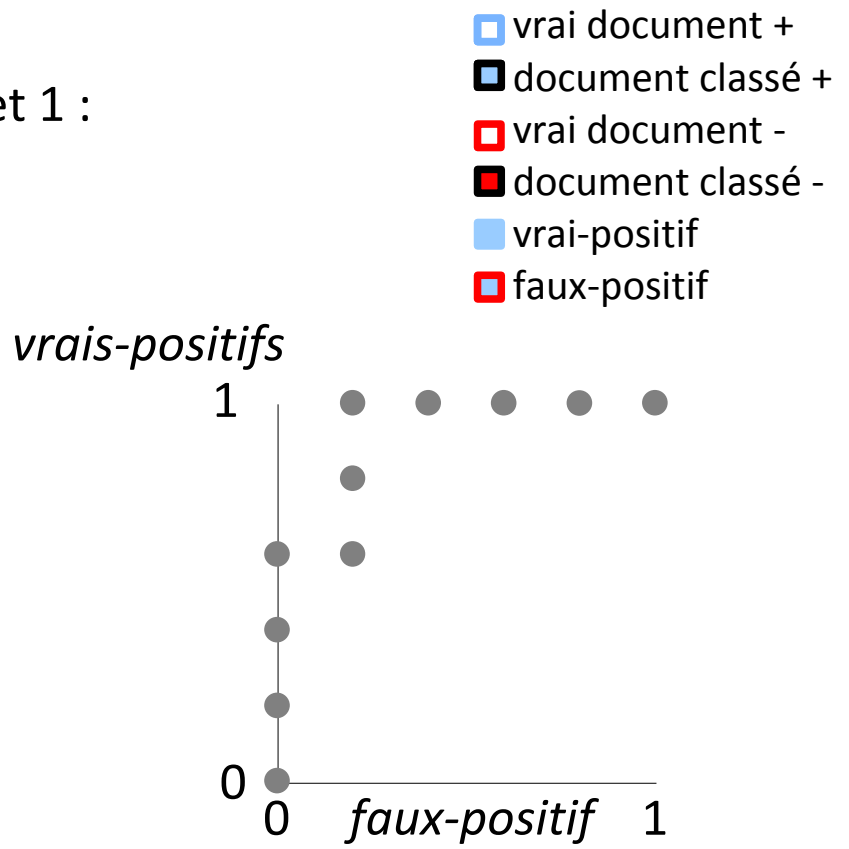
Exemple : *s*=-0.4 (5 vrais-positifs → fraction des vrais + classés + : 1)

(1 faux-positif → fraction des vrais - classés + : 0.2)

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :

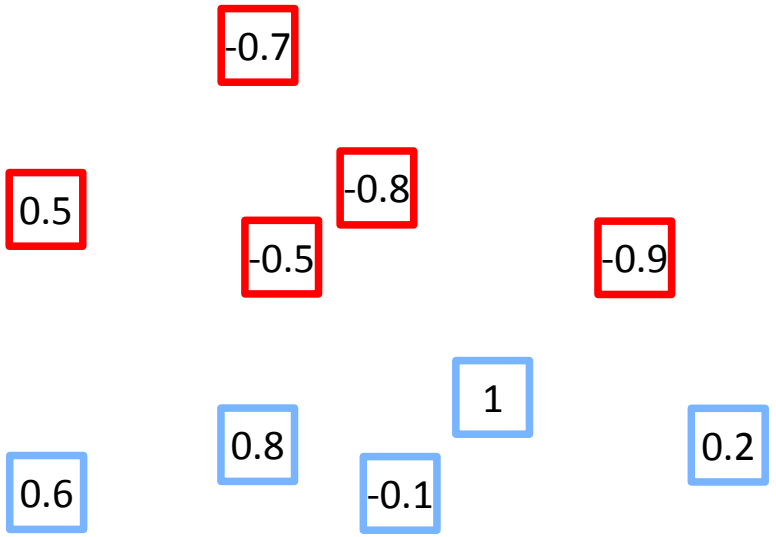


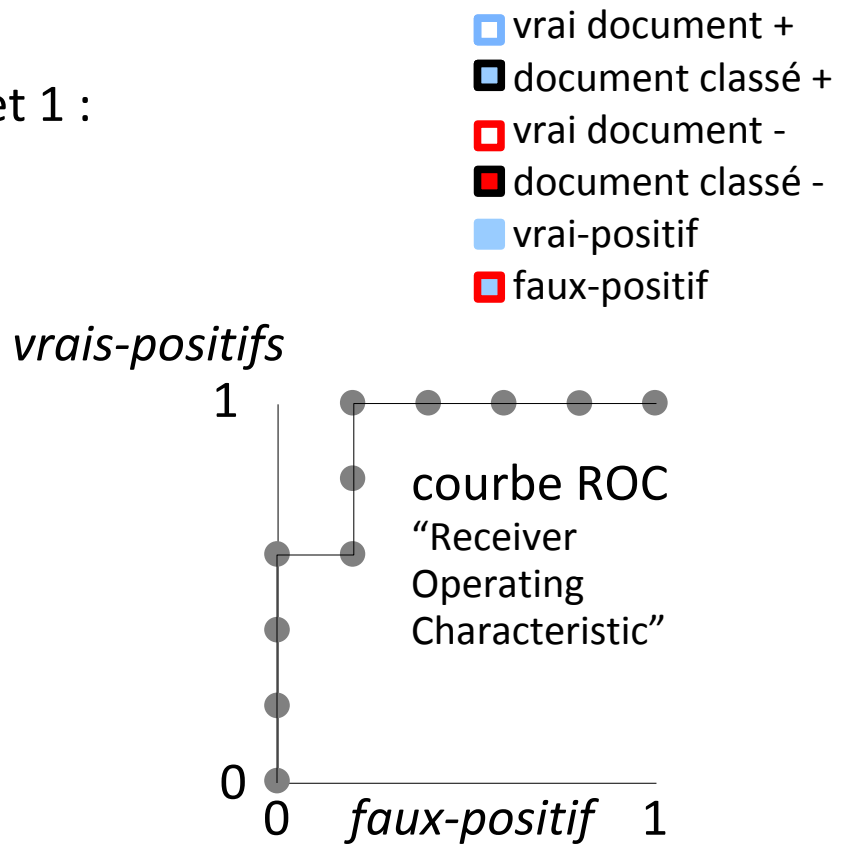


→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

Classification supervisée à deux classes

Étiquettes non binaires, scores entre -1 et 1 :





→ choix d'un **seuil** *s* pour attribuer l'étiquette – ou +

La courbe ROC doit **se rapprocher le plus possible de (0,1)** pour un bon classifieur.

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Approche du centroïde

Idée :

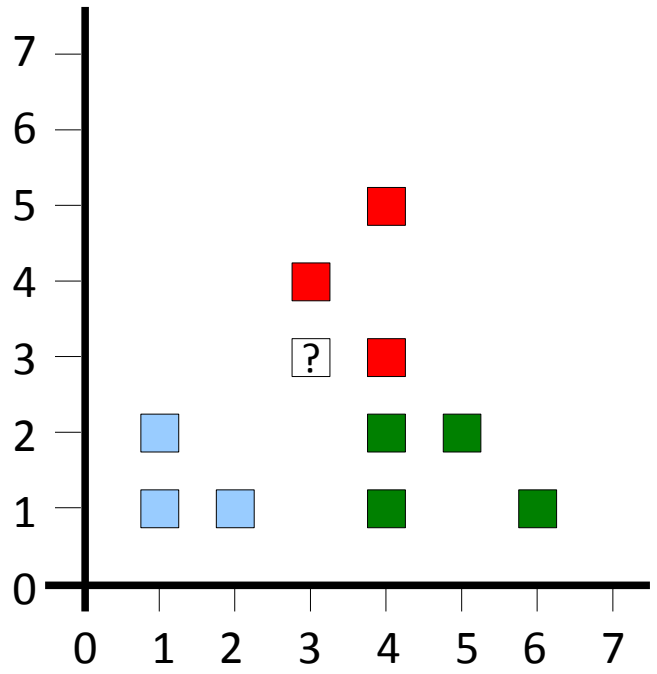
Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres.

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres.

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

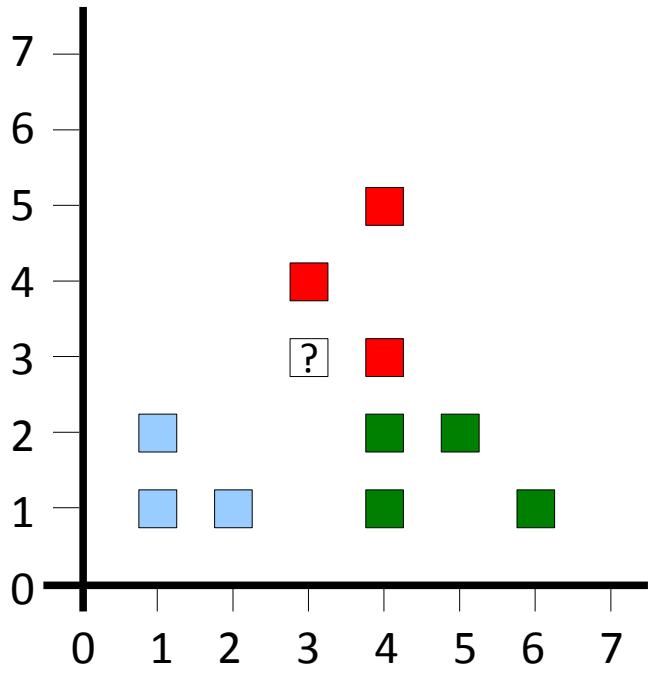
Document D11(3,3) à classer

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3

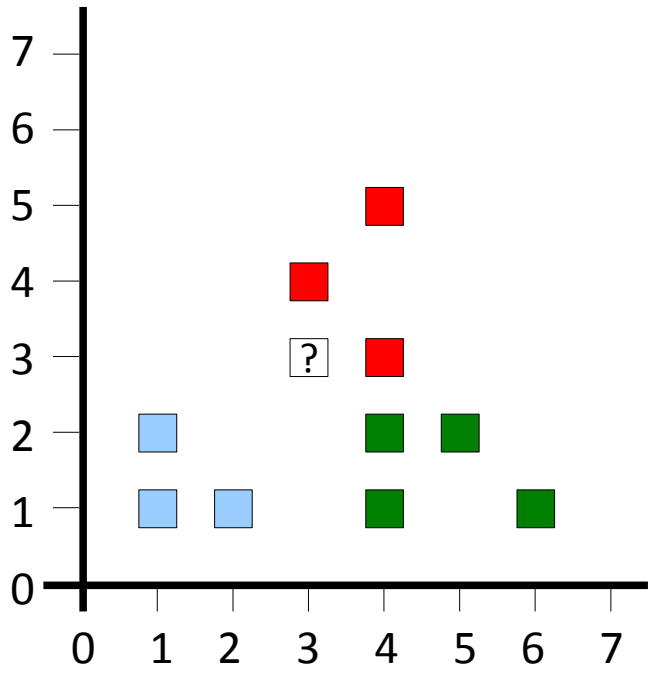
Document D11(3,3) à classer

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3

Document D11(3,3) à classer

i-ième coordonnée du centroïde de la classe *k* :

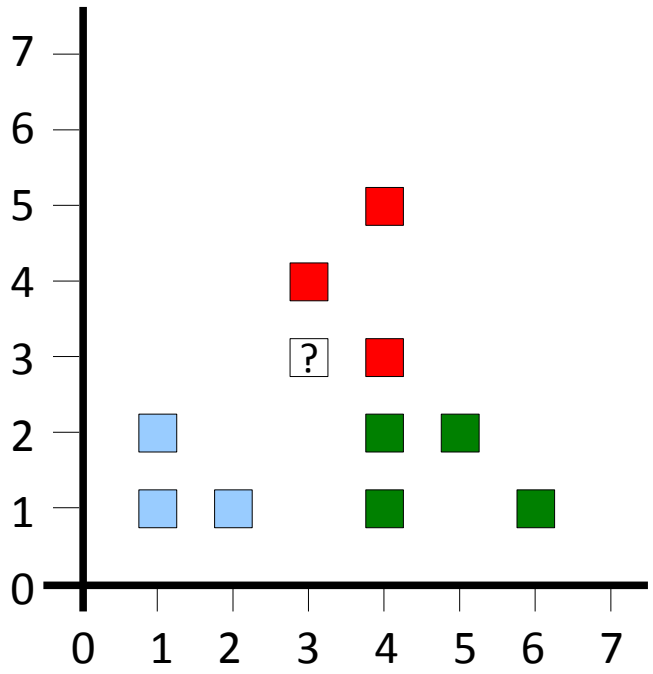
$$Ck_i = \frac{\sum_{Dj \in \text{classe } k} Dj_i}{|\text{classe } k|}$$

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1(1.333,1.333)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2(3.667,4)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3(4.75,1.5)

Document D11(3,3) à classer

i-ième coordonnée du centroïde de la classe *k* :

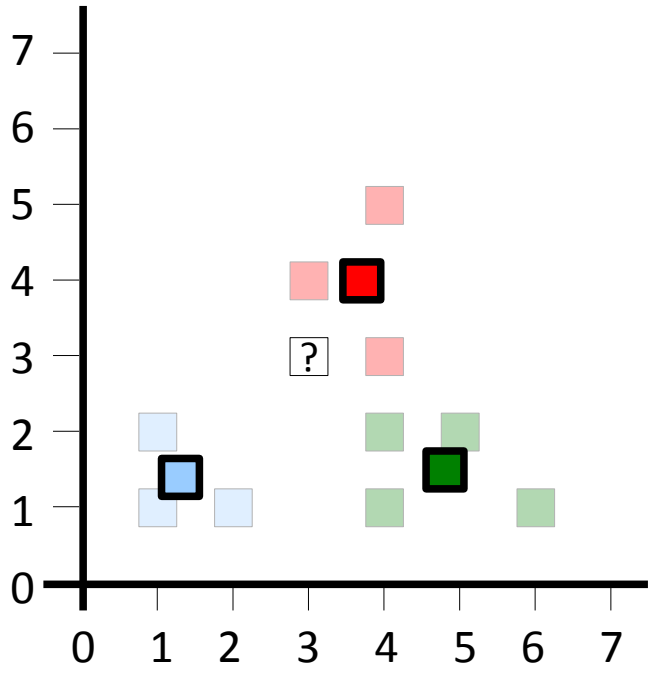
$$Ck_i = \frac{\sum_{Dj \in \text{classe } k} Dj_i}{|\text{classe } k|}$$

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1(1.333,1.333)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2(3.667,4)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3(4.75,1.5)

