Projet Industriel

Exploration des méthodes d'apprentissage semi-supervisé

10-02-2016

- Yassine LANDA | 3A ICM – Informatique & Science des données
- Olivier ROUSTANT | Tuteur EMSE
- Michel LUTZ | Tuteur OCTO

Plan

1. Introduction à l'apprentissage Semi-supervisé

2. Algorithmes d'apprentissage Semi-supervisé
   - Generative Models
   - Self Training
   - Graph-Based Algorithms
   - S3VM

3. Cas Pratique

4. Conclusion

Pourquoi l'apprentissage semi-supervisé?

Parce qu'on cherche la performance au moindre coût.

Le point de vue traditionnel

- Les données non étiquetées sont disponibles.
- Les données étiquetées sont difficiles à obtenir.
- Une opération ennuyante !
- Peut nécessiter des experts.
- Peut nécessiter un matériel spécifique.

Exemple d'application

Web Page Classification

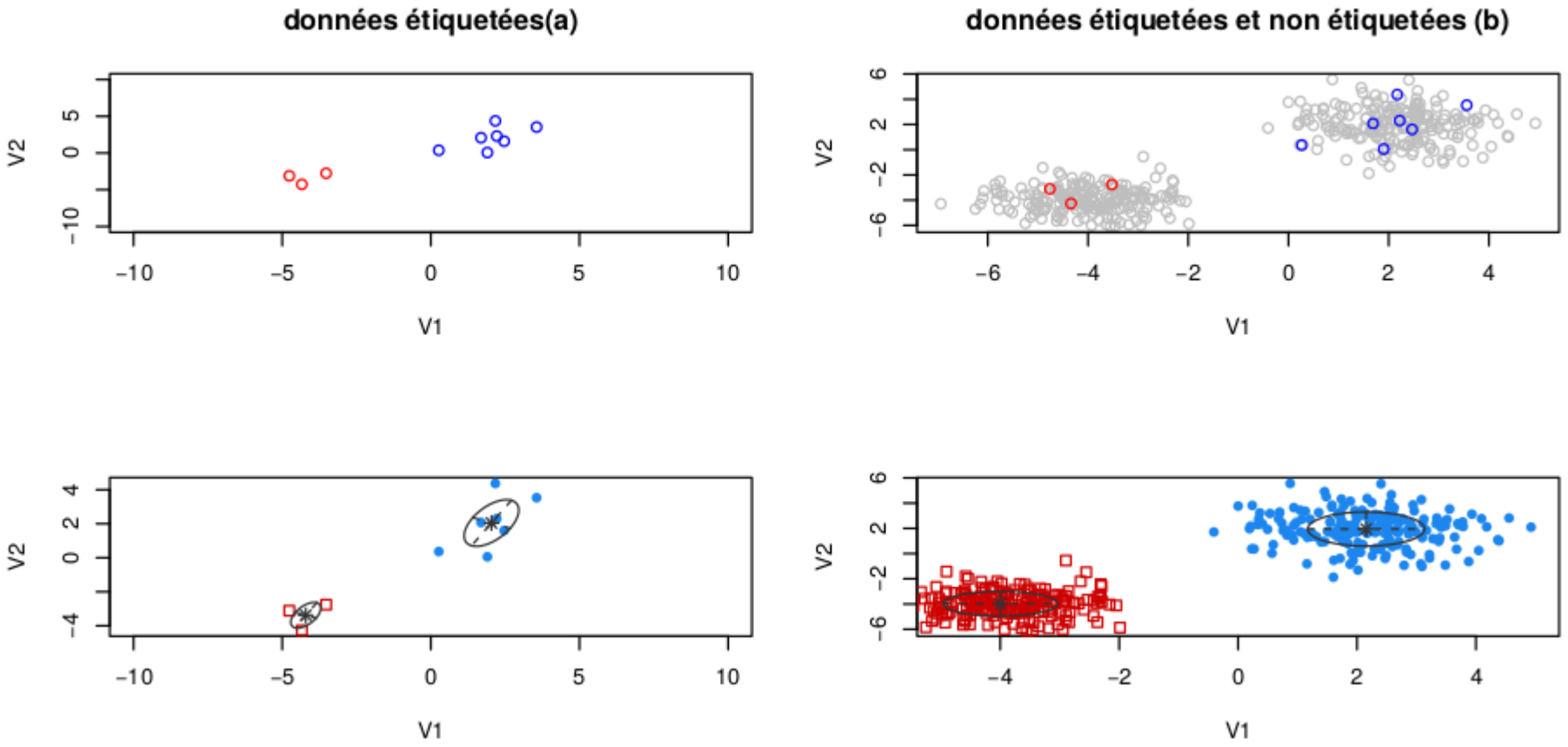
"Educational", " Shopping", "Forum"...

