Apprentissage Semi-Supervisé

et Apprentissage Transductif

Arnaud Revel
revel.arnaud@gmail.com

Plan

Plan

Pourquoi s'intéresser à l'apprentissage semi-supervisé ?

Les données annotées sont chères

- C'est ennuyeux à faire...
- Cela nécessite l'avis d'un expert
- Il est parfois nécessaire d'utiliser des dispositifs spécifiques

Idée de l'apprentissage semi-supervisé

Utiliser les données non-annotées pour compléter l'apprentissage supervisé

Apprentissage Semi-supervisé vs Transductif

L'apprentissage semi-supervisé est souvent associé au concept d'apprentissage **transductif**.

Apprentissage transductif

L'apprentissage s'effectue sur les données de la base d'apprentissage dans le but de faire des prédictions sur les observations de la base de test, et uniquement celles ci.
Le but n'est donc pas de déterminer la fonction qui minimise l'erreur en généralisation, mais celle qui minimise l'erreur moyenne sur la base de test.

Discussion Scholkopf, Vapnik

La distinction entre apprentissage transductif et semi-supervisé n'est pas si tranchée...

Problématique

Comment apprendre à partir de données non-annotées ?

L'idée de l'approche semi-supervisée est d'adapter le modèle à la structure du problème.
S'il existe des degrés de liberté sur les paramètres du modèles il s'agit de trouver le modèle adapté à la fois aux données d'apprentissage et aux données de test.

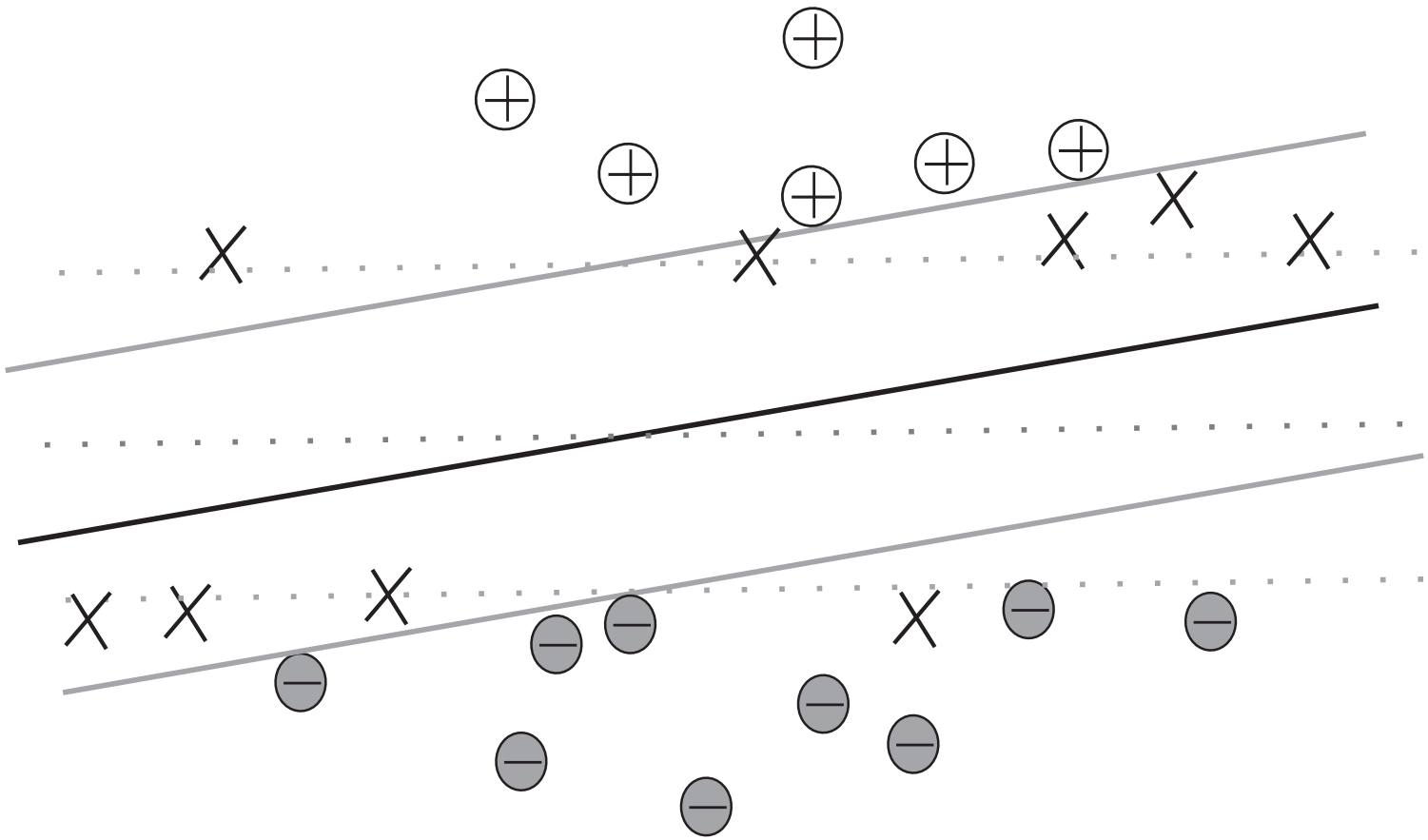
L'utilisation de données non-annotées est-elle toujours utilisable ?

