SEMI-SUPERVISED LEARNING

Jasneet Sabharwal
jasneet.sabharwal@sfu.ca

Types of Learning

- Supervised Learning - Uses only labelled data for training a classifier.

- Semi-Supervised Learning - Uses both labelled and unlabelled data for training a classifier.

- Unsupervised Learning - Uses only unlabelled data.

Why do we need Semi-Supervised Learning?

- Labelled data is hard to get
  - Annotation takes time and is boring
  - Domain experts are required
  - Undergraduates are on a holiday
- Unlabelled data is cheap

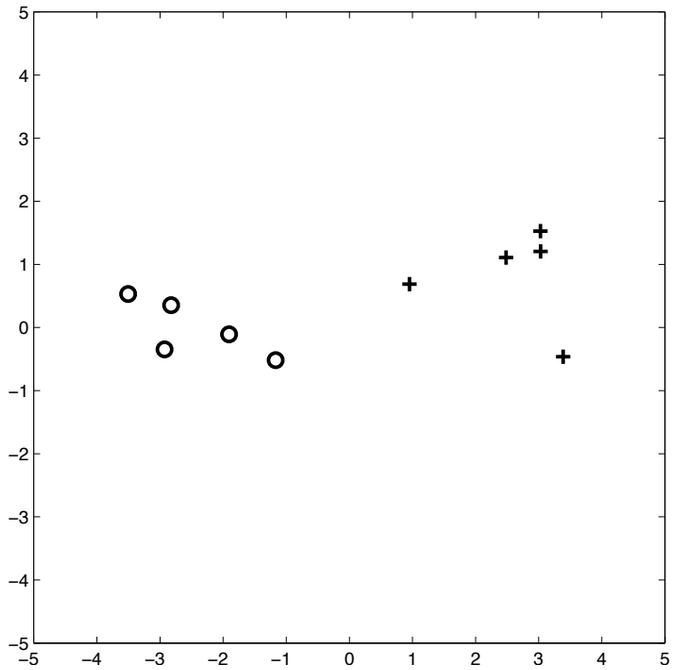
Generative Models

- $p(x,y) = p(y)\, p(x|y)$, where $p(x|y)$ is an identifiable mixture model (example - Gaussian Mixture Model)

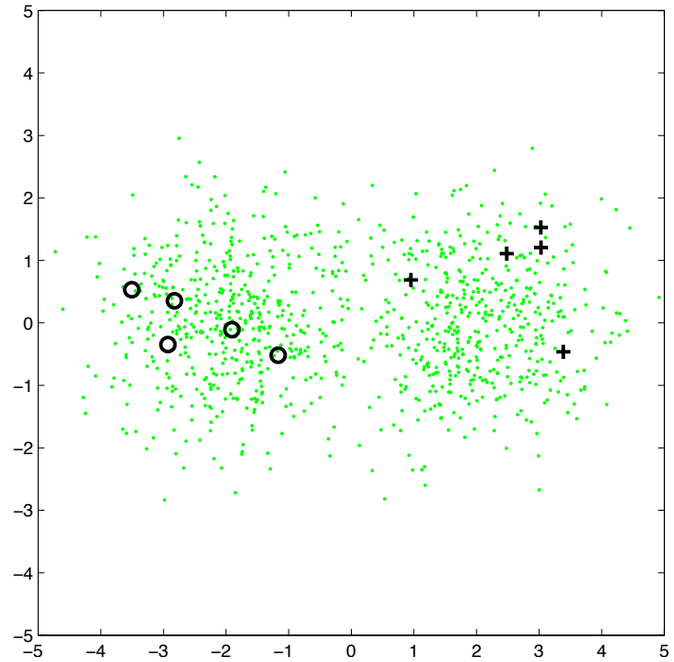
- With large unlabelled data, we can identify mixture components and then we

- Only require one labelled example per mixture component to identify the mixture distribution