- Coûteux d'annoter une dizaine de milliers de pages par un humain.

- Abondance des pageweb (un crawler pour récupérer des millions de pages en quelques Heures).

Apprentissage semi-supervisé

Exemple simple



Le problème d'apprentissage

Objectif

Utiliser l'ensemble des données étiquetées et l'ensemble des données non étiquetées pour construire de meilleurs classificateurs, que si on utilise chacun des ensembles seul.

Notations

- Donnée d'entrée x, étiquette y
- Fonction objectif $f : X \rightarrow Y$
- Données étiquetées $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- Données non étiquetées $X_u = \{x_{l+1:n}\}$, *disponible durant l'apprentissage*
- En général $l << n$
- Donnée Test $X_{test} = \{x_{n+1:}\}$, *non disponible durant l'apprentissage*

Generative methods

Hypothèse

Le modèle génératif entier $p(X,Y|\theta)$.

Somme de distributions Gaussiennes (GMM)
- Classification d'images
- EM algorithm

Somme de distributions multinomiales (Näive Bayes)
- Catégorisation de texte
- EM algorithm

Hidden Markov Models (HMM)
- Reconnaissance vocale
- Baum-Welch algorithm

Generative methods

Pour la simplicité, on considère un classificateur binaire avec une somme de distributions Gaussiennes utilisant l'estimation du maximum de vraisemblance (MLE).

Données étiquetées seulement:

$$\log p(X_l, Y_l|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i,\theta)$$

- MLE pour θ triviale

Données étiquetées et non étiquetées:

$$\log p(X_l, Y_l, X_u|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i,\theta) + \sum_{i=l+1}^{l+u} \log \left(\sum_{y=1}^{2} p(y|\theta)p(x_i|y,\theta) \right)$$

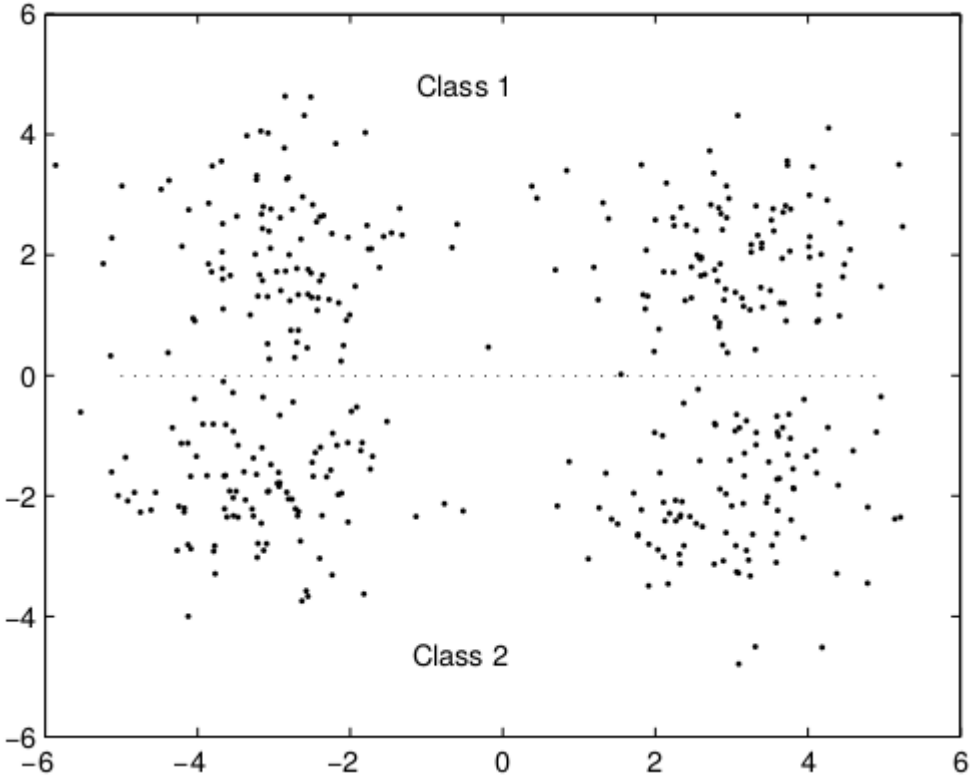
- MLE plus difficile (variables cachées)
- L'algorithme Expectation-Maximization (EM) est une solution pour trouver un maximum local.

Avantages des méthodes génératives

- Claires, ont une base probabiliste très bien étudiée.
- Peuvent être extrêmement efficace, si le modèle correspond au données.

