# Data Mining:

# Concepts and Techniques

## (3rd ed.)

## — Chapter 9 —
## Classification: Advanced Methods

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University

# Semi-Supervised Learning

- Definition
  - We are given a set of labeled data as well as unlabeled data
  - We want to be able to predict the correct labels of those unlabeled data

# Semi-Supervised Learning

- Why unlabeled data?
  - Data label is often expensive to obtain
  - “Fraud or not?”, “Cancer or not?”, “credit rating”

# Semi-Supervised Learning

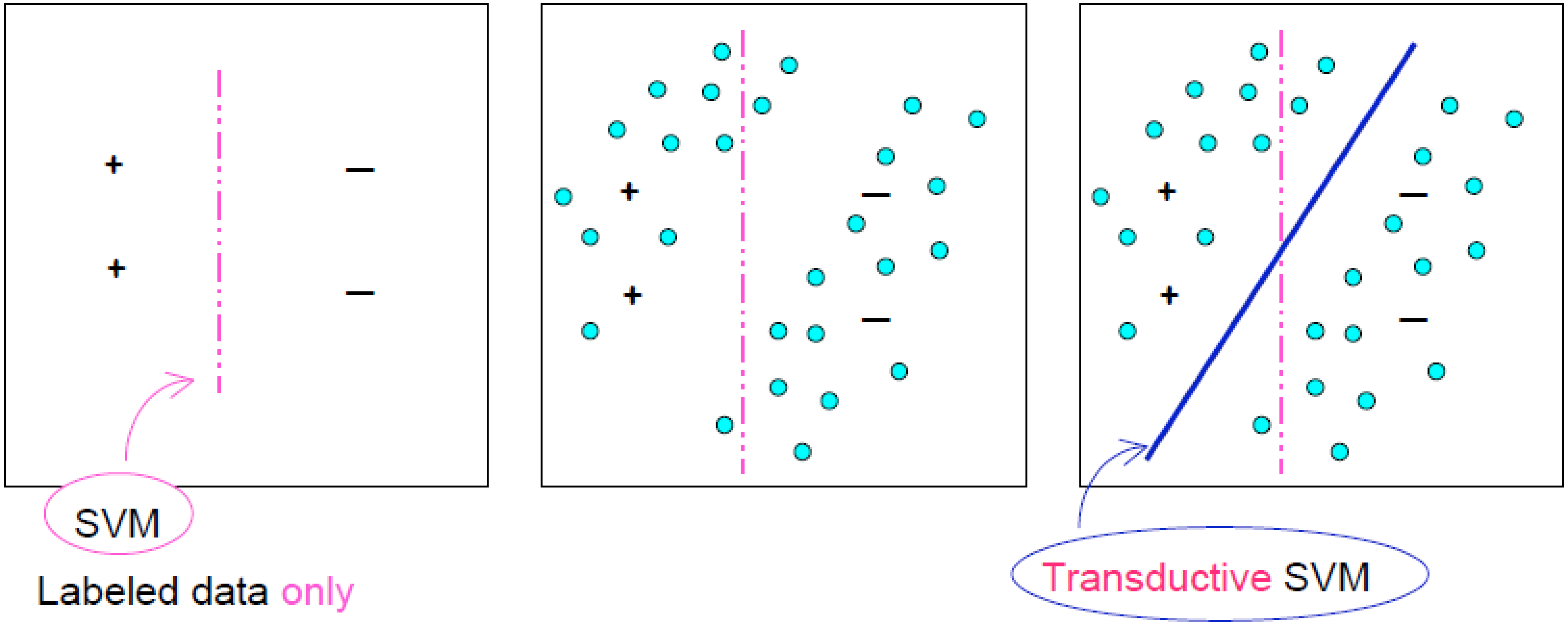
- An expectation
  - the classification performance better than
    - discarding the unlabeled data and doing supervised learning
    - discarding the labels and doing unsupervised

# Semi-Supervised Learning

- Transductive learning
  - is to infer the correct labels for the given unlabeled data  only.
- Inductive learning
  - is to infer the correct mapping from  x to y.