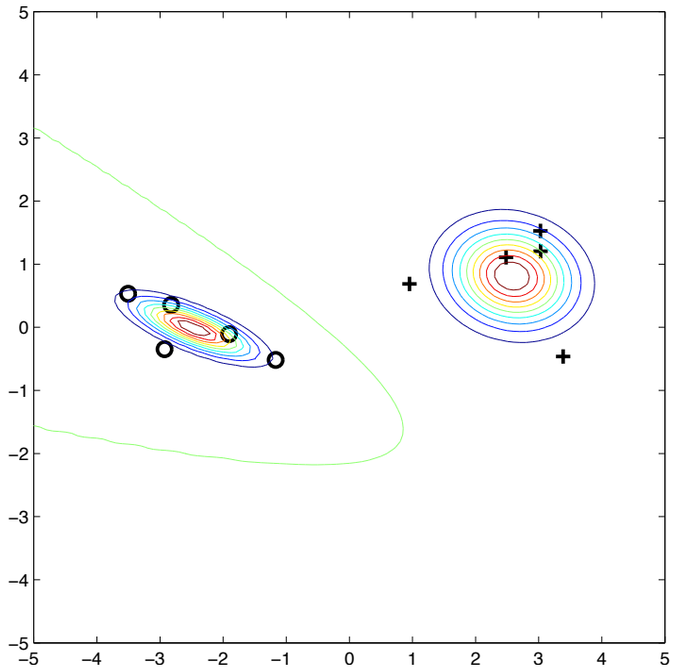
Generative Models



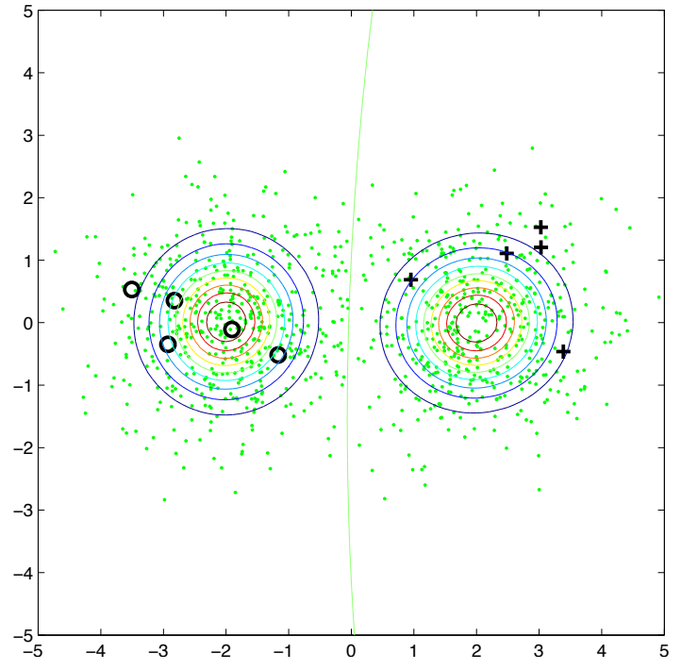
(a) labeled data



(b) labeled and unlabeled data (small dots)

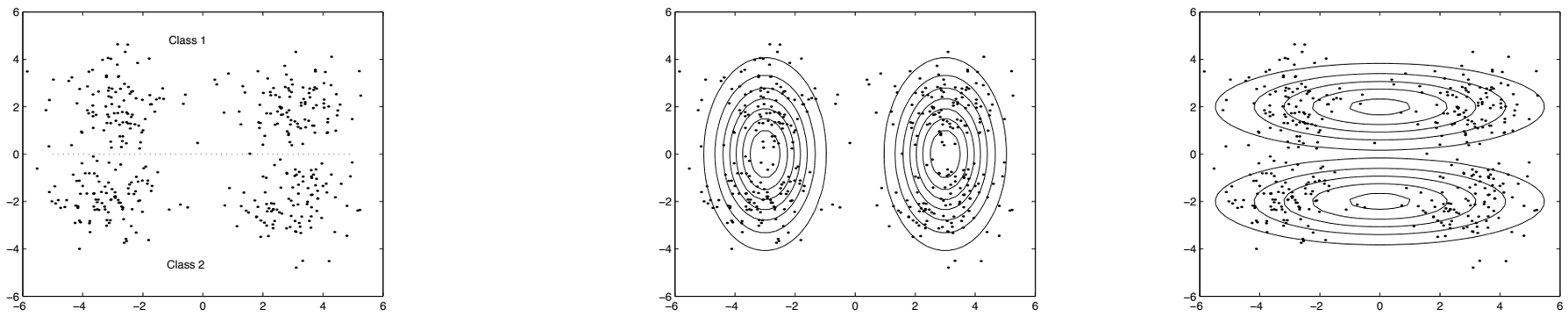


(c) model learned from labeled data



(d) model learned from labeled and unlabeled data

- Note
  - If mixture model assumptions are correct then unlabelled data will improve accuracy. - *Castelli & Cover, 1995; Castelli & Cover, 1996; Ratsaby & Venkatesh, 1995*
  - If mixture models are wrong then unlabelled data may hurt accuracy. - *Cozmen et. al. 2003*



(a) Horizontal class separation (b) High probability (c) Low probability

Figure 3: If the model is wrong, higher likelihood may lead to lower classification accuracy. For example, **(a)** is clearly not generated from two Gaussian. If we insist that each class is a single Gaussian, **(b)** will have higher probability than **(c)**. But **(b)** has around 50% accuracy, while **(c)**'s is much better.