Document D11(3,3) à classer

i-ième coordonnée du centroïde de la classe *k* :

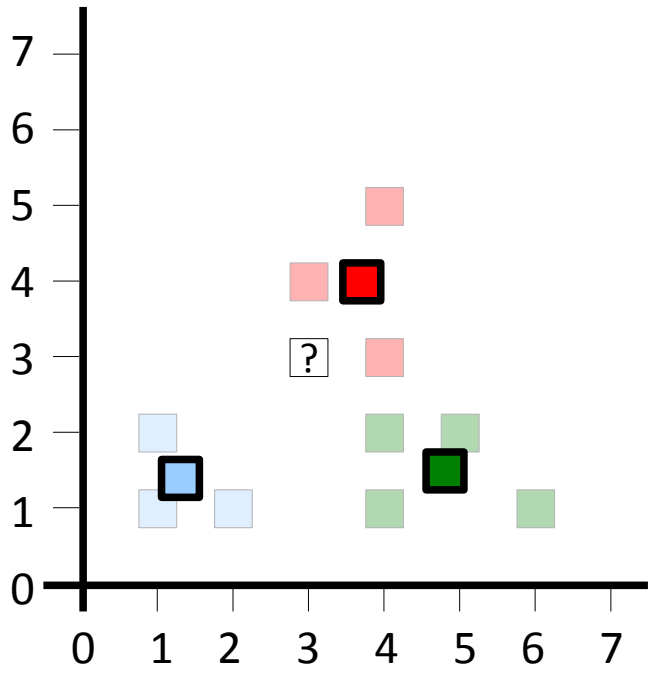
$$Ck_i = \frac{\sum_{Dj \in \text{classe } k} Dj_i}{|\text{classe } k|}$$

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1(1.333,1.333)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2(3.667,4)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3(4.75,1.5)

Document D11(3,3) à classer

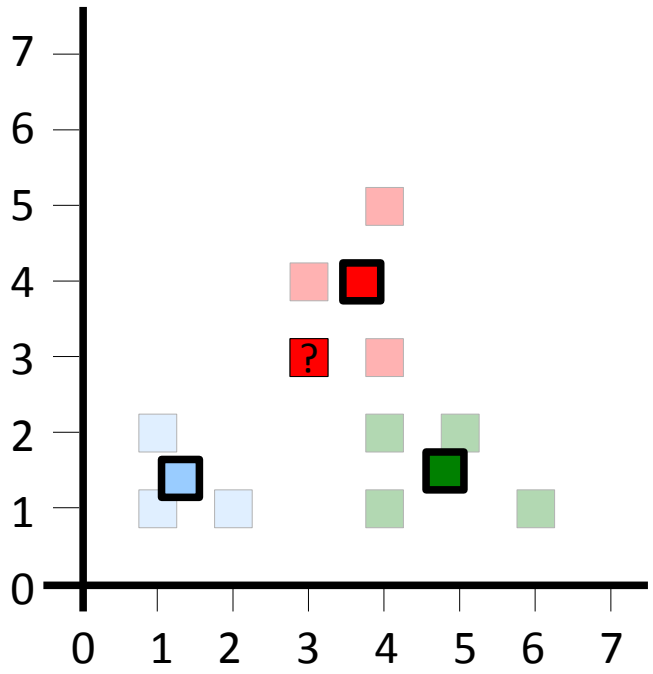
Carrés des distances euclidiennes aux centroïdes

Approche du centroïde

Idée :

Représenter chaque classe par son centre et classer le nouvel élément en fonction de sa distance aux centres (= centroïdes, barycentres, moyennes).

Exemple en dimension 2 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)
C1(1.333,1.333)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)
C2(3.667,4)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)
C3(4.75,1.5)

Document D11(3,3) à classer

Carrés des distances euclidiennes aux centroïdes
→ Classe 2 !

Plan

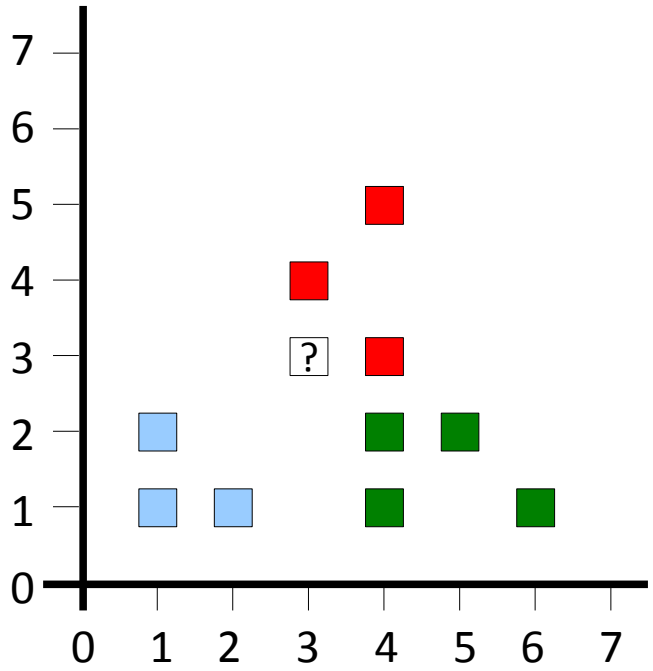
- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=3 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

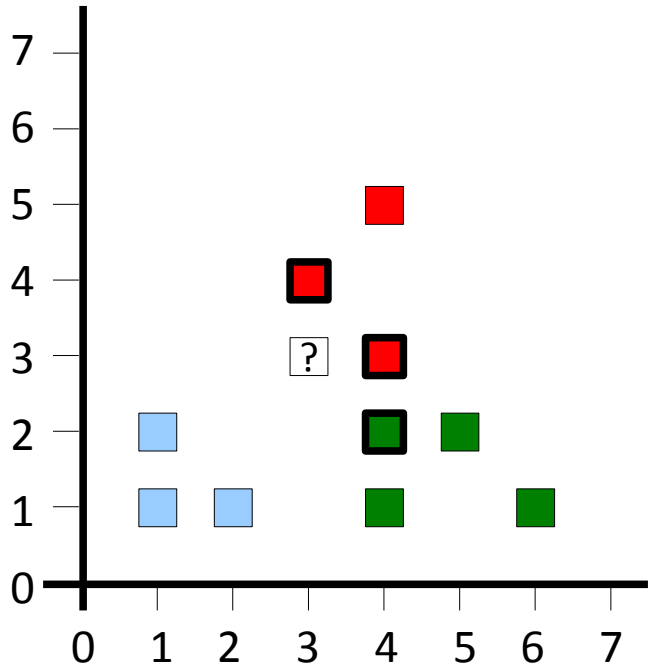
Document D11(3,3) à classer

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=3 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

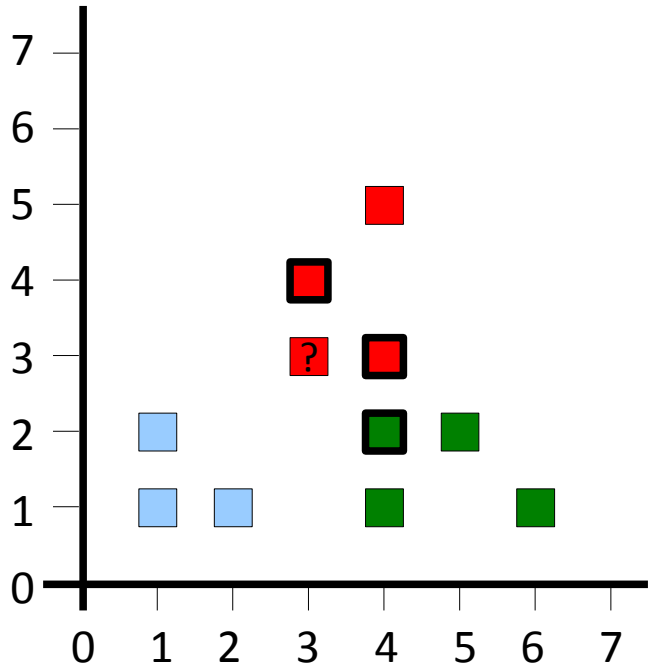
Document D11(3,3) à classer

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=3 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

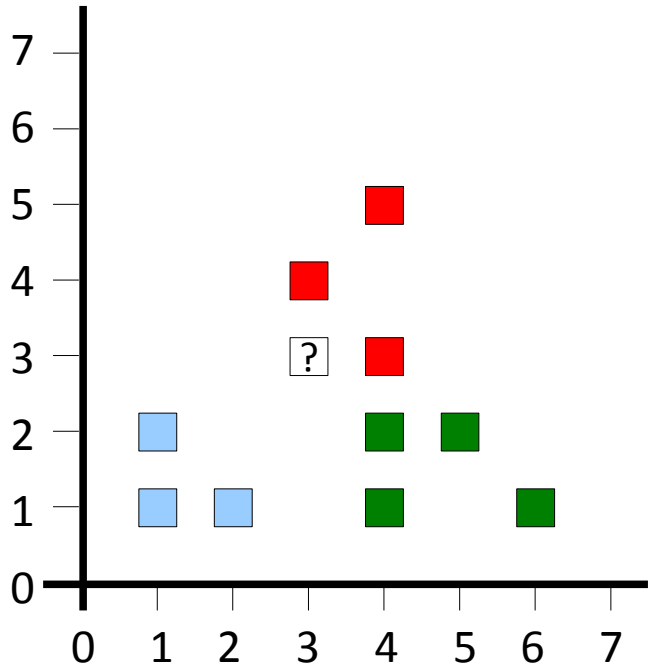
Document D11(3,3) à classer

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=10 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

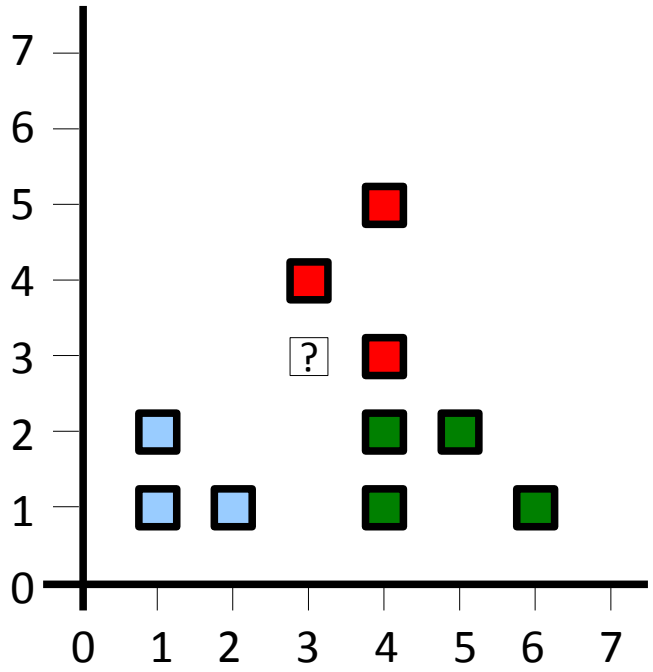
Document D11(3,3) à classer

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=10 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

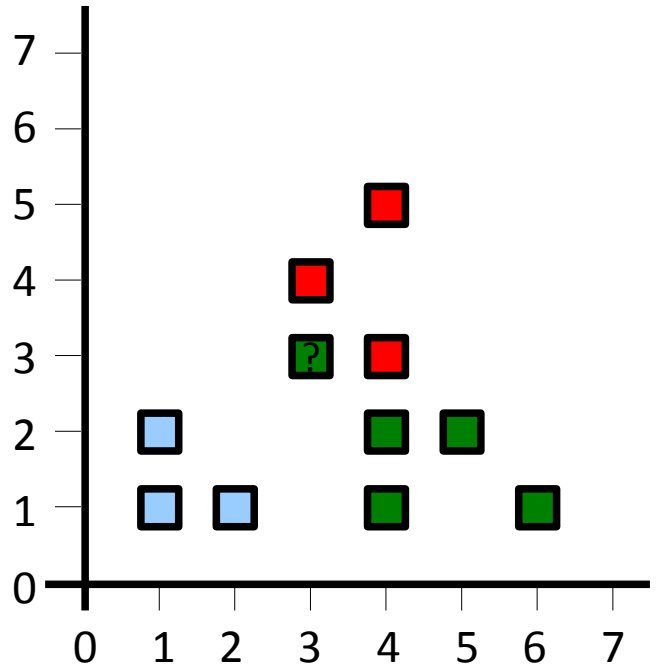
Document D11(3,3) à classer

Approche des *k* plus proches voisins

Idée :

Choisir pour chaque sommet la classe majoritaire parmi ses *k* plus proches voisins.

Exemple en dimension 2, *k*=10 :



Classe 1 :
D1 (1,1)
D2 (1,2)
D3 (2,1)

Classe 2 :
D4 (3,4)
D5 (4,5)
D6 (4,3)

Classe 3 :
D7 (4,2)
D8 (5,2)
D9 (6,1)
D10(4,1)

Document D11(3,3) à classer

Approche des *k* plus proches voisins

Cas d'égalité ?

• Augmenter k de 1 ? Fonctionnera si classification à 2 classes, risque d'échouer sinon.

• Tirage au hasard.

• Pondération des voisins par rapport à leur distance au point à classer.