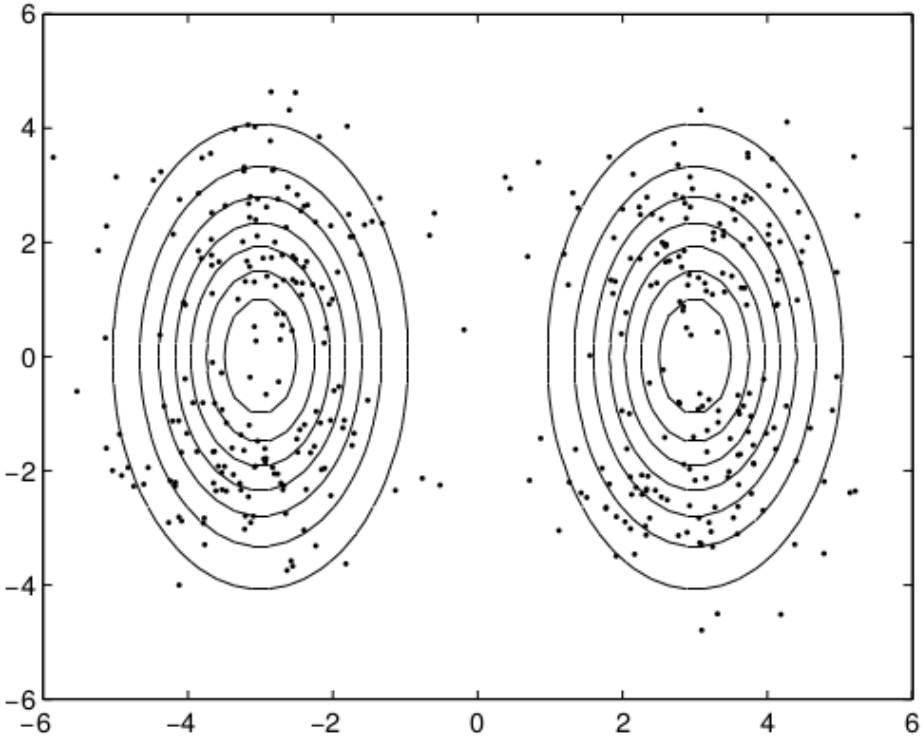
Inconvénients des méthodes génératives

- Souvent il est difficile de vérifier que le modèle est correcte.
- Locaux maximaux de l'algorithme EM.
- Les données non labellisées peuvent nuire aux résultat si le modèle génératif est erroné.

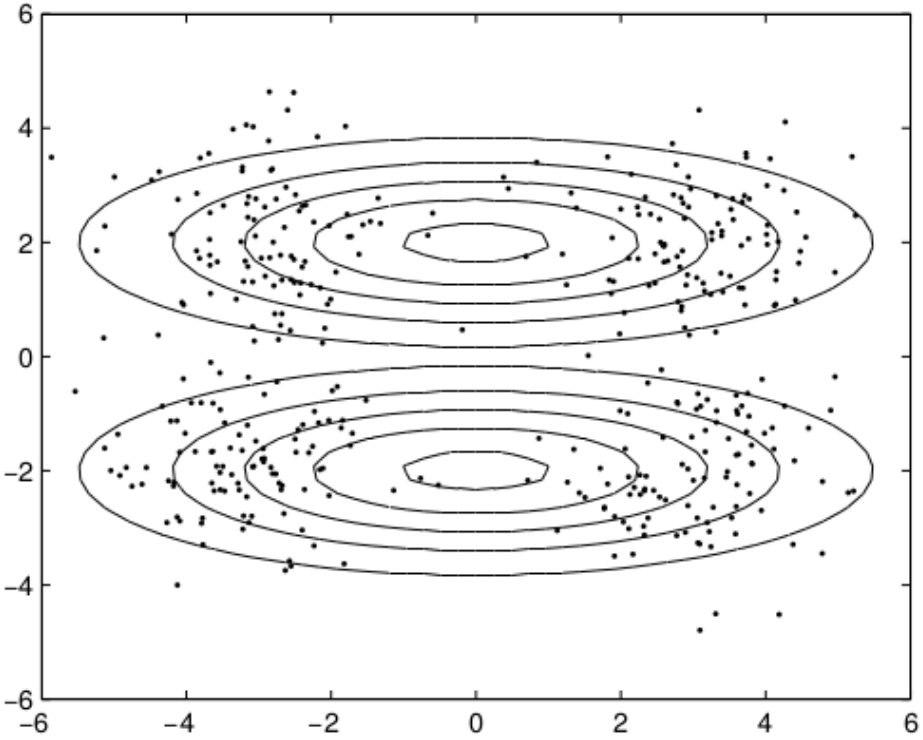


Inconvénients des méthodes génératives

high likelihood
wrong



low likelihood
correct



Images : “Semi-Supervised Learning” Olivier Chapelle,Bernhard Schölkopf, and Alexander Zien

Self-Training

Hypothèse

Les prédictions trouvés avec une confidence importante sont correctes.