# Why Semi-Supervised Learning?

- Sparsity in data: training examples cannot cover the data space well
- unlabeled data can help to address sparsity



# Semi-Supervised Learning Methods

- Many methods exist:
  - self-training,
  - co-training,
  - EM with generative mixture models,
  - data-based methods,
  - transductive SVM,
  - graph-based methods, …

# Semi-Supervised Learning Methods

- Inductive methods and Transductive methods
  - Transductive methods: only label the available unlabeled data – not generating a classifier
  - Inductive methods: not only produce labels for unlabeled data, but also generate a classifier

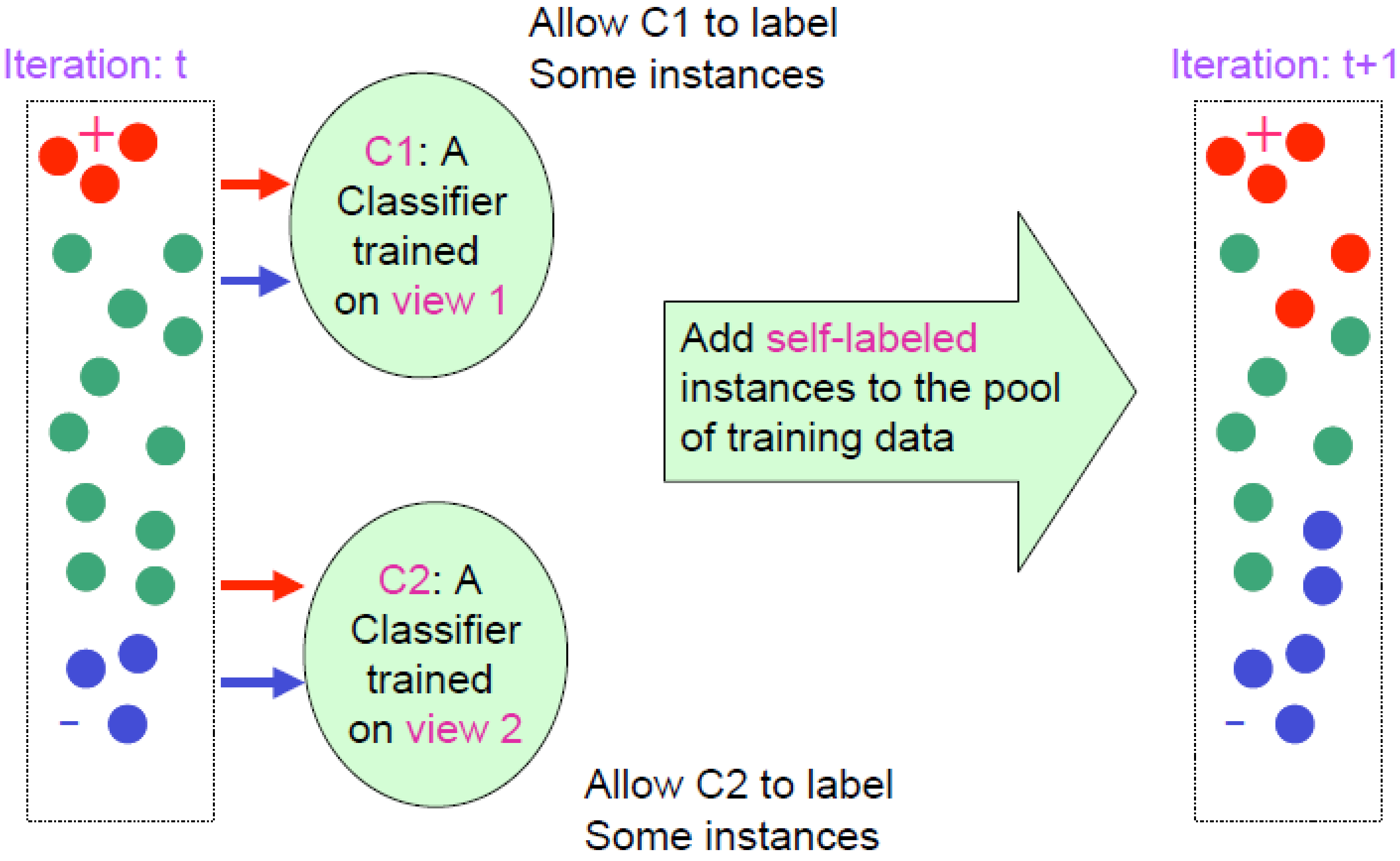
# Semi-Supervised Learning Methods

- Algorithmic methods
  - Classifier-based methods: start from an initial classifier, and iteratively enhance it
  - Data-based methods: find an inherent geometry in the data, and use the geometry to find a good classifier

# Self-training

- Build a classifier using the labeled data
- Use it to label the unlabeled data, and those with the most confident label prediction are added to the set of labeled data
- Repeat the above process
- Adv: easy to understand; disadv: may reinforce errors

# Co-Training



Allow C1 to label Some instances

Iteration: t

C1: A Classifier trained on view 1

Add self-labeled instances to the pool of training data

Iteration: t+1

C2: A Classifier trained on view 2

Allow C2 to label Some instances

# Co-training

- Each learner uses a mutually independent set of features of each tuple to train a good classifier, say $f_1$ and $f_{2.}$
- Then $f_1$ and $f_2$ are used to predict the class label for unlabeled data X
- Teach each other: The tuple having the most confident prediction from $f_1$ is added to the set of labeled data for $f_2$, & vice versa