Si le modèle est finalement inadapté par rapport aux structures des données, les performances du classifieur peuvent être altérées.

Comment utiliser les données non-annotées

Le principe de l'apprentissage semi-supervisé est soit, de modifier, soit de réorganiser les hypothèses effectuées sur le modèle à partir des données d'apprentissage.

Problématique



Plan

Auto-Apprentissage

1. L'auto-apprentissage (self-training) [Zhu05] consiste à entraîner un classifieur avec les données étiquetées (DL).
2. Le classifieur est, ensuite, utilisé pour étiqueter les données incomplètes (DU).
3. Les données étiquetées avec un haut degré de confiance sont ajoutées aux données d'apprentissage (DL).
4. Le classifieur est ré-entraîné sur les données de DL et la procédure est répétée jusqu'à satisfaire un critère d'arrêt.

Co-apprentissage

Co-apprentissage

L'idée du co-apprentissage est que s'il existe 2 projections indépendantes d'un même espace de données, deux classifieurs entraînés selon ces 2 projections, doivent étiqueter de manière identique la même donnée.

Algorithme

- L'ensemble d'attributs est divisé en 2 ensembles indépendants
- Deux classifieurs sont entraînés en utilisant ces jeux de paramètres sur les données d'apprentissage DL
- Ces classifieurs sont utiliser pour étiqueter les données de la base de test DU
- Les données étiquetées avec une bonne confiance sont ajoutées à DL
- La phase d'apprentissage des classifieurs est réitérée sur le nouvel ensemble d'apprentissage.
- Lorsque l'apprentissage est terminé, les deux classifieurs sont combinés.

Séparateur Semi-Supervisé à Vaste Marge (S^3VM)

- Dans cette approche, deux contraintes sont ajoutées au problème quadratique des SVM
- Ces contraintes sont définies pour maintenir les données non-étiquetées à l'extérieur de la marge tout en minimisant l'erreur de classification supposée :
  - $\frac{1}{2}||\mathbf{w}||^2 + C(\sum_{i=1}^{l} \xi_i + \sum_{j=l+1}^{N} \min(\xi_j^u, \xi_j^{u^*}))$
  - $y_i(\mathbf{w}^* \cdot \mathbf{x_i} + b) \geq 1 - \xi_i$
  - $(\mathbf{w}^* \cdot \mathbf{x_j} + b) \geq 1 - \xi_j^u$
  - $(\mathbf{w}^* \cdot \mathbf{x_j} + b) \leq -1 + \xi_j^{u^*}$
  - $\xi_j^u / \xi_j^{u^*}$ Coefficient de Lagrange lié à une erreur de classification dans le cas où l'échantillon est classé +1/-1

T-SVM

T-SVM

L'idée du T-SVM est d'induire une fonction globale à l'aide des données annotées et des données de test.
Le problème peut alors se poser de la manière suivante :

- $\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i + C^*\sum_{j=l+1}^{N}\xi_j^*$
- $y_i(\mathbf{w}^* \cdot \mathbf{x_i} + b) \geq 1 - \xi_i$
- $y_j^*(\mathbf{w}^* \cdot \mathbf{x_j} + b) \geq 1 - \xi_j^*$

Cette minimisation doit s'effectuer pour tous les cas de catégorisation possibles !