Plan

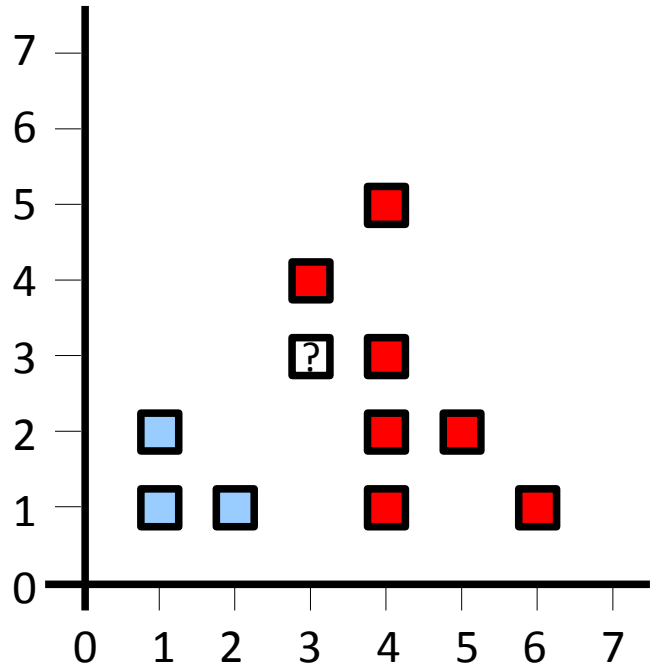
- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Classifieurs linéaires

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :

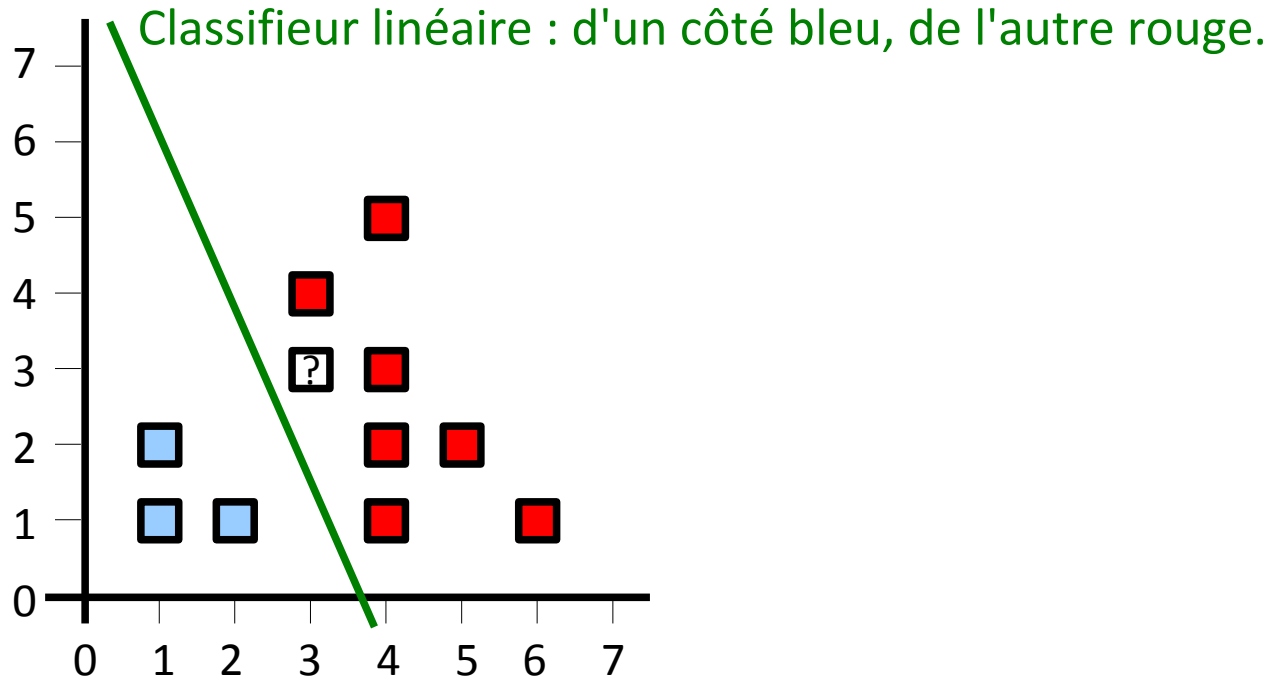


Classifieurs linéaires

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :

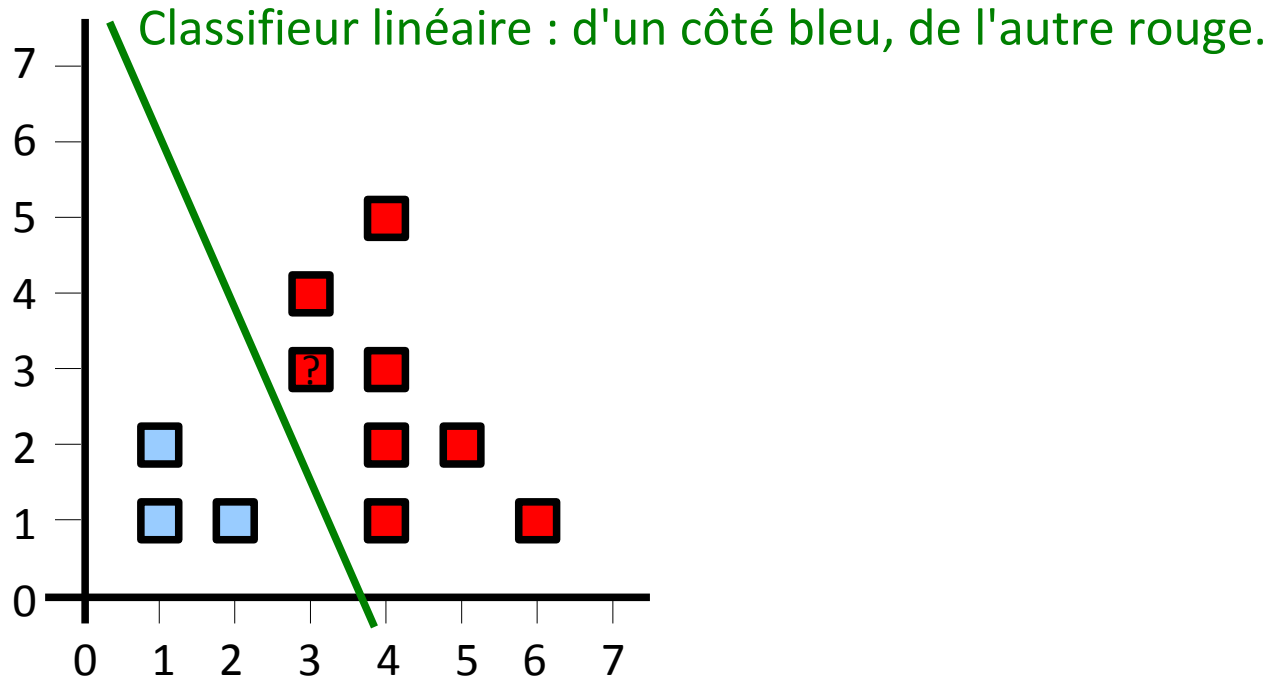


Classifieurs linéaires

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :

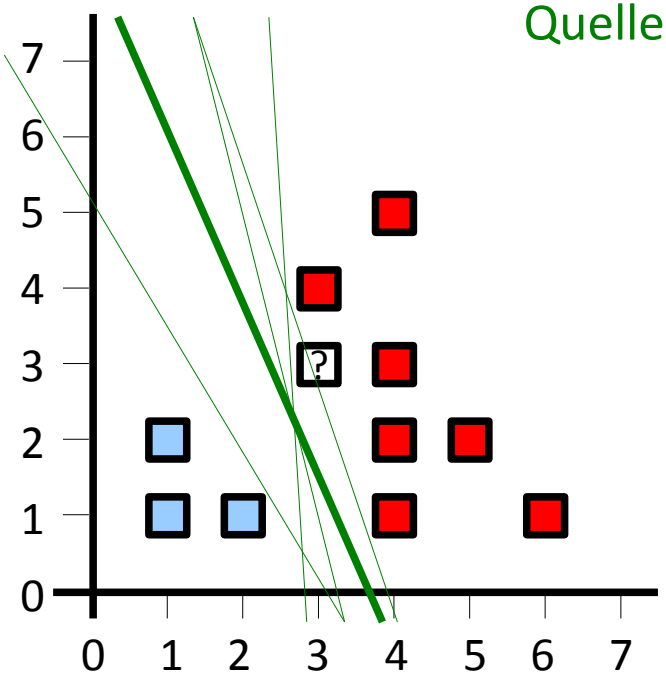


Classifieurs linéaires

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :

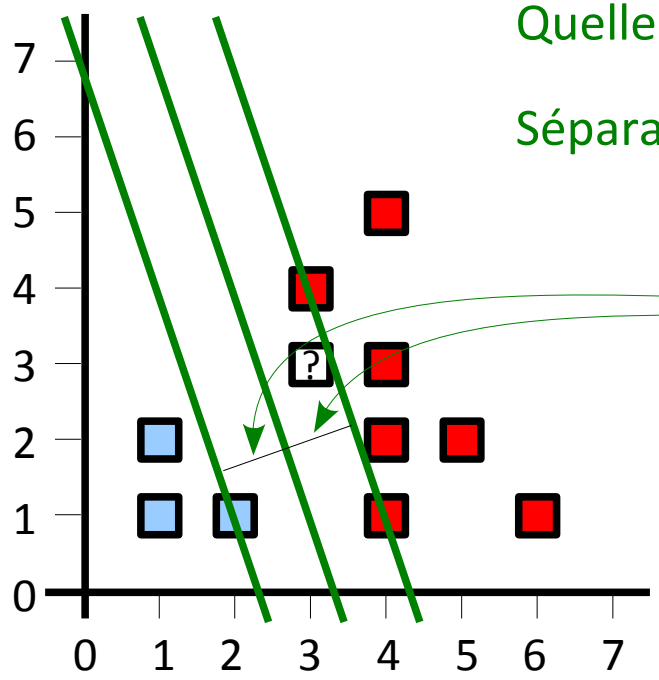


Classifieurs linéaires - SVM

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :



Quelle droite choisir ?

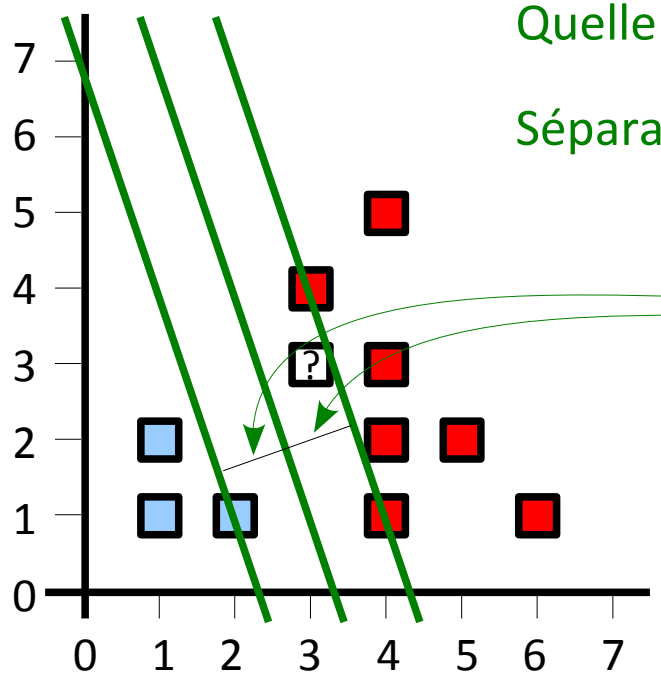
Séparateur à Vastes Marges : maximiser la marge de part et d'autre du classifieur linéaire.

Classifieurs linéaires - SVM

Idée pour une classification supervisée à 2 classes :

Choisir une ligne qui sépare le mieux les deux classes

Exemple en dimension 2 :



Quelle droite choisir ?

Séparateur à Vastes Marges : maximiser la marge de part et d'autre du classifieur linéaire.

+ généralisation à des classifieurs non linéaires par transformation dans un espace de plus grande dimension où il existe un hyperplan linéaire séparateur “kernel trick”

https://www.lri.fr/~sebag/Slides/Cachan_5_2013.pdf

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Classification non supervisée

Objectif : trouver une **partition des données**

Évaluation

Qualité des partitions calculées ?

• Formule de **Rand** entre deux partitions P1 et P2 :

Rand(P1,P2) = ratio de paires d'éléments se comportant pareil dans P1 et P2

http://en.wikipedia.org/wiki/Rand_index

Classification non supervisée

Objectif : trouver une **partition des données**

Évaluation

• Si aucune partition de référence n'est connue, vérifier que :

- les éléments proches sont dans un même ensemble (ou “cluster”) de la partition

- les éléments éloignés sont dans deux ensembles différents de la partition

• Si une partition de référence est connue :

→ distance avec la partition de référence ?

Classification non supervisée

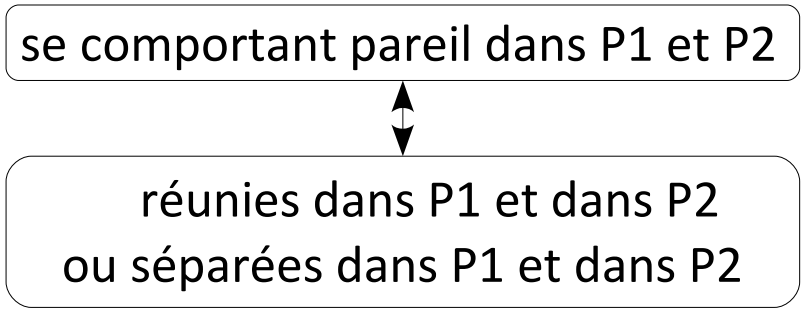
Objectif : trouver une **partition des données**

Évaluation

Qualité des partitions calculées si une partition de référence P1 est connue ?

• Formule de **Rand** entre deux partitions P1 et P2 :

Rand(P1,P2) = ratio de paires d'éléments



http://en.wikipedia.org/wiki/Rand_index

Classification non supervisée

Objectif : trouver une **partition des données**

Évaluation

Qualité des partitions calculées si une partition de référence P1 est connue ?