L'algorithme Self-training:

1. Apprendre f sur (X_l, Y_l)
2. Prédire les $x \in X_u$
3. Ajouter $(x, f(x))$ aux données étiquetées
4. Répéter

Avantages du Self-Training

- La méthode d'apprentissage semi-supervisée la plus simple.
- Une méthode qui peut être utilisée avec des classificateurs (complexes) déjà existants.

Inconvénients du Self-Training

- Des erreurs initiales « s'auto-renforcent » toutes seules.
  - Solutions heuristiques, par exemple enlever l'étiquette d'une instance si sa confiance descend au dessous d'un seuil.

- On peut rien dire sur la convergence.
  - Mais il y a des cas particuliers où l'algorithme de self-training est équivalent à l'algorithme d'Expectation-Maximization (EM).

  - Il y a aussi des cas particuliers (par exemple., fonctions linéaires) où la solution est connue.

Graph-Based Algoritms

Hypothèse

Un graphe est donné sur toutes les données.
Les instances reliées avec des segments qui ont un poids important auront tendance à avoir la même étiquette.

- Création du graphe à partir des données (Matrice de distance, Knn puis Floyd Algorithm pour la grande dimension…)

- On estime f sur le graphe qui vérifie:
  - f proche des étiquettes données y_l sur les nœuds labellisés
  - f doit être lisse sur l'ensemble du graphe.

Avantages Graph-Based Algoritms

- Base mathématique claire
- Bonne Performance quand le graphe est bien adéquat au cas qu'on veut traiter.

Inconvénients Graph-Based Algoritms

- Mauvaise performance si le graphe est mauvais.
- Sensible à la structure du graphe et aux poids des segments.

Machine à Vecteurs de Support pour l'apprentissage semi-supervisé

- Semi-supervised SVMs (S3VMs) = Transductive SVMs (TSVMs)
- Maximiser “la marge” des données non étiquetées.

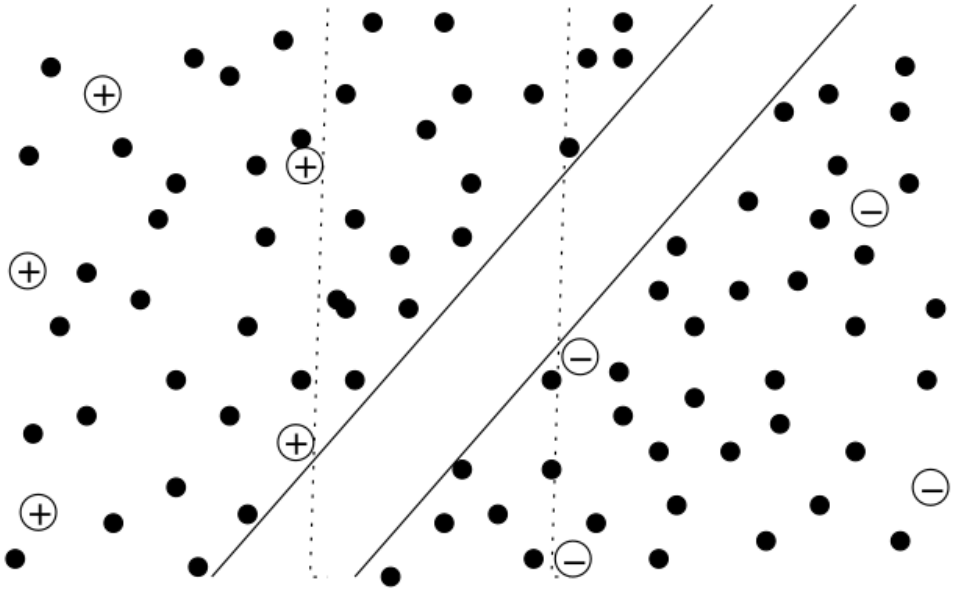


Image : “Semi-Supervised Learning” Olivier Chapelle,Bernhard Schölkopf, and Alexander Zien

S3VM

Hypothèse

Les données non-étiquetées des différentes classes sont séparées avec une grande marge.

L'idée derrière S3VM:

- Énumérer toutes les 2^u possibilités d'étiquetage possible de $X_{u.}$
- Construire une SVM standard pour chaque possibilité.
- Choisir la SVM avec la plus grande marge.