Approche itérative pour réduire la complexité

L'idée de [Joa99a, Joa02] est de partir d'une fonction locale et d'étiqueter les données (x_j, y_j^*) puis de généraliser petit à petit. Pour ce faire, le paramètre C^* est progressivement incrémenté :

- Dans le cas $C^* = 0$, l'apprentissage est purement inductif
- Dans le cas $C^* = 1$, l'apprentissage est complètement transductif.

A chaque itération, on cherche à minimiser la fonction quadratique puis on réapprend le SVM en augmentant C^*.

Algorithme

1. La fonction objectif est minimisée en cherchant un couple de données non étiquetées (x_m, x_l) se situant dans la marge (ou dans la zone où $\xi_m + \xi_l > 2$) et tels que $y_m^* \neq y_l^*$.
2. Les étiquettes sont alors permutées pour replacer les données dans une zone plus plausible, minimisant $(\xi_m + \xi_l)$.
   - Si un couple d'étiquettes a été permuté, un nouveau SVM est appris et la procédure est répétée.
   - Si, par contre, aucun couple n'a été trouvé, C^* est augmenté et l'algorithme passe à la prochaine itération.
3. Les itérations s'arrêtent lorsque C^* a atteint un seuil fixé a priori.

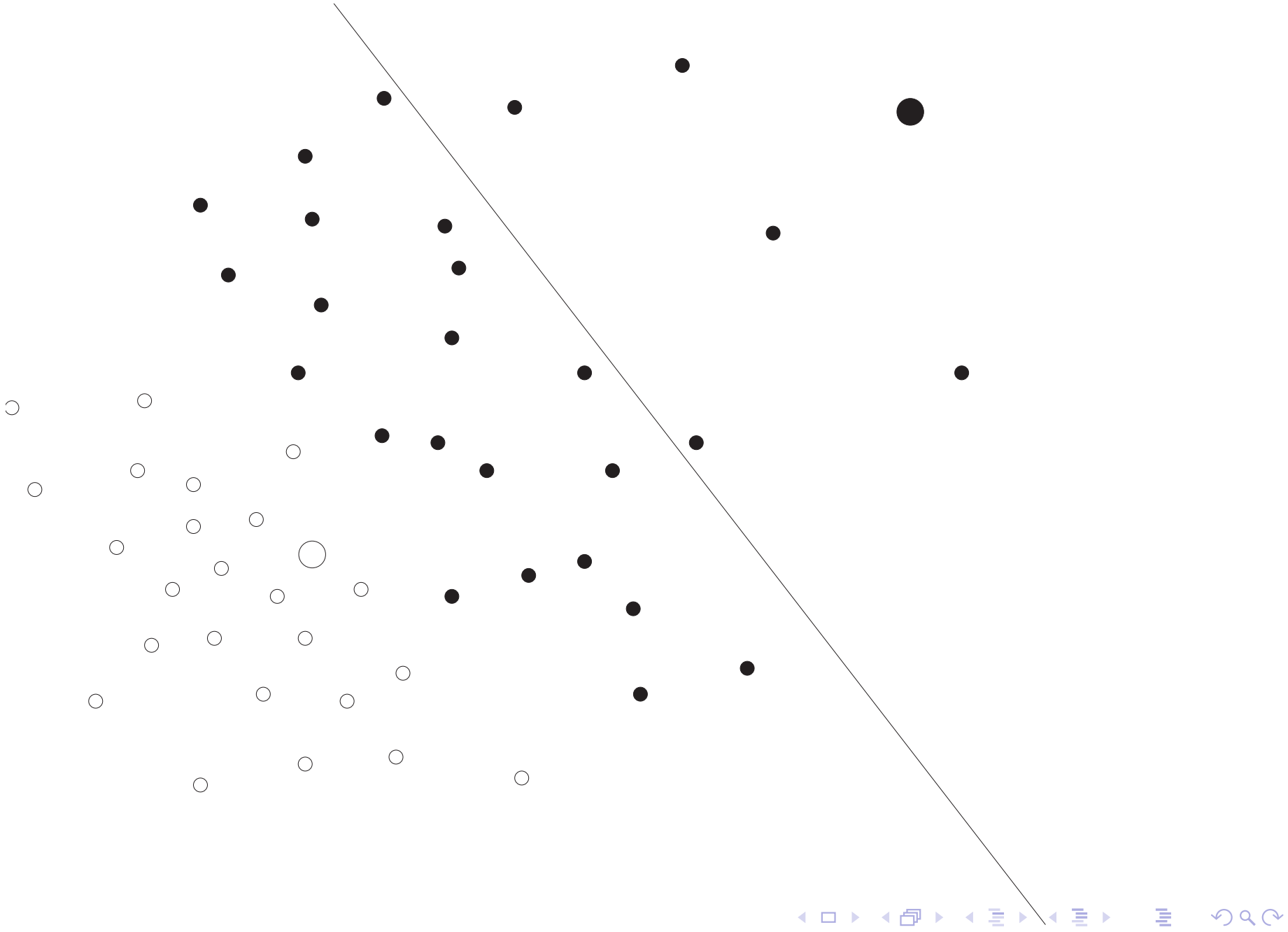
Plan

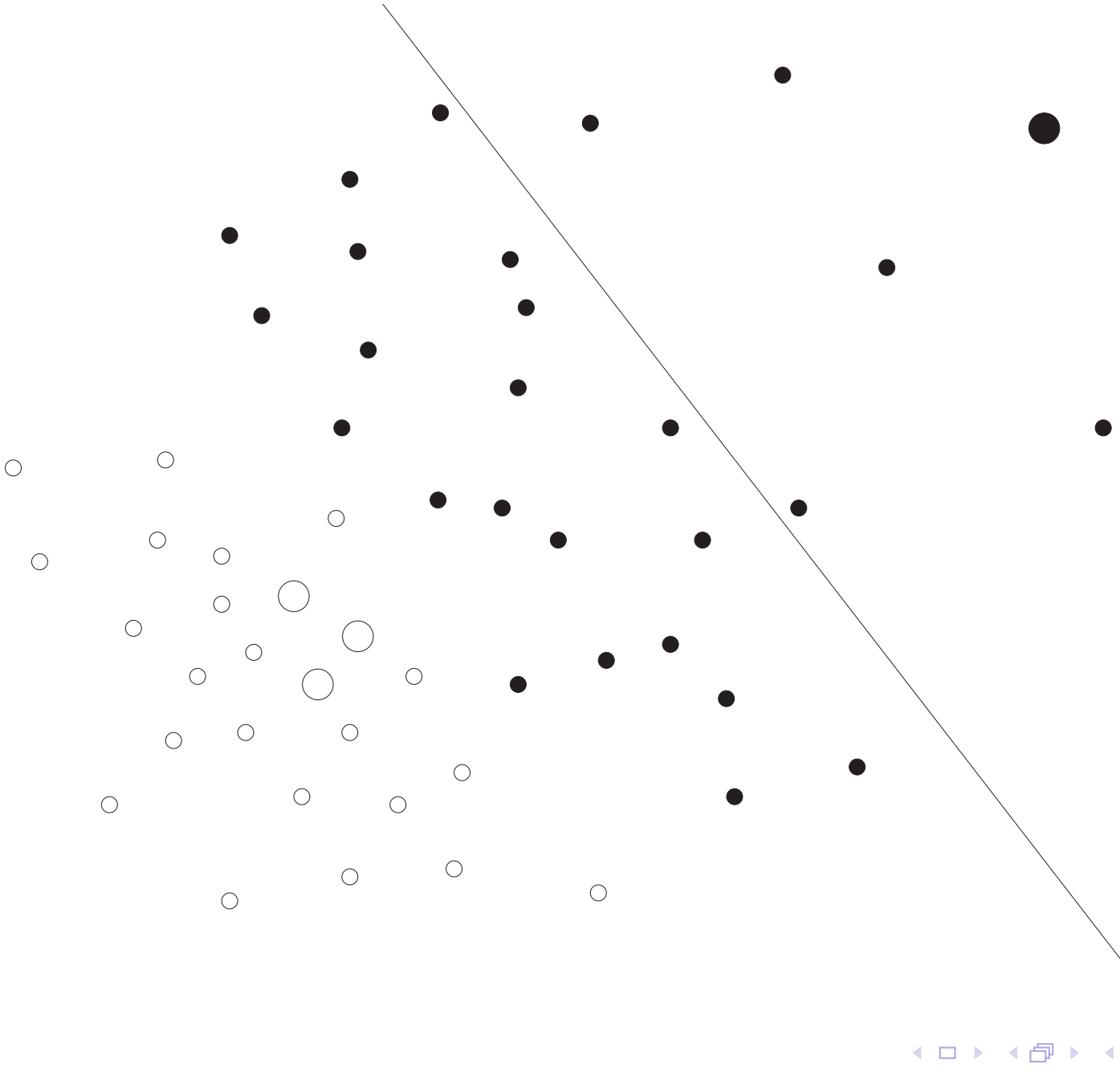
Apprentissage actif

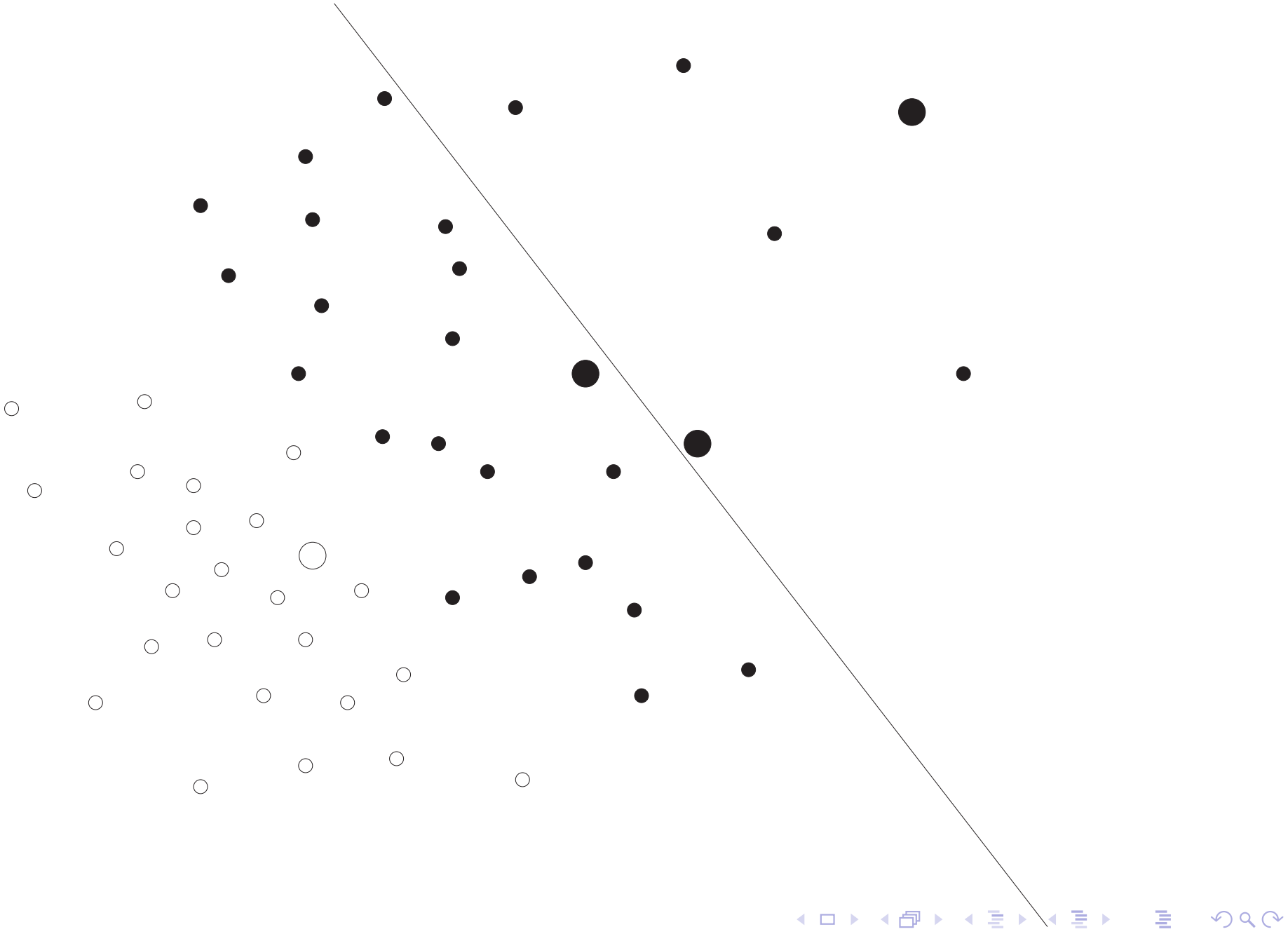
- L'apprentissage actif est une extension de l'apprentissage semi-supervisé
- Plutôt qu'exploiter les données non annotées...
- ...l'idée est de faire annoter de manière active les exemples qui apporteront le plus d'informations
- On minimise ainsi l'effort d'annotation