• Formule de **Rand** entre deux partitions P1 et P2 :

Rand(P1,P2) = ratio de paires d'éléments se comportant pareil dans P1 et P2

• Formule **"adjusted Rand"** entre deux partitions P1 et P2 :

Prise en compte du fait que certaines paires se comportent pareil par hasard.

http://en.wikipedia.org/wiki/Rand_index

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

k-moyennes

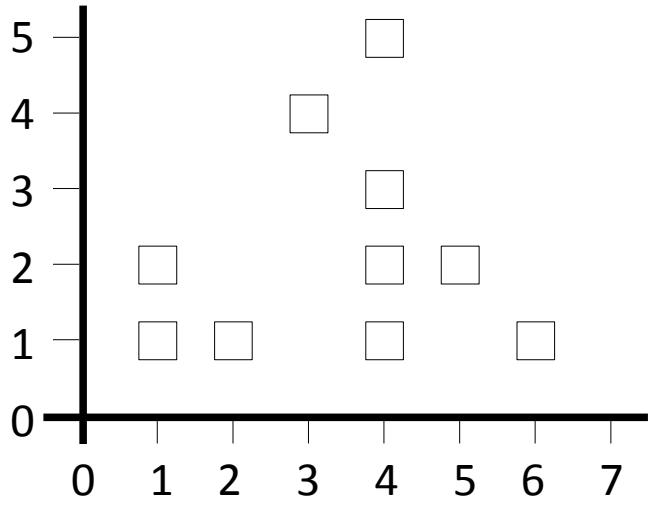
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :

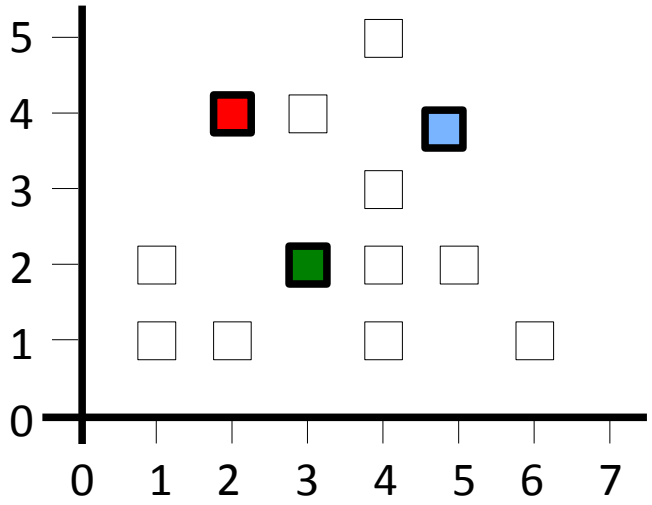


k-moyennes

Idée :

- **Choisir *k* points au hasard, et les considérer comme des centroïdes**
- Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde
- Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

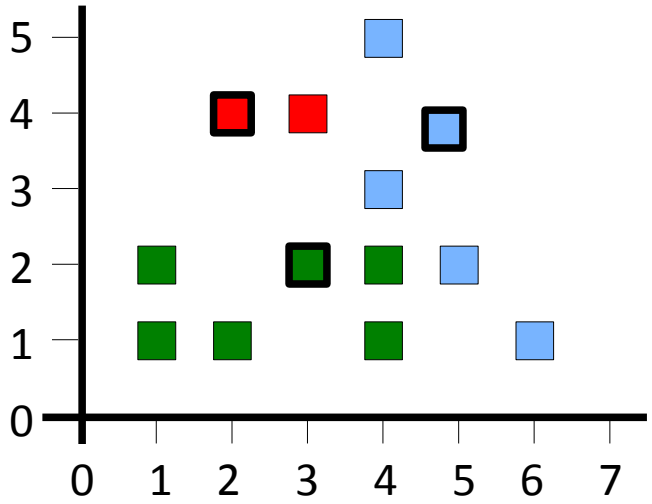
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• **Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde**

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

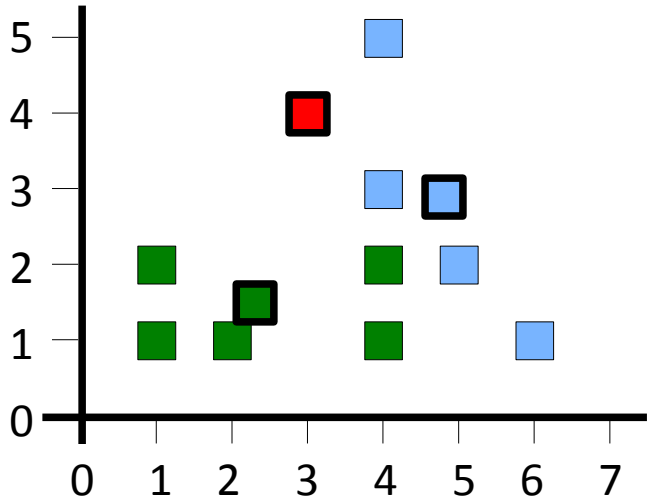
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• **Utiliser les barycentres des classes comme nouveaux centroïdes** et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

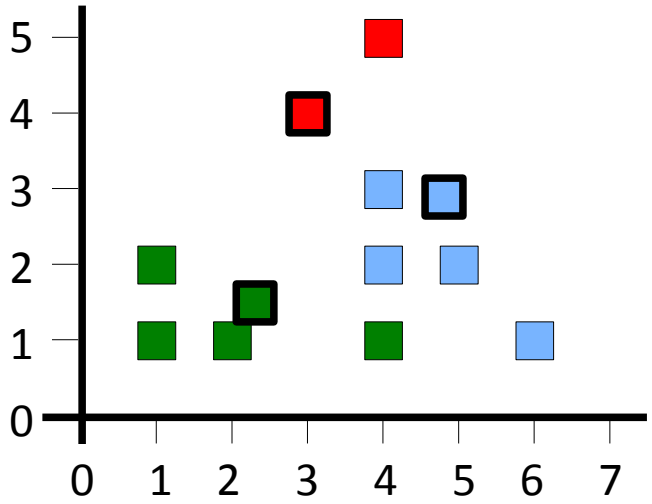
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• **Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde**

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

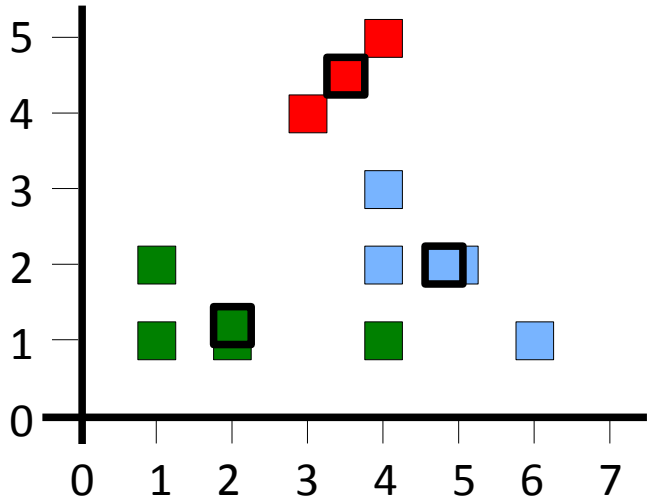
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• **Utiliser les barycentres des classes comme nouveaux centroïdes** et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

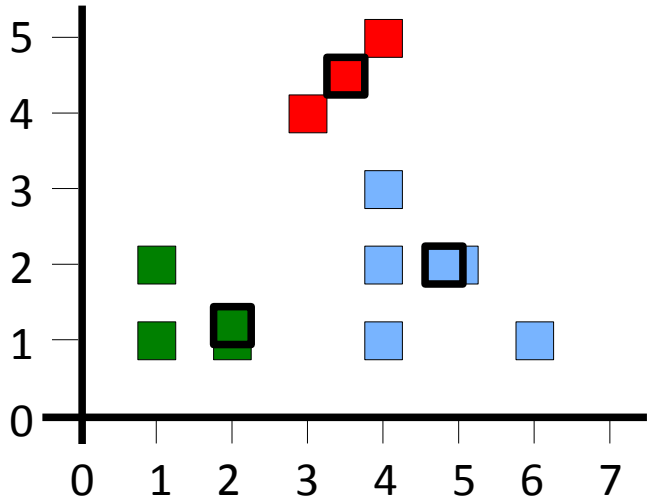
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• **Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde**

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

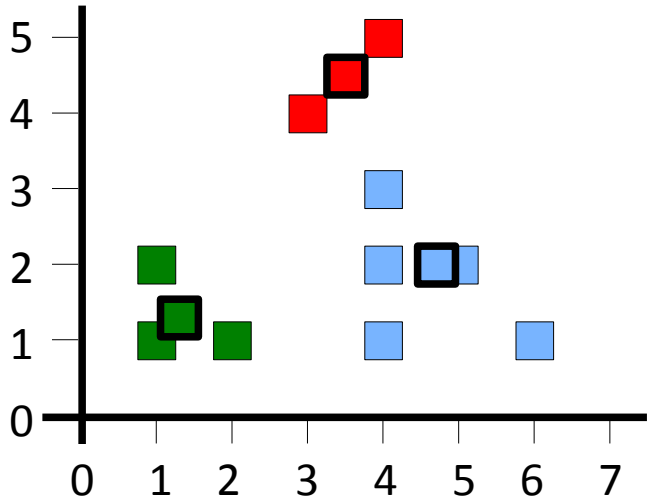
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• **Utiliser les barycentres des classes comme nouveaux centroïdes** et répéter jusqu'à ce qu'il n'y ait plus de changement.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

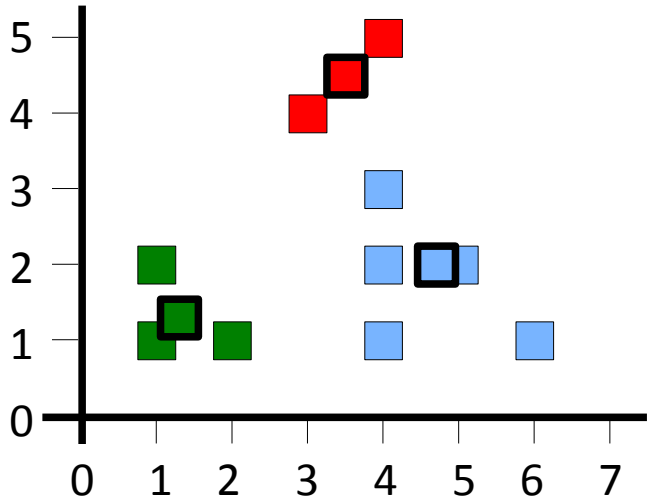
Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait **plus de changement**.

Exemple en dimension 2 avec *k*=3 :



k-moyennes

Idée :

• Choisir *k* points au hasard, et les considérer comme des centroïdes

• Distribuer les points dans les *k* classes ainsi formées selon leur proximité au centroïde

• Utiliser les barycentres des classes comme nouveaux centroïdes et répéter jusqu'à ce qu'il n'y ait **plus de changement**.

Exemple en dimension 2 avec *k*=3 :

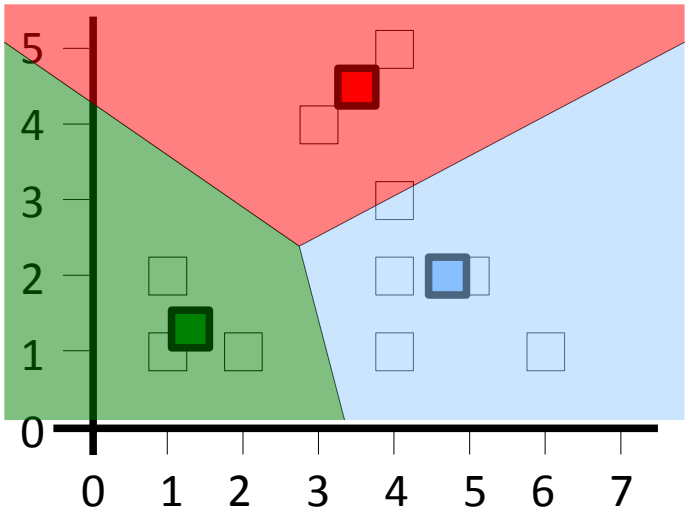


Diagramme de Voronoi

Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Classification hiérarchique

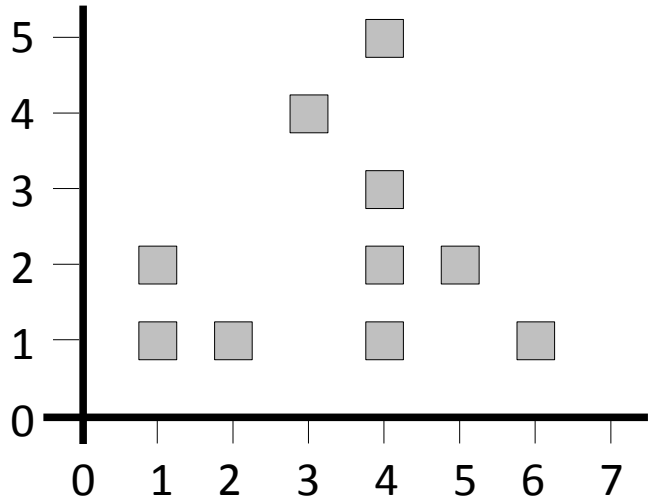
Idée :

- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :



Classification hiérarchique

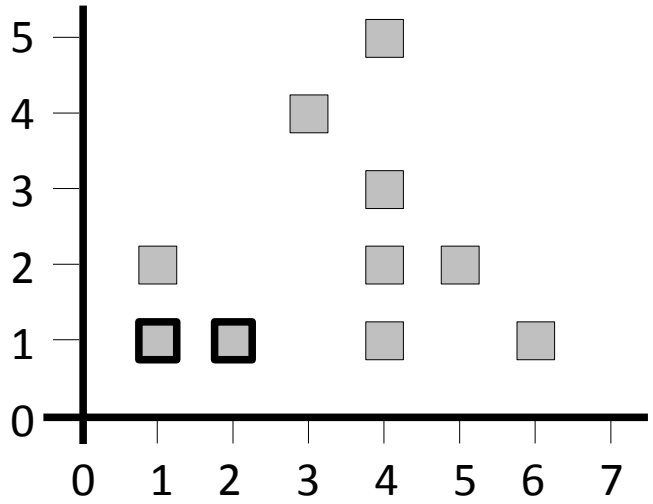
Idée :

- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :



Classification hiérarchique

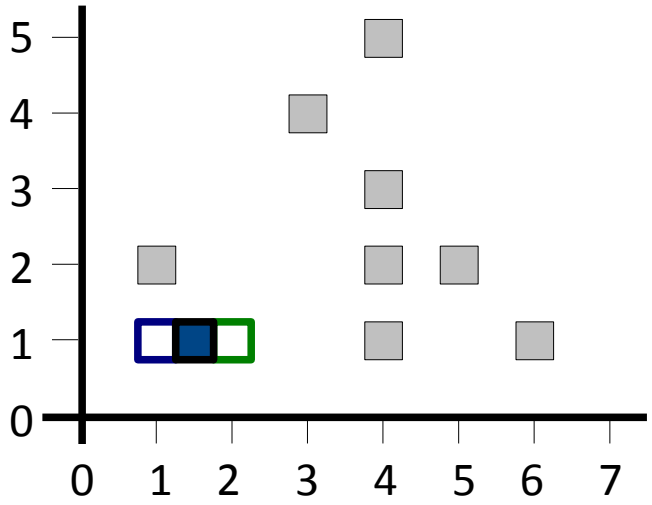
Idée :

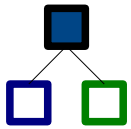
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

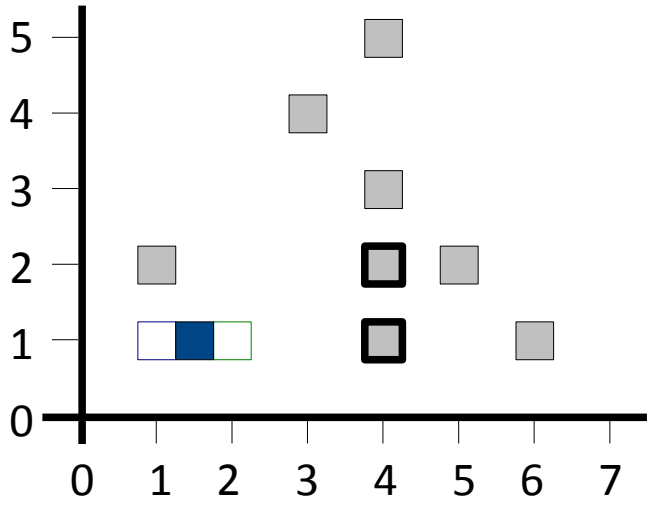
Idée :