Avantages S3VM

- Base mathématique claire
- Applicable quand SVMs est applicable

Inconvénients S3VM

- Optimisation difficile
- Peut être coincer dans un minima locale
- potentiellement moins de gains, car hypothèse plus modeste que les méthodes génératives ou celles basées sur les graphes

Cas Pratique

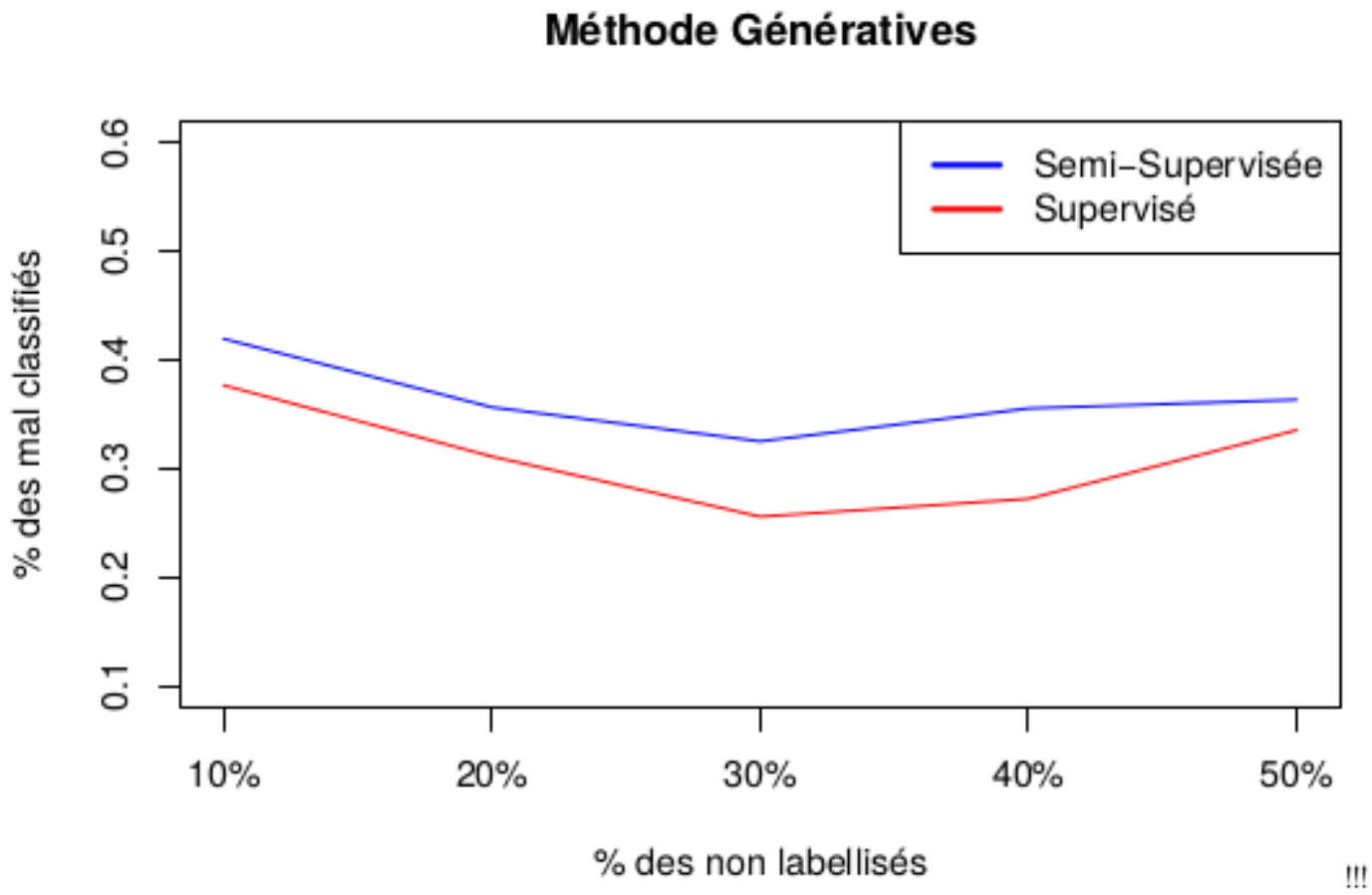
- Méthodes génératives
  -Package “upclass” (language R)
- Self-Training
  - Package “DmwR” (langage R)
- Graph-based methods
  - Package “spa” (langage R)
- S3VM
  - “svmlin” (librairie écrite en C++)