Generative Models

- Identifying Mixture Components
  - Expectation Maximization (Dempster et. al., 1977) is generally used
    - Prone to local maxima
    - If local maxima far away from global maxima then unlabelled data may hurt learning

Fischer Kernel for Discriminative Learning

- Train generative mixture models (one component per class)

- Use EM to incorporate unlabelled data

- For each component model, convert labelled examples into fixed length Fisher score vector (derivatives of log likelihood w.r.t model parameters)

- Use Fisher score vectors in discriminative classifier (example - SVM)

Self-Training

- Train classifier on labelled data
- Use classifier to classify unlabelled data
- Add predicted unlabelled with high confidence to training set
- Retrain classifier

Self-Training

- Advantages of Self-Training
  - Simplest form of semi-supervised learning method
  - Wrapper method, applied to other existing classifiers
  - Frequently used in real time tasks in NLP (example - Named Entity Recognition)
- Disadvantages of Self-Training
  - Mistakes can re-enforce themselves

Co-Training

- Assumptions
  - Features can be split into two sets.
  - Each sub-feature set is sufficient to train a classifier.
  - Sub-feature sets are conditionally independent given the class.

Co-Training

- Train 2 classifiers using the sub-feature sets of the labelled data.
- Classify unlabelled data using each classifier.
- Each classifier trains the other classifier using the predicted labels of the unlabelled data for which the confidence is high.
- Each classifier's high confidant data points are IID samples to the other classifier.

Co-Training

- Works well when conditional independence holds.

- Only works when one classifier correctly labels a data that the other classifier misclassified.

- If no natural feature split is present, then randomly split features into two parts.

Co-Training

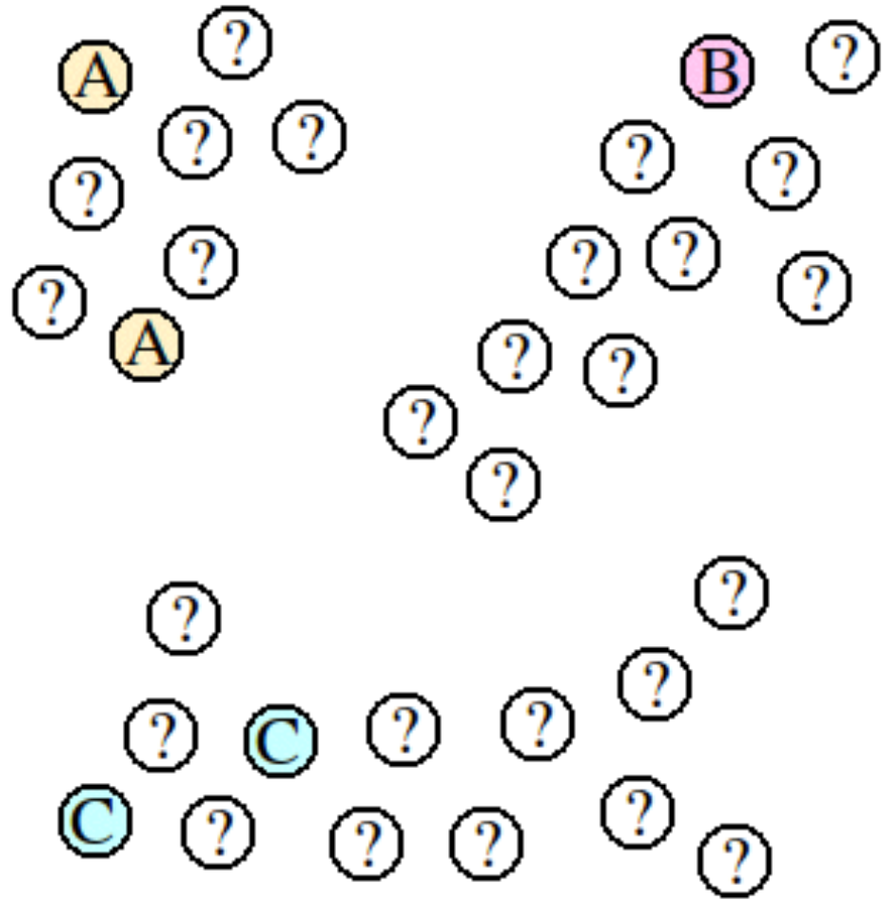
- Alternative to feature splitting
  - *Goldman and Zhou (2000)* - Use two different learners but use full features for both. Then use one learner's high confidence data points as training for other learner.
  - *Zhou and Goldman (2004)* - Ensemble of learners with different inductive bias.
  - *Zhou and Li (2005)* - Tri-learning