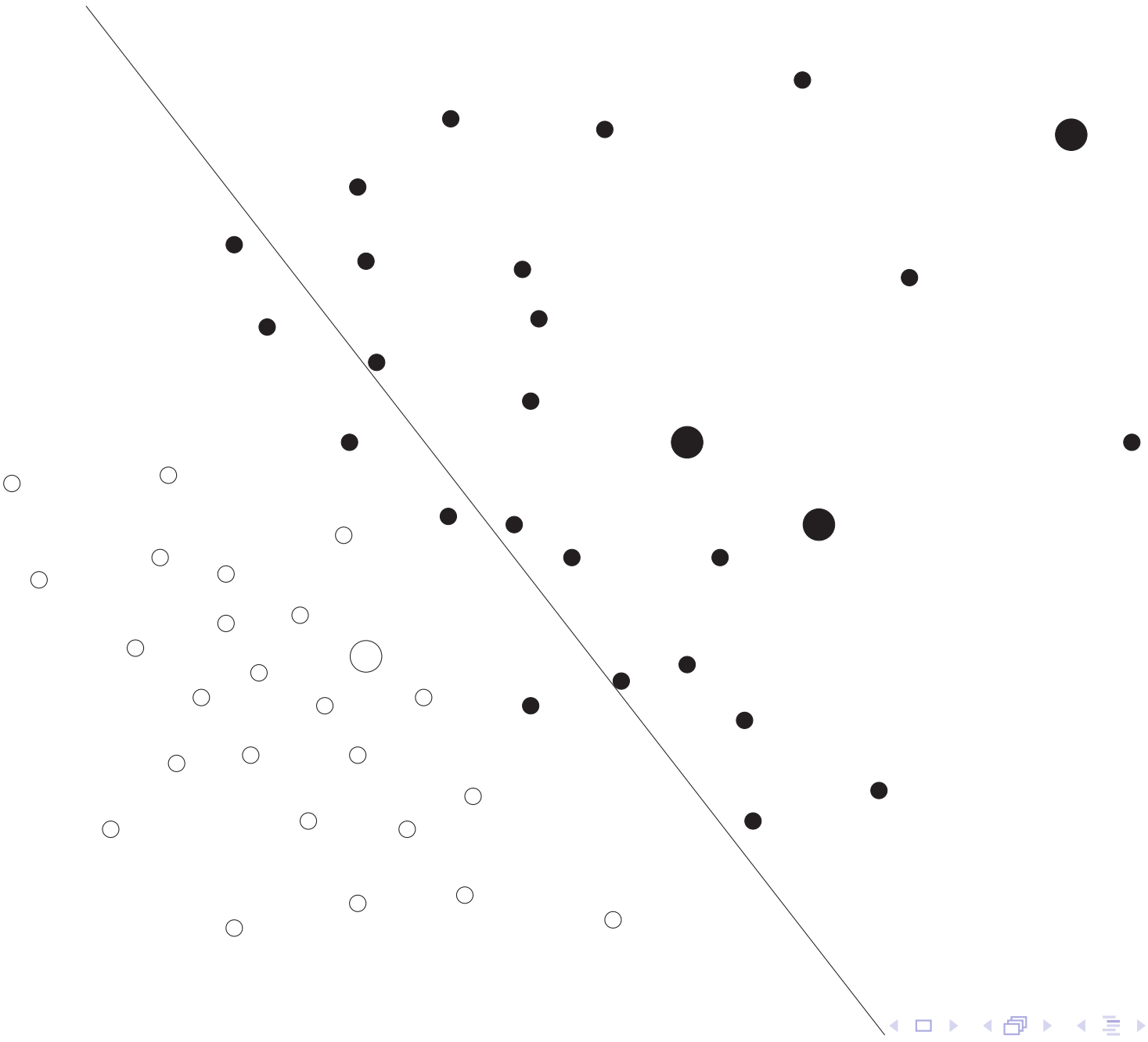
Supervisé/actif

Attention à la différence de l'apprentissage supervisé, les données apprises ne sont plus identiquement distribuées !









Méthodes actives basées sur l'incertitude

Méthode active basée sur l'incertitude

- Cette stratégie a pour objectif de sélectionner les documents à annoter parmi ceux dont le classifieur est le moins sûr.
- Si par exemple on dispose d'une mesure de probabilité de classement, on choisira les éléments dont la probabilité est proche de 0.5
- En pratique, on peut utiliser directement la sortie d'un SVM ou d'un k-ppv comme fonction de pertinence $f(x)$

Hélas, c'est dans la zone où la fonction de pertinence est proche de 0 où l'on est le classifieur est le moins pertinent !

Méthodes actives basées sur l'incertitude (suite)

Méthode active basée sur l'incertitude (suite)

Une autre possibilité consiste à utiliser plusieurs modèles et sélectionner les éléments pour lesquels les modèles se contredisent le plus.

Méthodes basées sur la réduction de l'erreur

Méthodes basées sur la réduction de l'erreur

- Cette stratégie vise à sélectionner les éléments qui, une fois ajoutés à la base d'apprentissage, minimisent l'erreur de généralisation.
- Soit $P(y|x)$ la probabilité que la donnée x soit de la classe y
- Soit $P(x)$ la distribution des images
- Soit $\hat{P}(y|x)$ l'estimation de $P(y|x)$ avec le classifieur courant