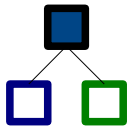
- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

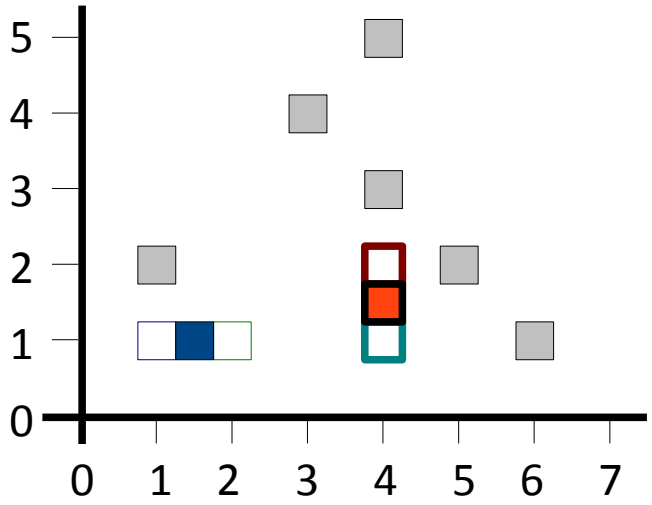
Idée :

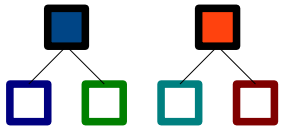
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

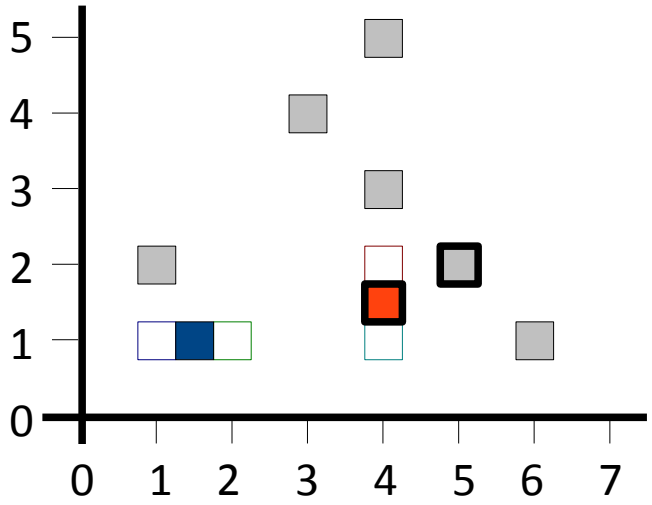
Idée :

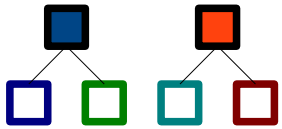
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

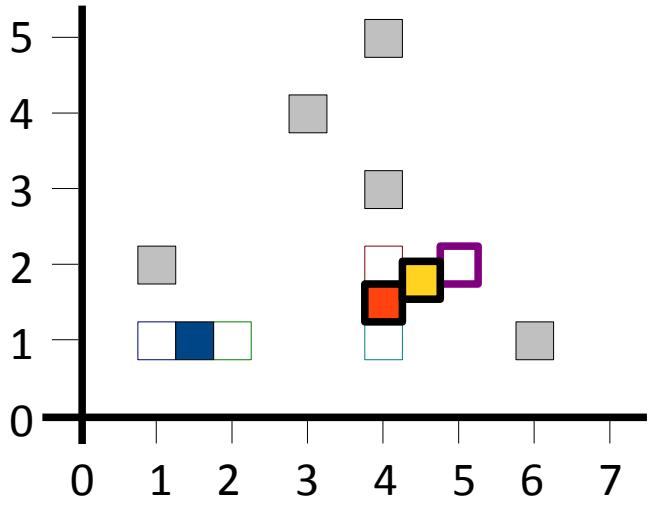
Idée :

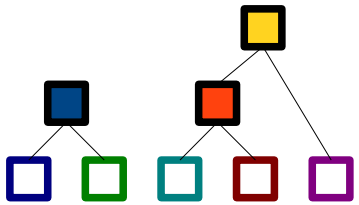
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

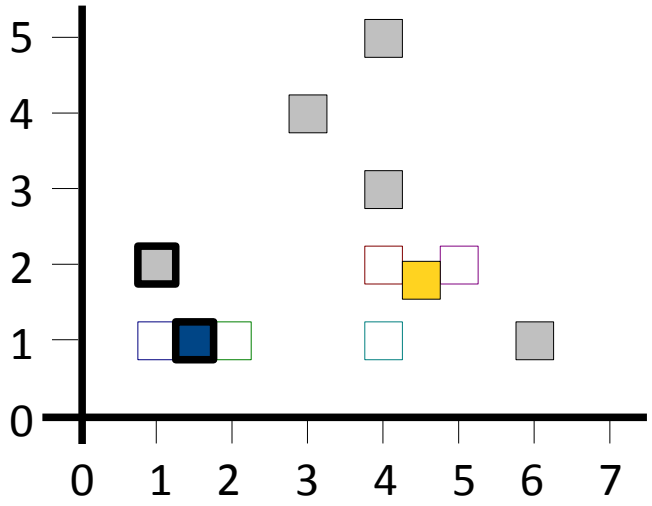
Idée :

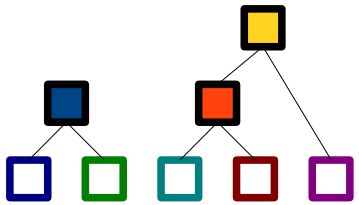
- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

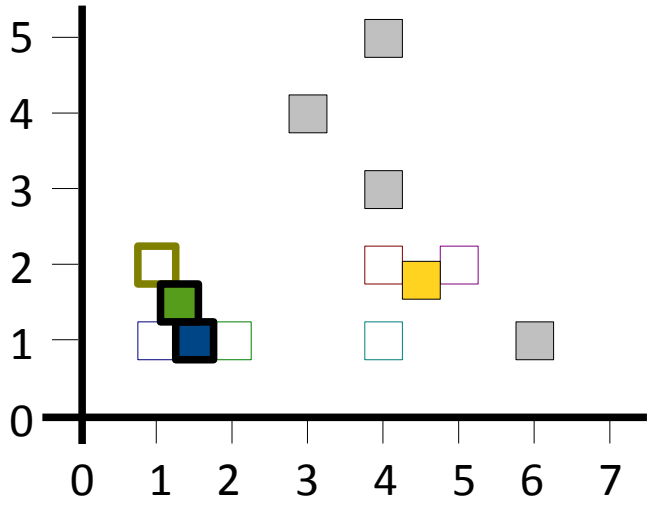
Idée :

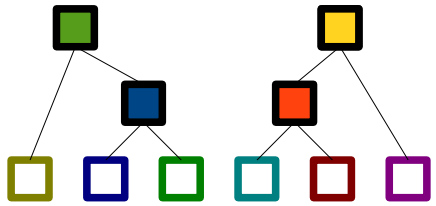
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

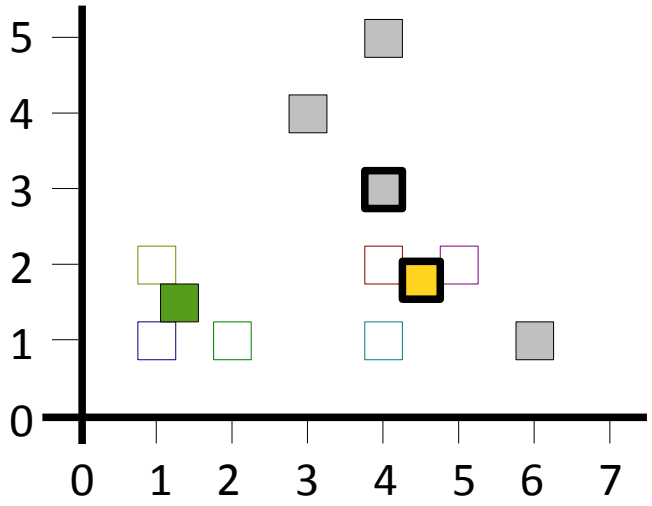
Idée :

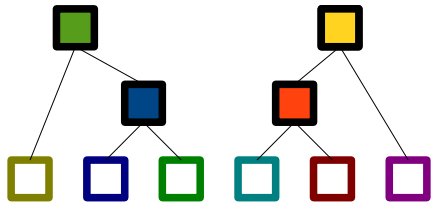
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

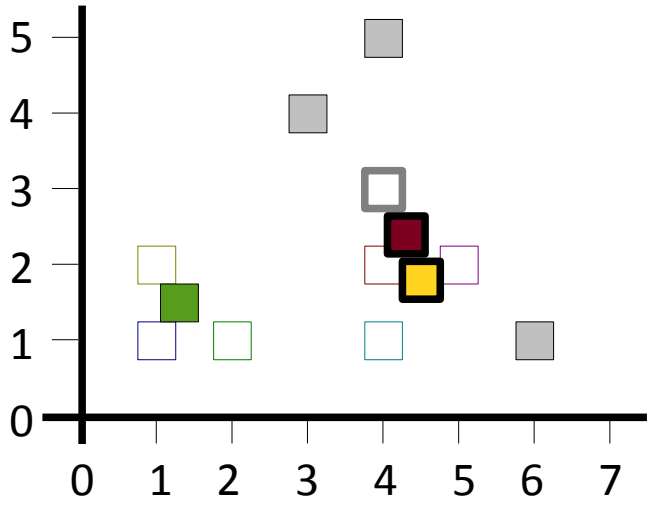
Idée :

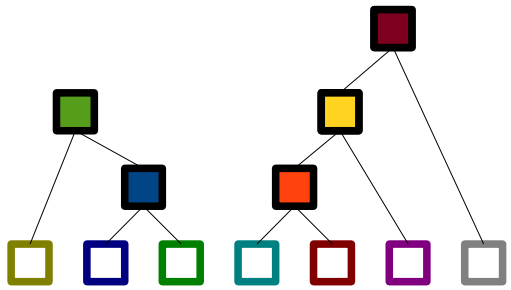
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

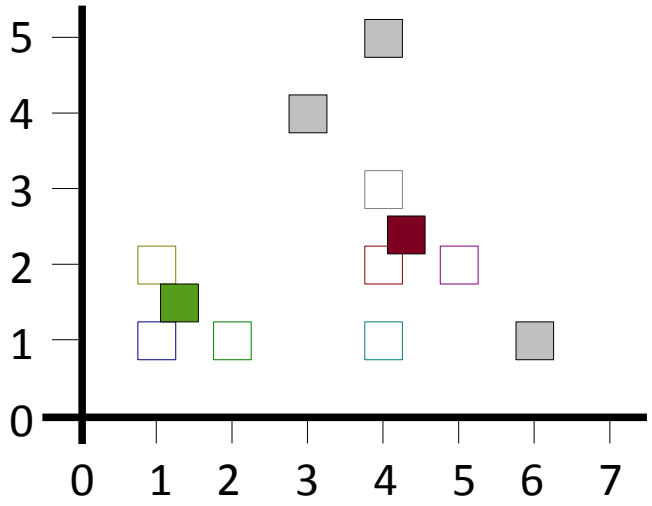
Idée :

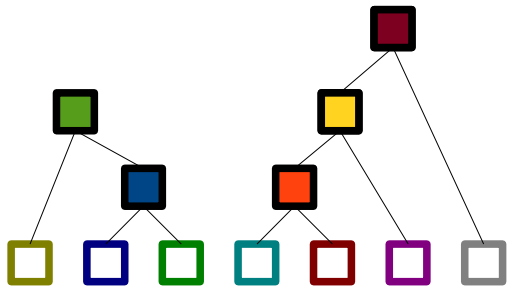
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

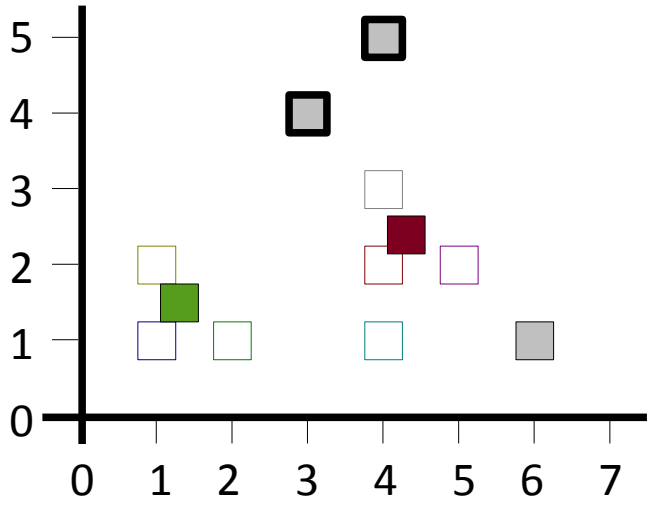
Idée :

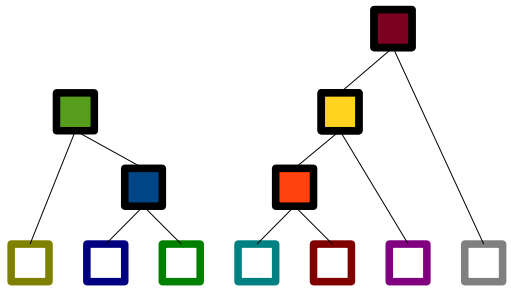
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

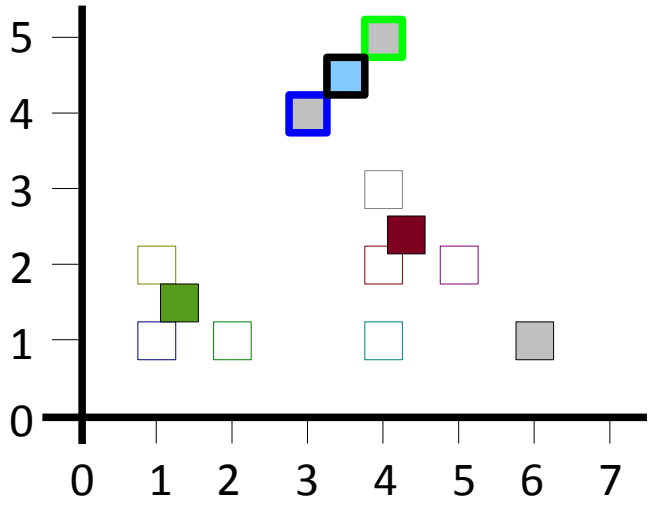
Idée :