Semi-Supervised vs Transductive Learning

- Semi-Supervised Learning
  - Uses both labelled and unlabelled data
  - Contrasts Supervised or Unsupervised learning
- Transductive Learning
  - Only works on labelled and unlabelled data
  - Cannot handle unseen data

Transductive Learning

- Inductive approach - Use labelled data and train a supervised classifier.
  - Only 5 labelled points
  - Tough to capture the structure of data
- Transductive approach - Consider all points
  - Label the unlabelled according to clusters that they belong to
  - If unseen data points are added, then entire transductive algorithm needs to be repeated



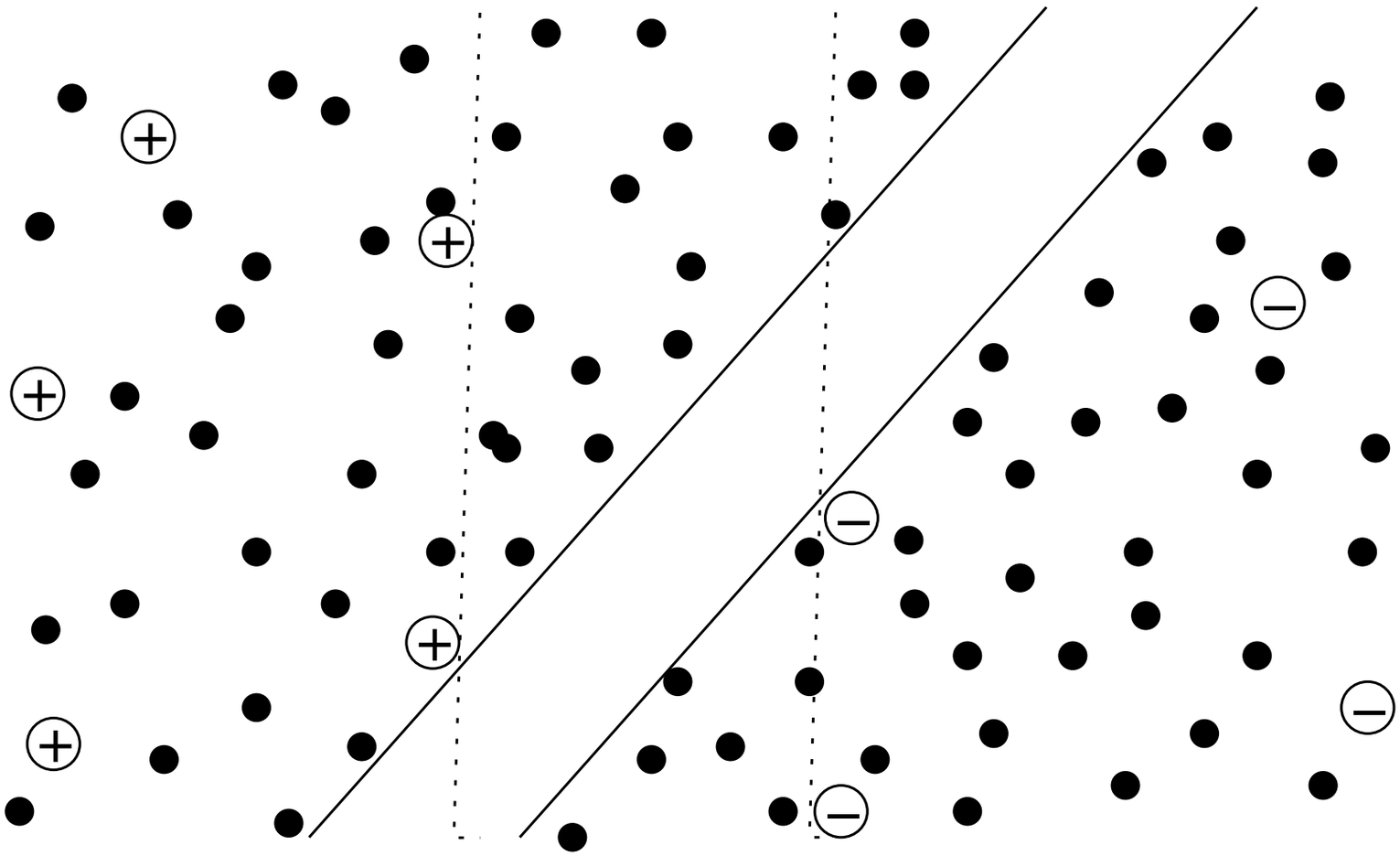
Transductive SVMs

- SVMs only use labelled data and we find maximum margin boundary

- Use of both labelled and unlabelled data

- Goal - Find labelling of unlabelled data such that linear boundary has maximum margin on both labelled and unlabelled data

Transductive SVMs



Unlabelled data guides the linear boundary
away from dense regions

Thank You