Homesite Dataset

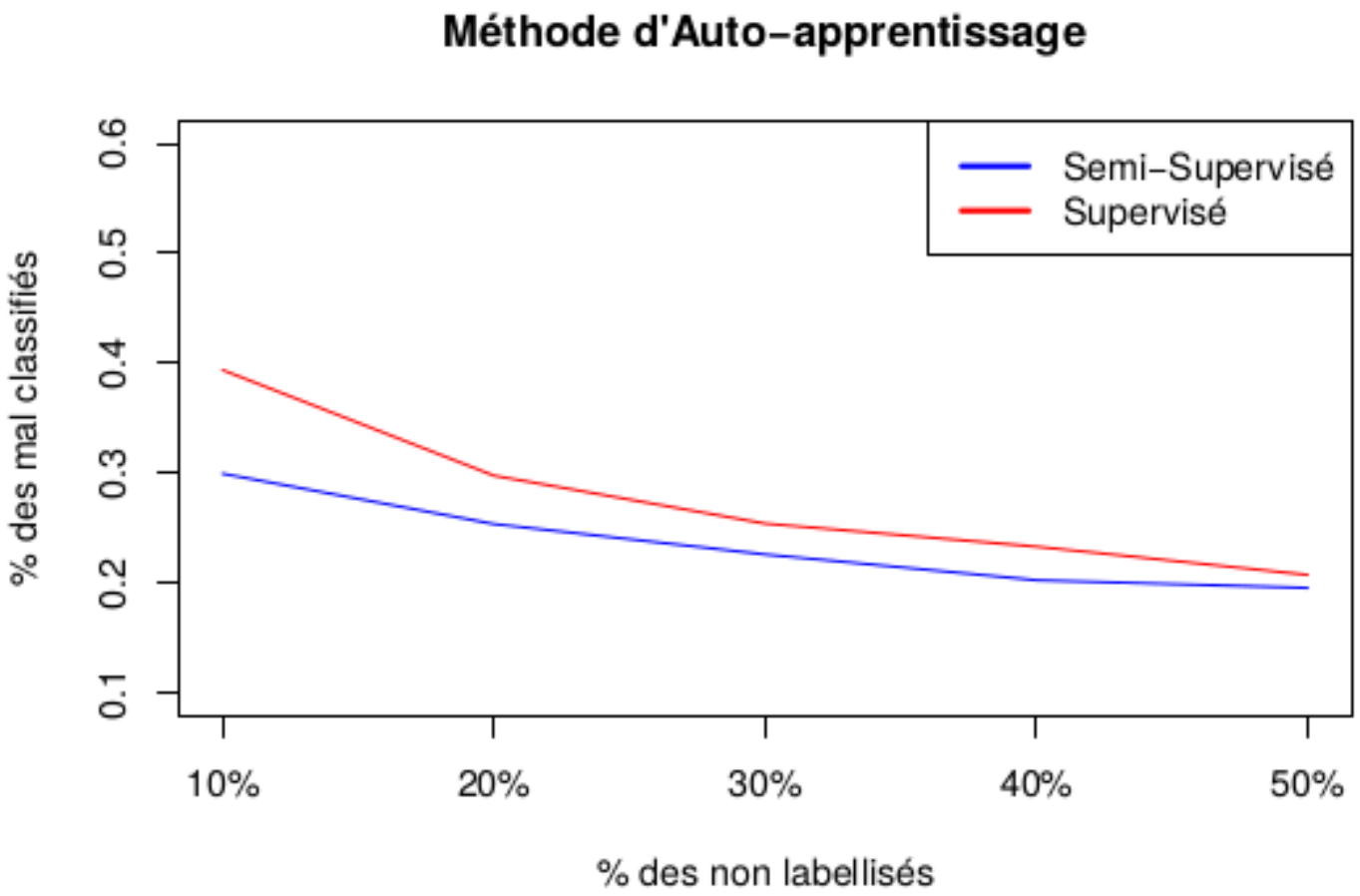
- L'ensemble de donnée contient quelques 300 000 lignes et 290 variables
- Pour la fonction binaire objectif il y a ~18% dans la première classe (client qui ont acheté l'assurance) et ~82% dans la deuxième classe (qui n'ont pas acheté)
- Pour les variables elles peuvent être regroupées en 5 grandes catégories : données temporelles, données personnelles du client, données de ventes, données géographiques du client, données sur l'habitation du client. A
- À noter que 9% des variables sont qualitatives.