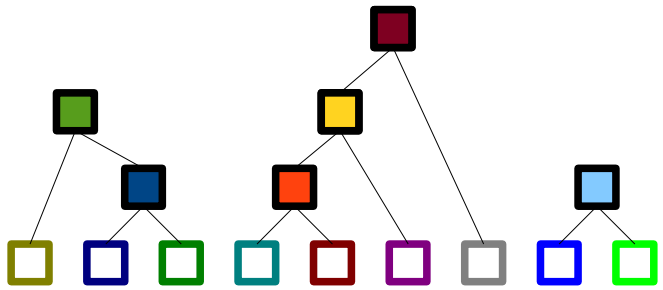
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

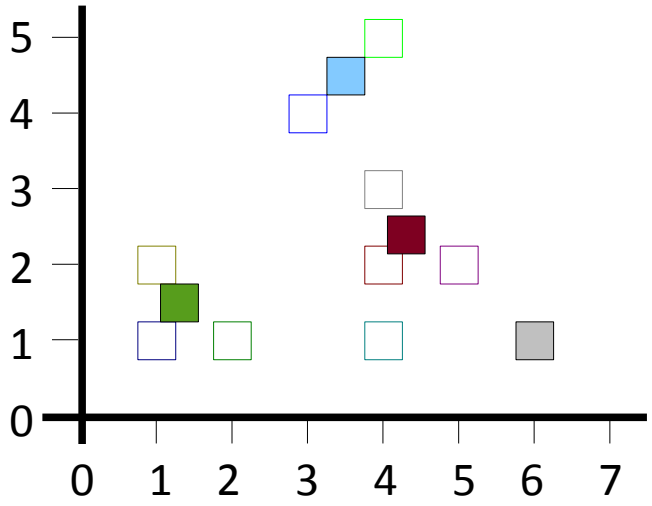
Idée :

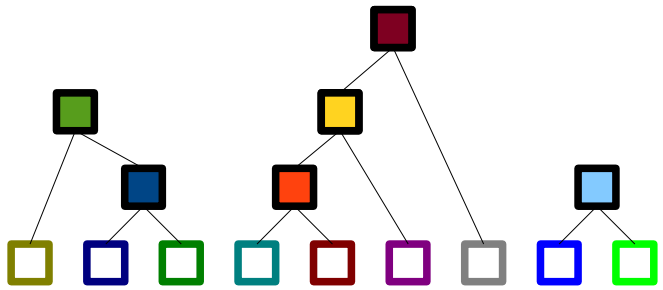
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

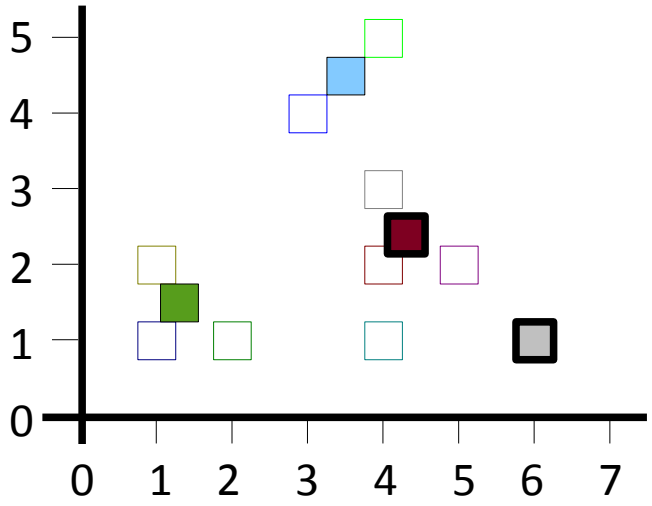
Idée :

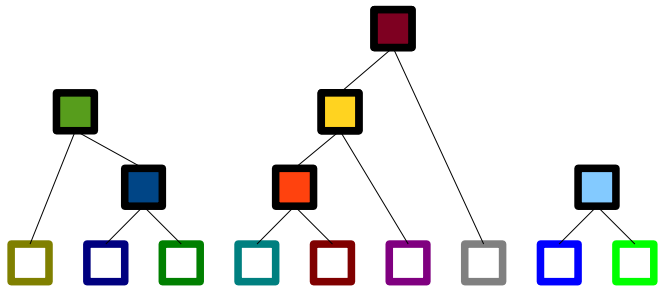
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

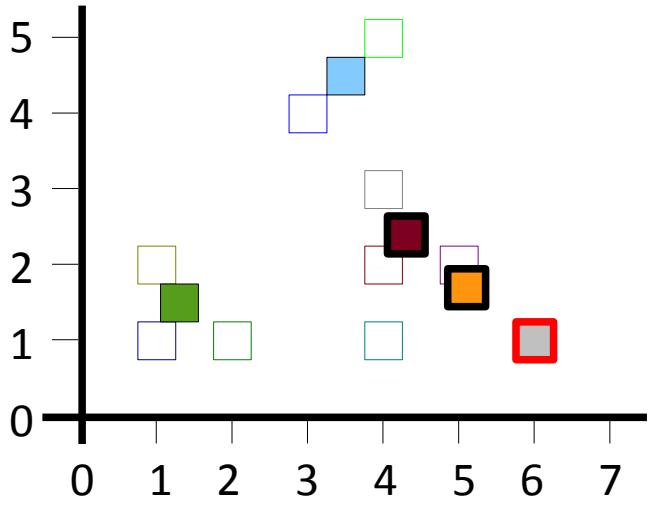
Idée :

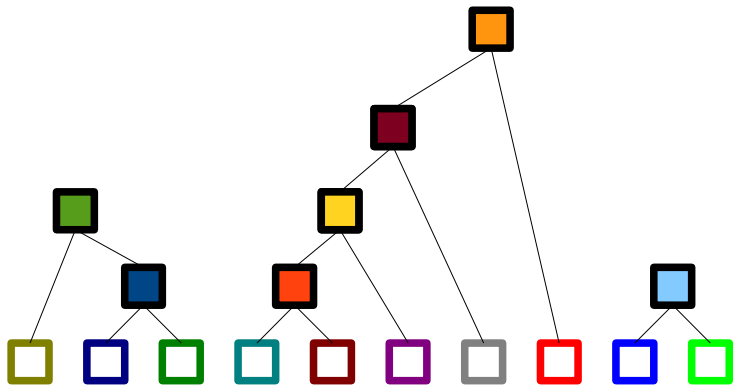
• Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.

• Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

• Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

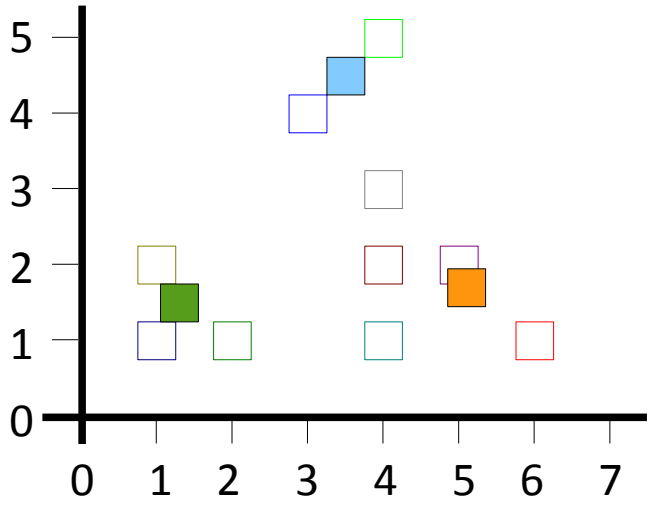
Idée :

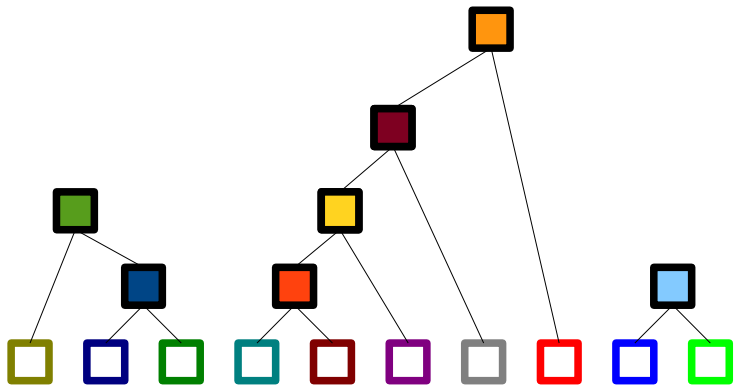
- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :





Classification hiérarchique

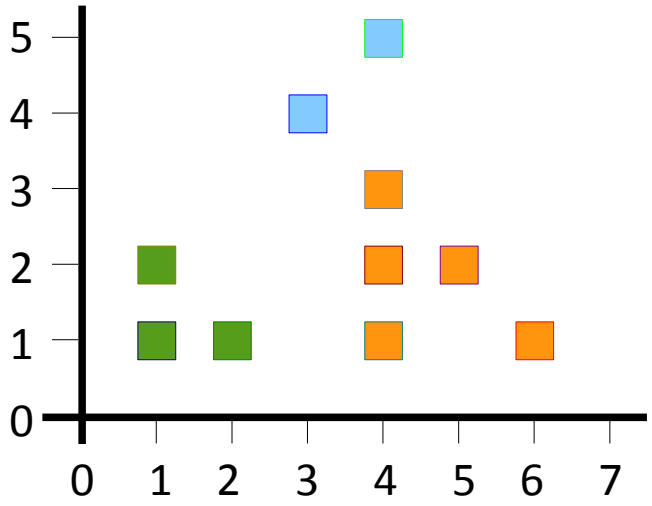
Idée :

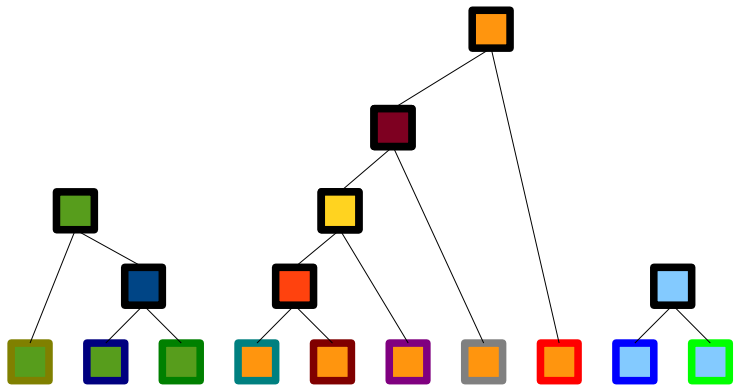
- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**

- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)

- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :



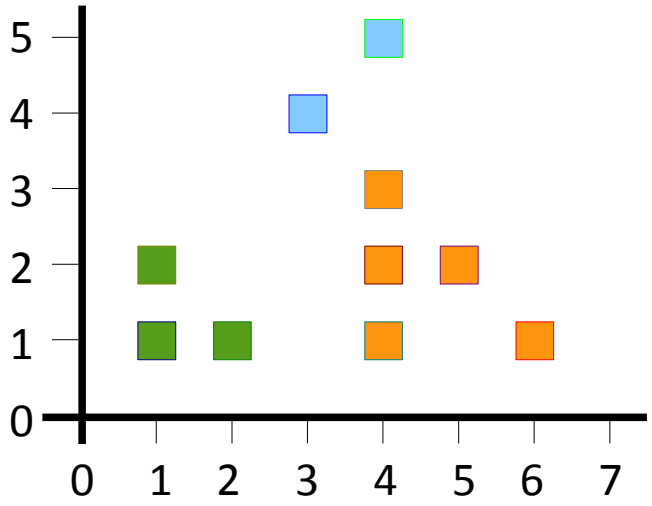


Classification hiérarchique

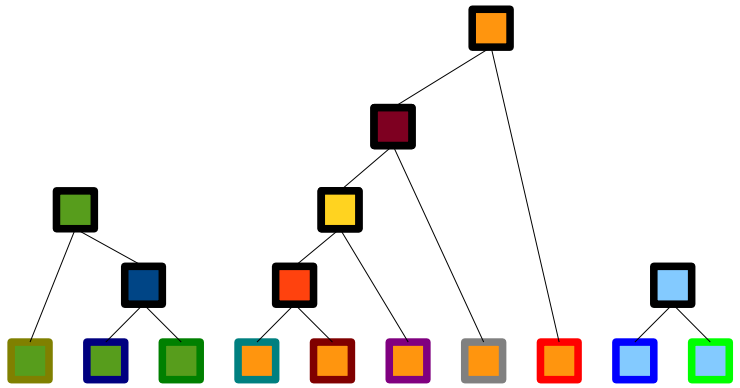
Idée :

- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**
- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)
- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :



Fonctionne aussi si on n'a pas les coordonnées des points mais seulement les **distances entre paires de points !**

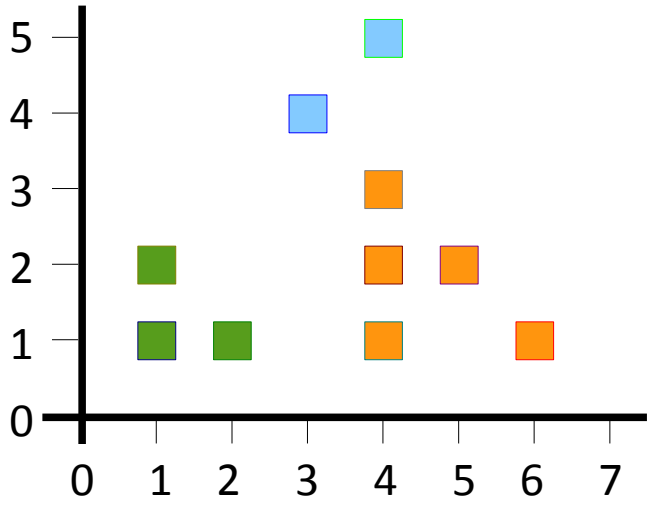


Classification hiérarchique

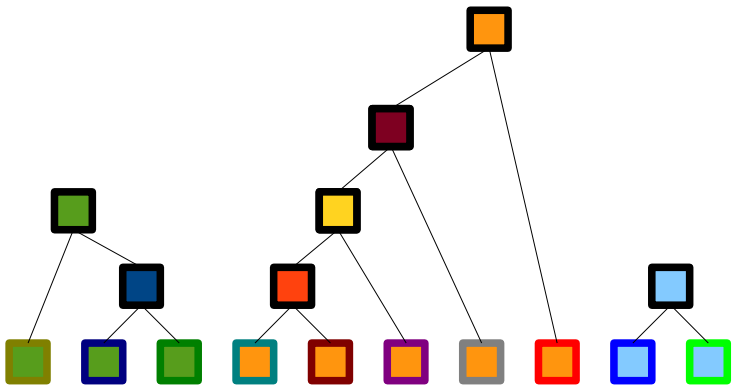
Idée :