- L'erreur de généralisation est :

$$E_{\hat{P}} = \int_x L(P(y|x), \hat{P}(y|x)) dP(x) \quad (1)$$

- Avec $L()$ une fonction mesurant la perte entre $P(y|x)$ et $\hat{P}$
- La donnée sélectionnée est celle qui minimise $E_{\hat{P}}$ sur les données non encore annotées $\overline{I}$

Méthodes basées sur la réduction de l'erreur (suite)

Méthodes basées sur la réduction de l'erreur

- En pratique l'erreur de généralisation n'est pas praticable et $P(x)$ est approximée à partir des données non-annotées.
- $P(y|x)$ est estimée à partir de la pertinence prédite par le classifieur

$$\hat{E}_{\hat{P}} = \frac{1}{|\overline{I}|} \sum_{x_i} (1 - \max \hat{P}(y|x_i)) \tag{2}$$

- Par ailleurs, comme les annotations sont inconnues sur $\overline{I}$ elles sont estimées à partir de la prédiction sur chaque label

$$\hat{P}(y|x) = \frac{y}{2}(f(x) + y) \tag{3}$$

Plan

Bibliographie


ASEERVATHAM, S. (2007).
Apprentissage à base de Noyaux Sémantiques pour le Traitement de Données Textuelles.
PhD thesis, Universite Paris 13 – Institut Galilée.
http ://www-lipn.univ-paris13.fr/ aseervatham/pub/these_aseervatham.pdf.


Chapelle, O., Scholkopf, B., and Zien, A. (2006).
Semi-Supervised Learning.
MIT Press, Cambridge, MA.


Cord, M. and Cunningham, P., editors (2008).
Machine Learning Techniques for Multimedia.
Springer.


Zhu, X.
Semi-supervised learning literature survey.
http ://www.cs.wisc.edu/ jerryzhu/pub/ssl_survey.pdf.