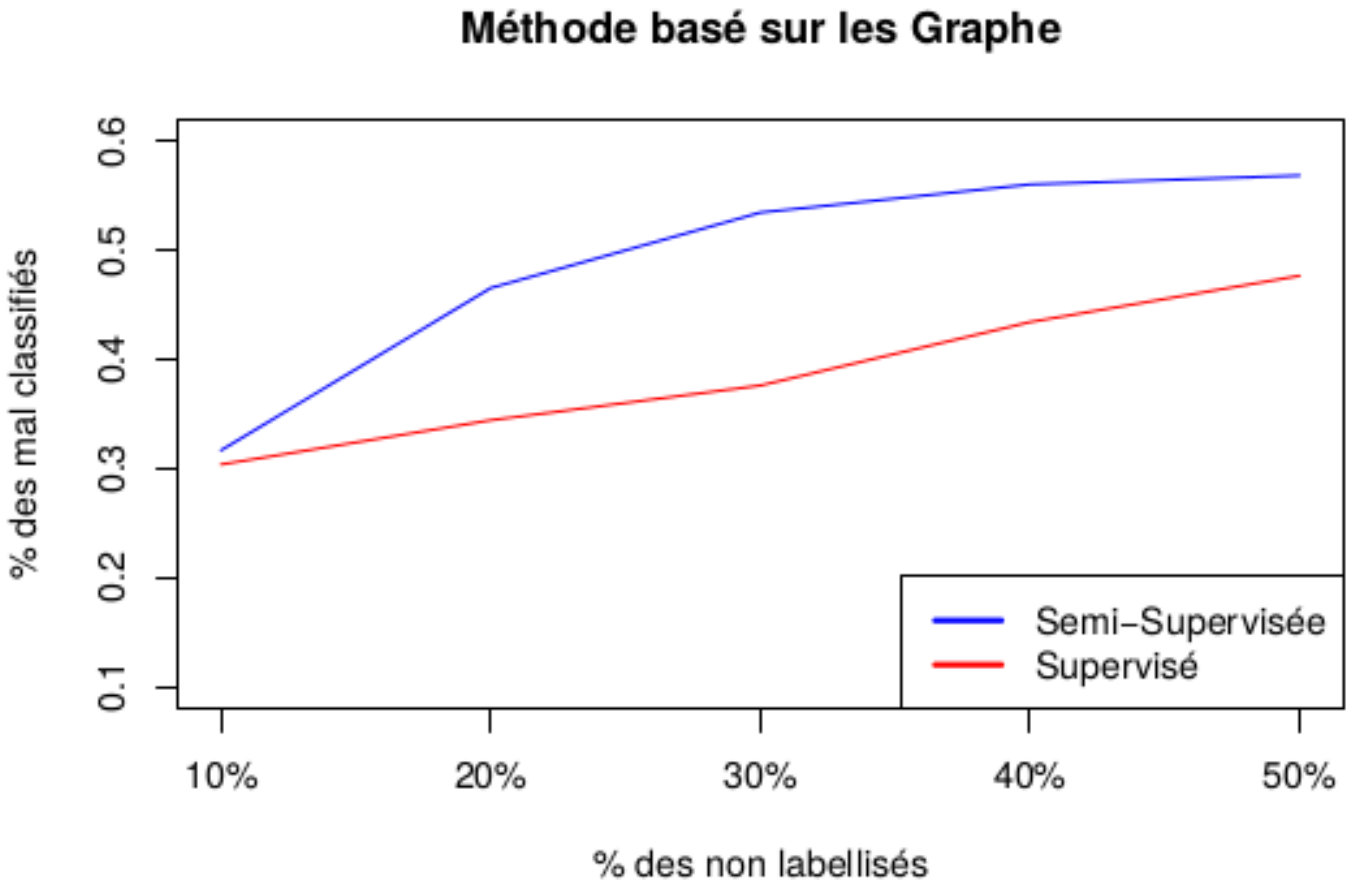
Expérience:

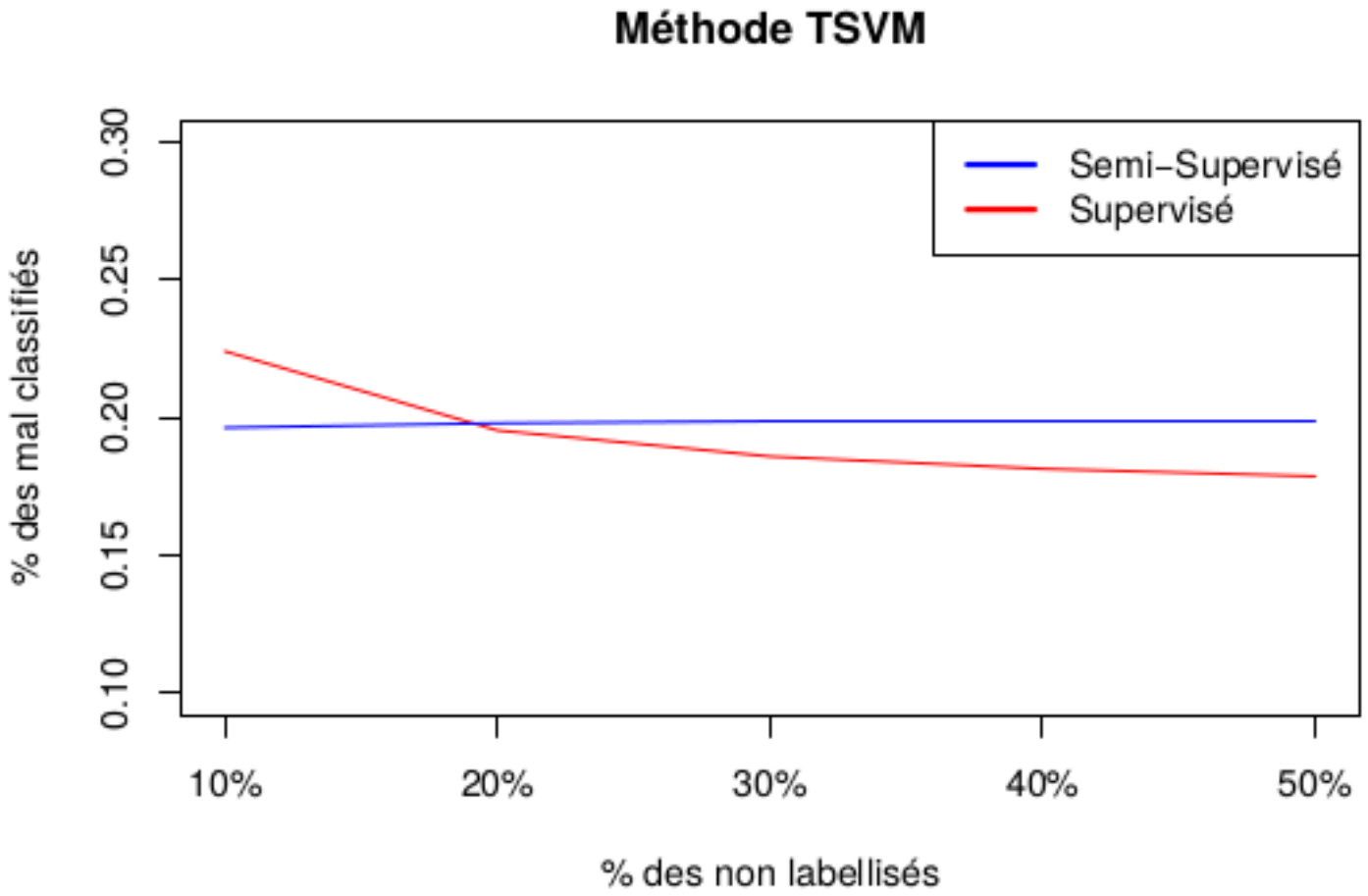
- Échantillonnage aléatoire 10% (labellisées)
- Apprentissage S sur 10% .
- Apprentissage SS sur les 10% labellisées + 90% non labellisées.
- Prédiction des labels des 90% avec les deux modèles.
- Les étapes au dessus sont répétées 40 fois avec différents échantillons de 10% - 90%.
- Pour chaque méthode on répète ces étapes pour 20%, 30%, 40% et 50% de données labellisées


Méthode Génératives
Semi-Supervisée
Supervisé
% des mal classifiés
0.1
0.2
0.3
0.4
0.5
0.6
10%
20%
30%
40%
50%
% des non labellisés

!!!


Méthode d'Auto-apprentissage
Semi-Supervisé
Supervisé
% des mal classifiés
0.1
0.2
0.3
0.4
0.5
0.6
10%
20%
30%
40%
50%
% des non labellisés


Méthode basé sur les Graphe
% des mal classifiés
0.1
0.2
0.3
0.4
0.5
0.6
10%
20%
30%
40%
50%
% des non labellisés
Semi-Supervisée
Supervisé


Méthode TSVM
Semi-Supervisé
Supervisé
% des mal classifiés
0.10
0.15
0.20
0.25
0.30
10%
20%
30%
40%
50%
% des non labellisés

Choix es techniques?

Algorithme	Précision	dimension	Temps	Notes
Generative	*	**	**	Précision dépend beaucoup de l'hypothèse du *clustering*
Self-Train	***	**	*	La précision dépend beaucoup de l'efficacité du classificateur utilisé. Faire attention à l'auto-renforcement de l'erreur.
Graph-based	*	*	*	Précision dépend beaucoup de l'hypothèse du *clustering* et de la qualité du graphe
TSVM	**	***	***	Adapté pour beaucoup de variables. La proportion des 2 classes si connue réduit les temps de calculs.

- Annotation coûteuse + abondance des données.
- Importance des hypothèses (données labellisées != gain automatique) :
  (« Fundamental Limitations of Semi-Supervised Learning » by Tyler (Tian) Lu)
- Limitations pratiques (grande dimension)

Autres techniques existantes

- Deep Neural Network pour l'apprentissage semi-supervisé.
- Co-Training.
- Semi-supervisé avec les règles d'association.

Ouverture

- Capacité cognitive humaine

Je vous remercie de votre attention.
Questions?