- **Prendre les deux points les plus proches, les fusionner, considérer leur moyenne par la suite.**
- Répéter jusqu'à un critère d'arrêt (par exemple distance supérieure à une certaine valeur)
- Ou bien découper l'arbre des fusions pour obtenir des classes

Exemple en dimension 2 :



Arbre construit de bas en haut
Méthode bottom-up



Classification hiérarchique

Variantes de la classification hiérarchiques

Pour calculer la **distance entre deux classes C1 et C2** :

• calculer la **moyenne** des distances entre les éléments de C1 et les éléments de C2

• calculer le **minimum** des distances entre les éléments de C1 et les éléments de C2

• calculer le **maximum** des distances entre les éléments de C1 et les éléments de C2

Méthodes par **division** (top-down) plutôt que par **agglomération** (bottom-up)

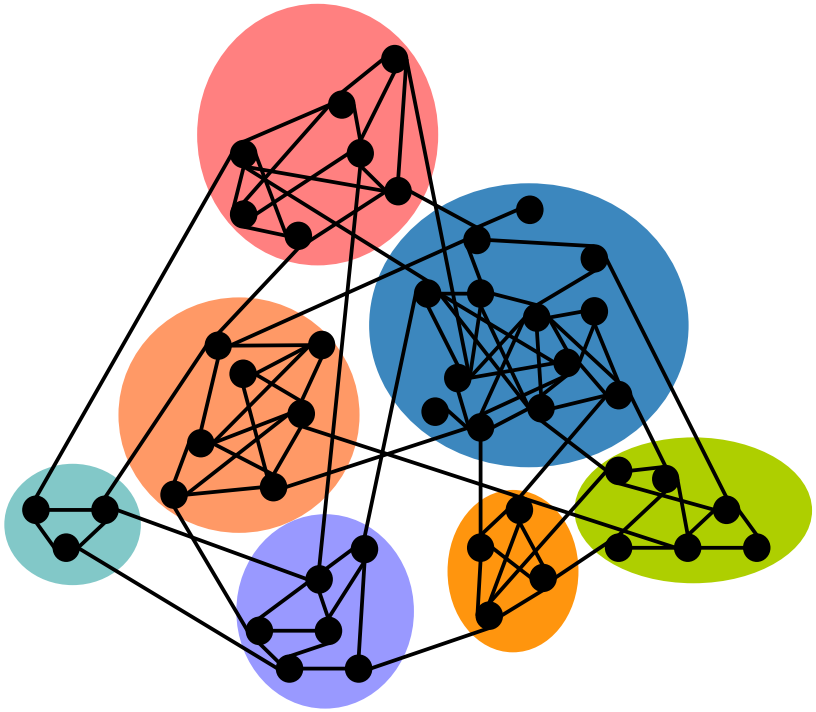
Plan

- Introduction
- Classification supervisée de documents
- Approche du centroïde
- *k*-plus proches voisins
- Classifieurs linéaires et SVM
- Classification non supervisée
- *k*-moyennes
- Classification hiérarchique
- Partitionnement de graphes et modularité

Partitionnement de graphes

Classification non supervisée des sommets d'un graphe :

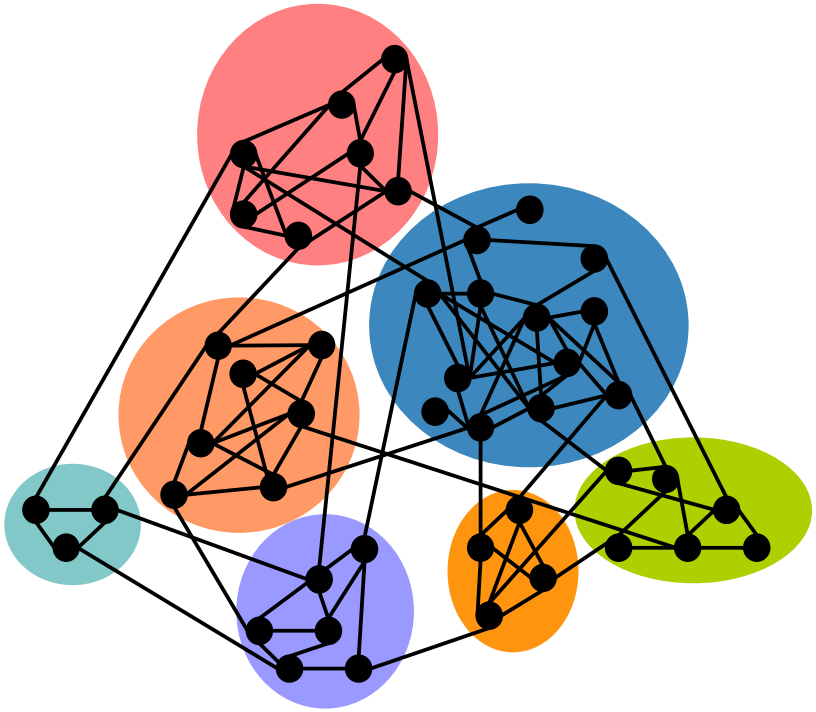
Une arête relie deux sommets à petite distance (forte similarité).



Partitionnement de graphes et modularité

Classification non supervisée des sommets d'un graphe :

Une arête relie deux sommets à petite distance (forte similarité).



Score de qualité du partitionnement ?

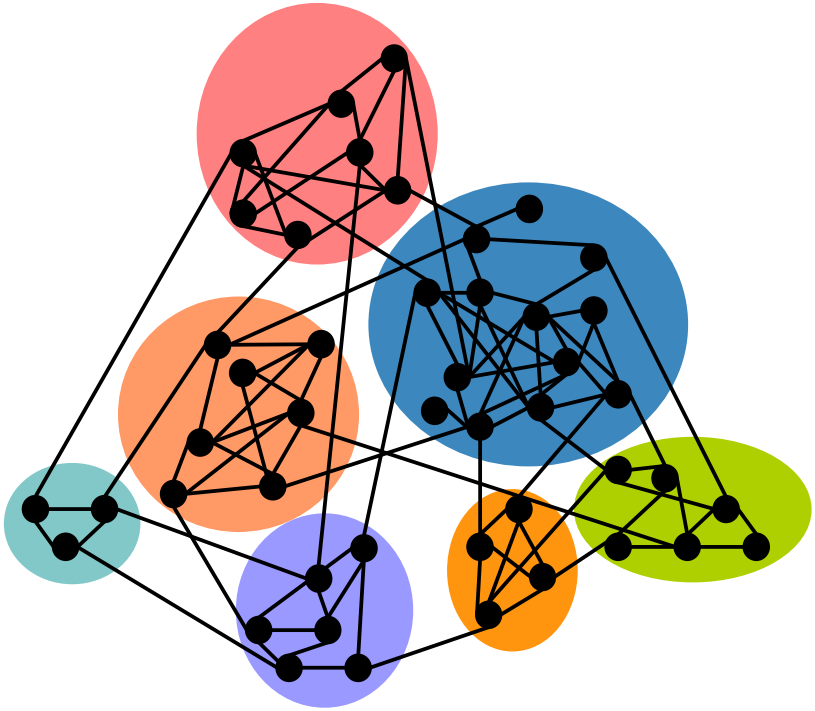
La *modularité*

Girvan & Newman 2004

Pas besoin de paramètres : nombre de classes et tailles des classes fixées automatiquement par l'optimisation de la modularité

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : qualité du partitionnement

Girvan & Newman 2004

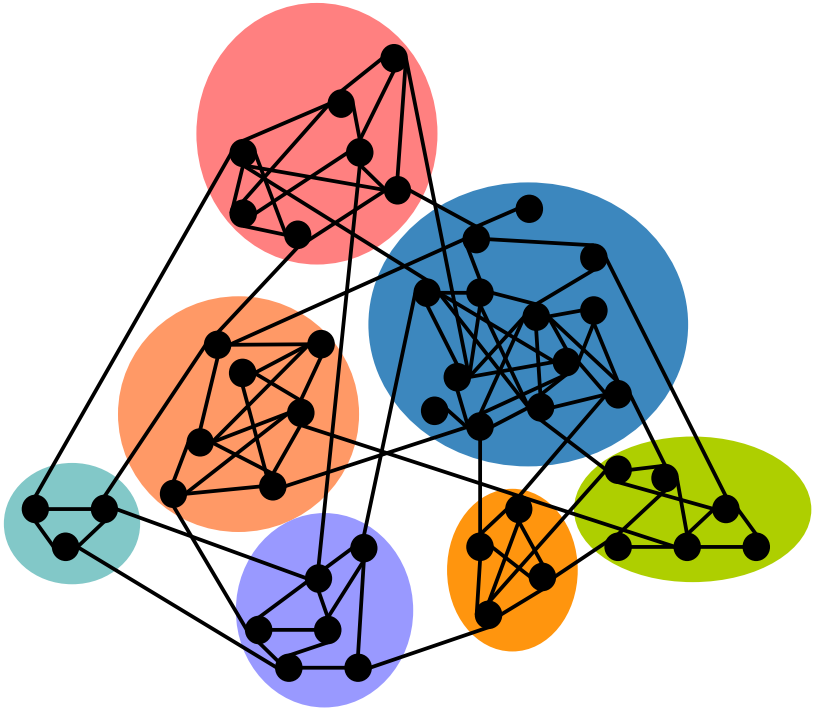
$$M(G,P) = \sum_{C_i \in P} M(C_i)$$

$$M(C_i) = e_{ii} - \left(e_{ii} + \sum_{j \neq i} e_{ij}/2\right)^2$$

e_{ij} = proportion d'arêtes avec un sommet dans C_i et l'autre dans C_j

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : qualité du partitionnement

Girvan & Newman 2004

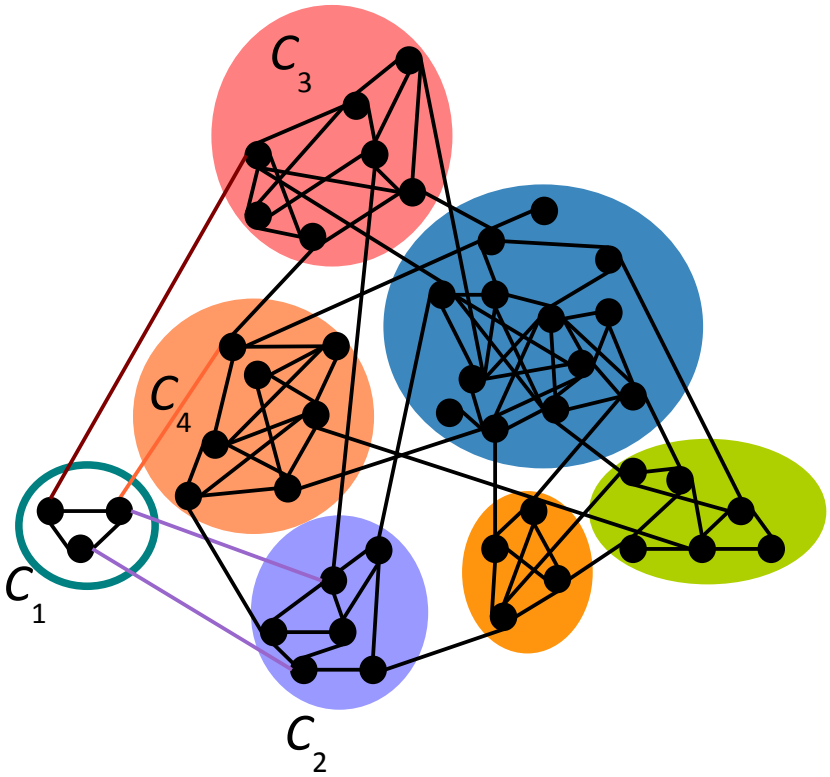
$$M(G,P) = \sum_{C_i \in P} M(C_i)$$

$$M(C_i) = \underbrace{e_{ii}}_{\text{proportion d'arêtes observées dans la classe } C_i} - \underbrace{\left(e_{ii} + \sum_{j \neq i} e_{ij}/2\right)^2}_{\text{proportion d'arêtes attendues dans la classe } C \text{ s'il n'y avait pas de communauté, et répartition au hasard en respectant les degrés}}$$

e_{ij} = proportion d'arêtes avec un sommet dans C_i et l'autre dans C_j

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



$e_{11} = 3$; $e_{12} = 2$; $e_{13} = 1$; $e_{14} = 1$

$M(C_1) = 3/100 - (5/100)^2$

$= 0.0275$

Modularité : qualité du partitionnement

Girvan & Newman 2004

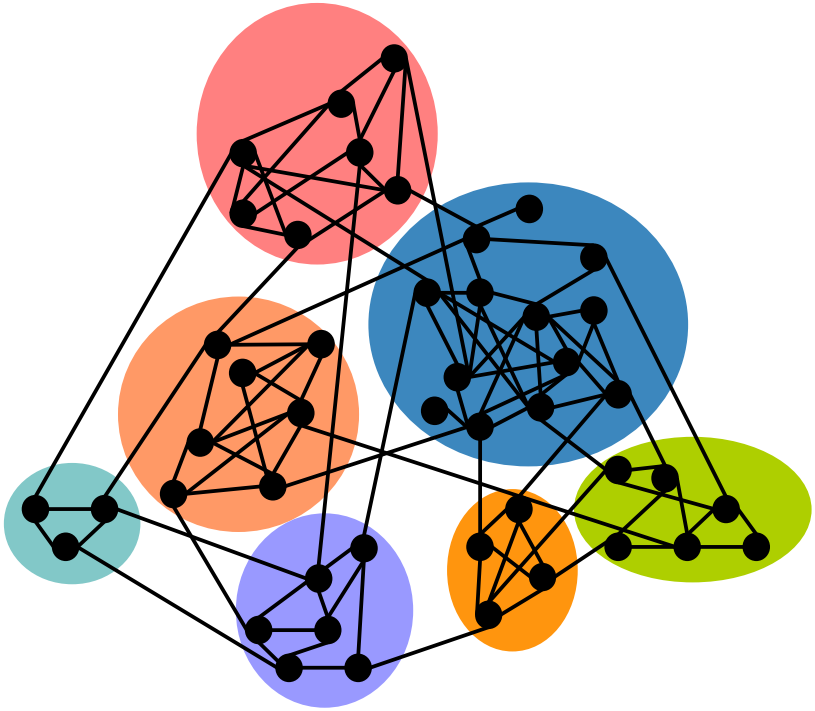
$$M(G,P) = \sum_{C_i \in P} M(C_i)$$

$$M(C_i) = \underbrace{e_{ii}}_{\text{proportion d'arêtes observées dans la classe } C_i} - \underbrace{\left(e_{ii} + \sum_{j \neq i} e_{ij}/2\right)^2}_{\text{proportion d'arêtes attendues dans la classe } C \text{ s'il n'y avait pas de communauté, et répartition au hasard en respectant les degrés}}$$

e_{ij} = proportion d'arêtes avec un sommet dans C_i et l'autre dans C_j

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : formule équivalente

Newman 2004

$$M(G,P) = \frac{1}{2m} \sum_{u,v} (G_{uv} - \frac{d(u)d(v)}{2m})\alpha_{uv}$$

G_{uv} = 1 si u et v adjacents

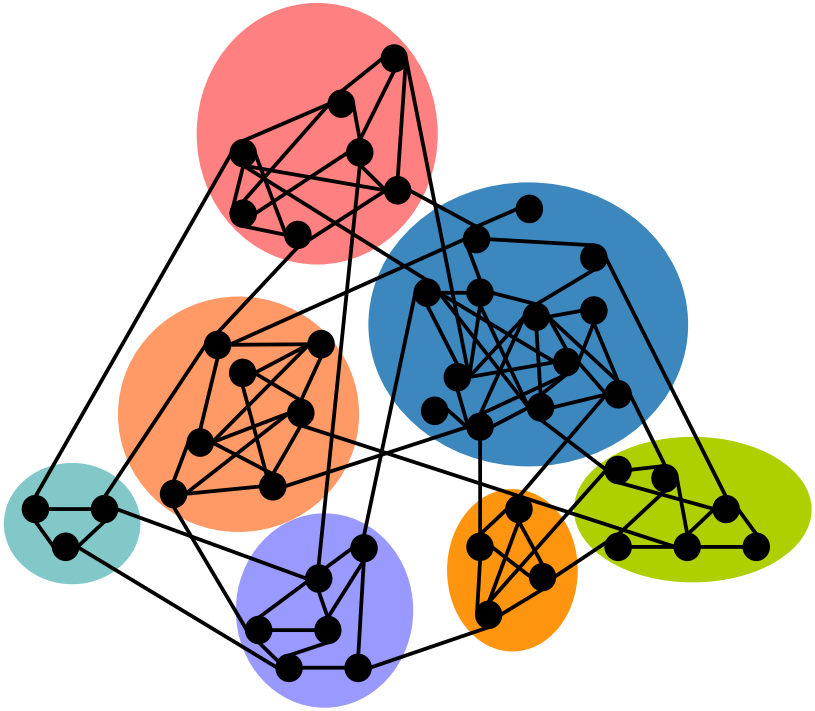
0 sinon

α_{uv} = 1 si u et v dans la même classe

0 sinon

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Possibilité d'étendre aux graphes aux arêtes pondérées

Modularité : formule équivalente

Newman 2004

$$M(G,P) = \frac{1}{2m} \sum_{u,v} \left(G_{uv} - \frac{d(u)d(v)}{2m} \right) \alpha_{uv}$$

G_{uv} = 1 si u et v adjacents

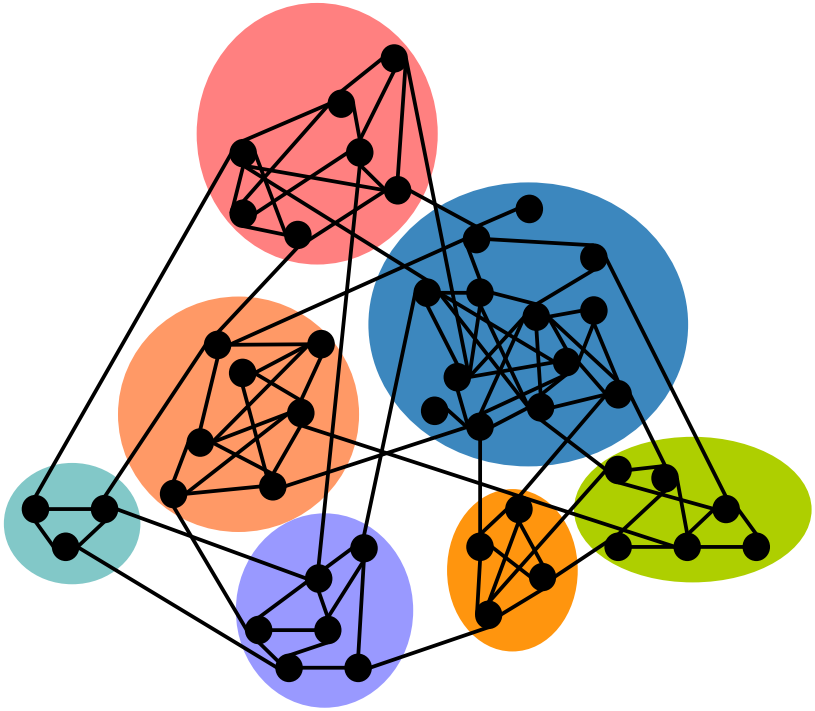
0 sinon

α_{uv} = 1 si u et v dans la même classe

0 sinon

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : formule équivalente

Newman 2004

$$M(G,P) = \frac{1}{2m} \sum_{u,v} \underbrace{\left(G_{uv} - \frac{d(u)d(v)}{2m}\right)}_{w_{uv}} \alpha_{uv}$$

G_{uv} = 1 si u et v adjacents

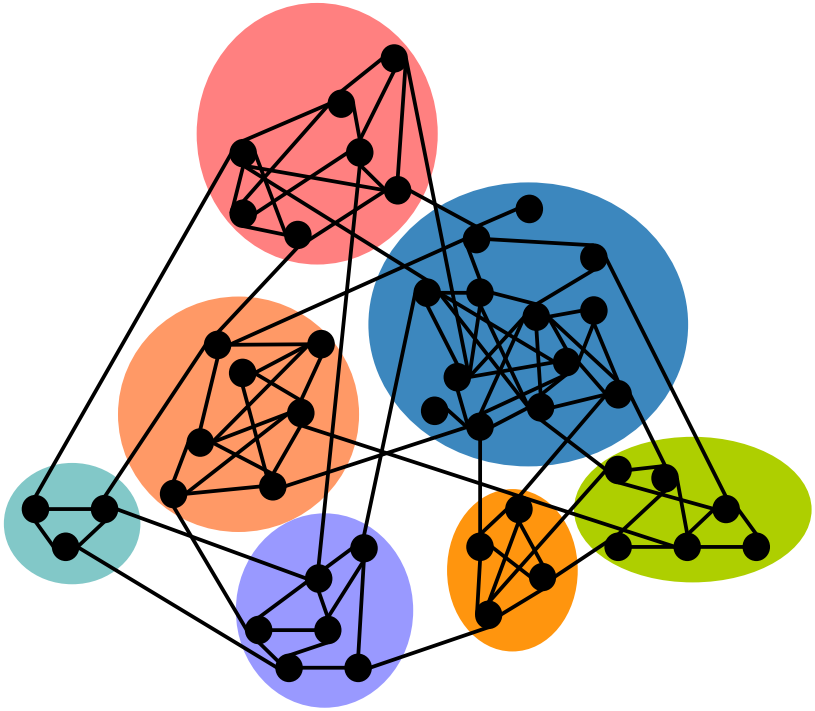
0 sinon

α_{uv} = 1 si u et v dans la même classe

0 sinon

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : formule équivalente

Newman 2004

$$M(G,P) = \frac{1}{2m} \sum_{u,v} \underbrace{\left(G_{uv} - \frac{d(u)d(v)}{2m}\right)}_{w_{uv}} \alpha_{uv}$$

Problème d'optimisation **NP-complet**

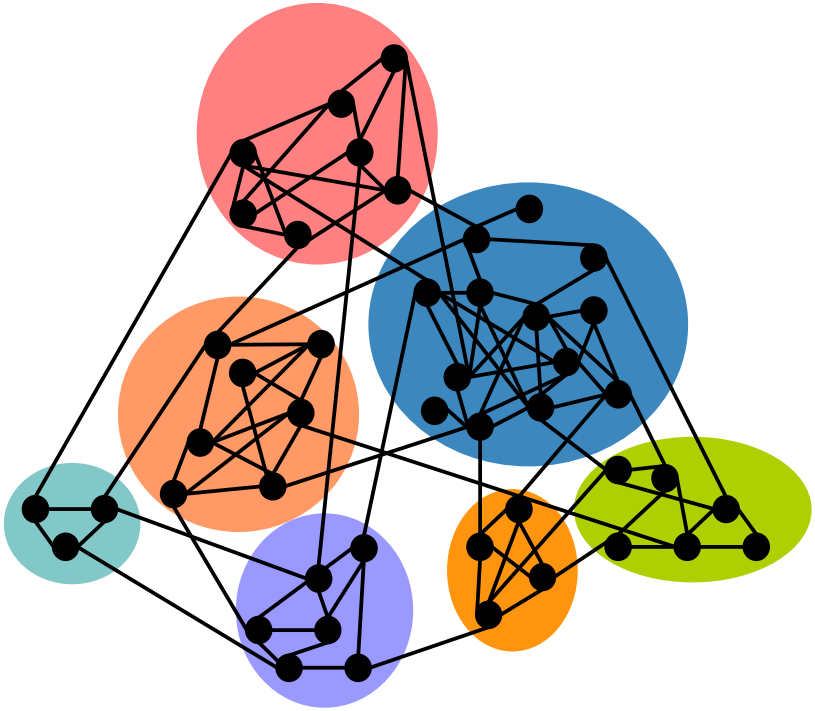
Brandes et al. 2008

- Même si la solution recherchée a 2 classes et que le graphe est peu dense

DasGupta & Desai, 2011

Partitionnement de graphes et modularité

Partitionnement d'un réseau : couvrir tous les sommets par des classes disjointes



Modularité : formule équivalente

Newman 2004

$$M(G,P) = \frac{1}{2m} \sum_{u,v} \underbrace{\left(G_{uv} - \frac{d(u)d(v)}{2m}\right)}_{w_{uv}} \alpha_{uv}$$

Problème d'optimisation **NP-complet**

Brandes et al. 2008

- Même si la solution recherchée a 2 classes et que le graphe est peu dense

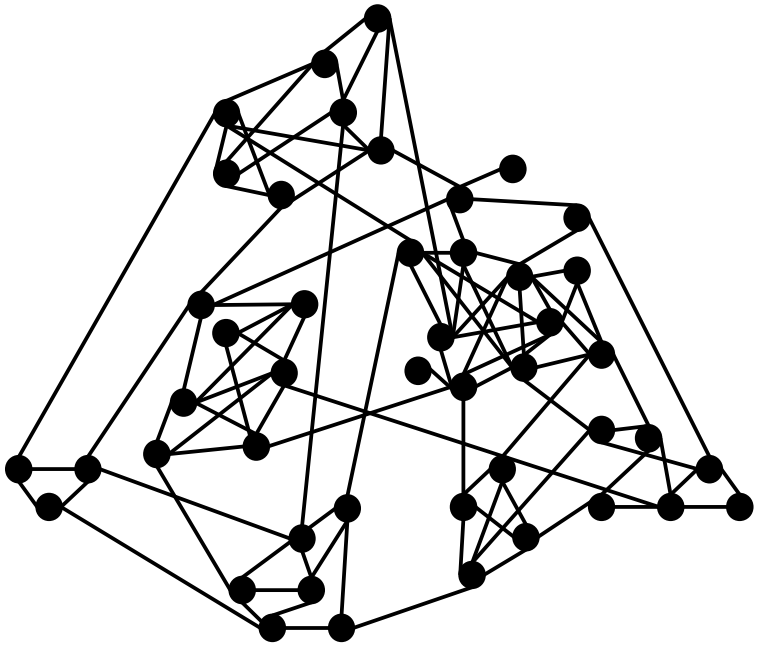
DasGupta & Desai, 2011

- Heuristiques très rapides (traitent des graphes de millions de sommets)

Blondel et al. 2008

Heuristiques d'optimisation de la modularité

Base similaire à la classification hiérarchique :

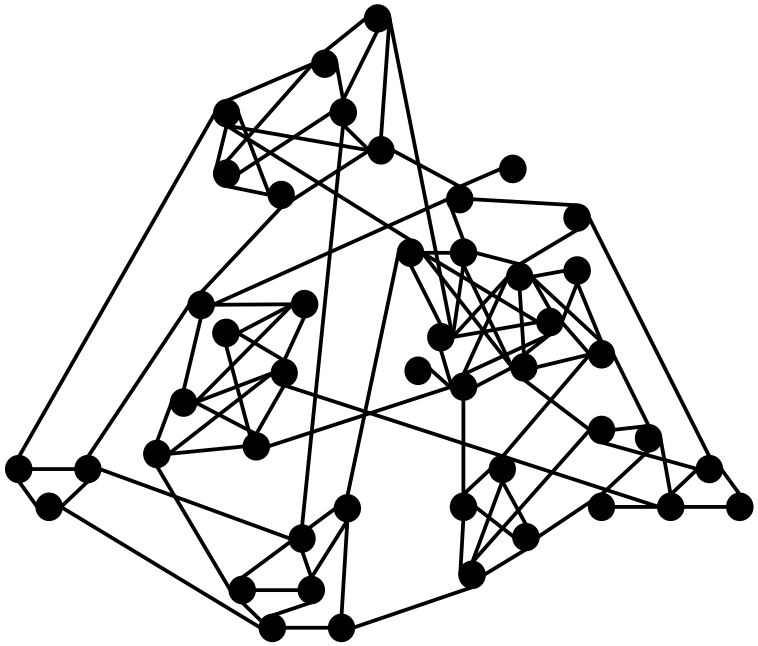


Tant qu'on augmente le score de modularité :

On transfère un sommet dans la classe qui fait le plus augmenter la modularité.

Heuristiques d'optimisation de la modularité

Base **similaire à la classification hiérarchique** :



Tant qu'on augmente le score de modularité :

> On transfère un sommet dans la classe qui fait le plus augmenter la modularité.

Si aucun transfert de sommet ne fait augmenter la modularité :

> On **contracte les sommets en un seul représentant** et on **recommence l'algorithme**.