GbR'2011 - Mini Tutorial: Graph Transduction

Aykut Erdem[1] and Marcello Pelillo[2]

[1] *Hacettepe University, Ankara, Turkey*
[2] *University of Venice, Venice, Italy*

May 20, 2001

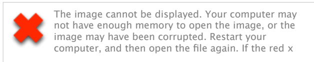
Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Semi-supervised learning (SSL)

- Unsupervised learning
  - learning with unlabeled data 

  - recovering the structure of the data

- Supervised learning
  - learning with labeled data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$
  - finding a mapping from a input space to an output space $f : \mathcal{X} \rightarrow \mathcal{Y}$

- Semi-supervised learning
  - learning with labeled and unlabeled data

    labeled data: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ *– hard to obtain*

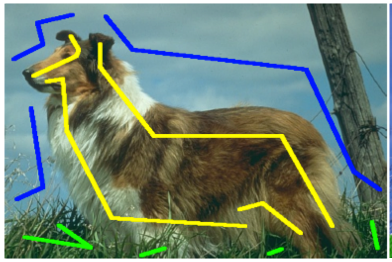
    unlabeled data: $\{\mathbf{x}_{\ell+1}, \ldots, \mathbf{x}_n\}$ *– is vast and cheap to obtain*

  - Can we find a better classifier  from both labeled and unlabeled data?

Inductive vs. Transductive SSL

labeled data:	$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$	$\mathcal{X}_\ell = \{\mathbf{x}_i\}_{i=1}^{\ell}$
unlabeled data:	$\{\mathbf{x}_{\ell+1}, \ldots, \mathbf{x}_n\}$	$\mathcal{X}_u = \{\mathbf{x}_j\}_{j=\ell+1}^{n}$
test data:	$\{\mathbf{x}_{n+1}, \ldots\}$	*available after training*

- Inductive SSL:
  - finding $f : \mathcal{X} \to \mathcal{Y}$
  - classification of newly observed data
- Transductive SSL:
  - finding $f : \mathcal{X}_\ell \cup \mathcal{X}_u \to \mathcal{Y}^{|\mathcal{X}_\ell \cup \mathcal{X}_u|}$
  - classification of unlabeled data
  - *e.g.* interactive segmentation



Original image with some labels

Segmentation result

How does SSL work?

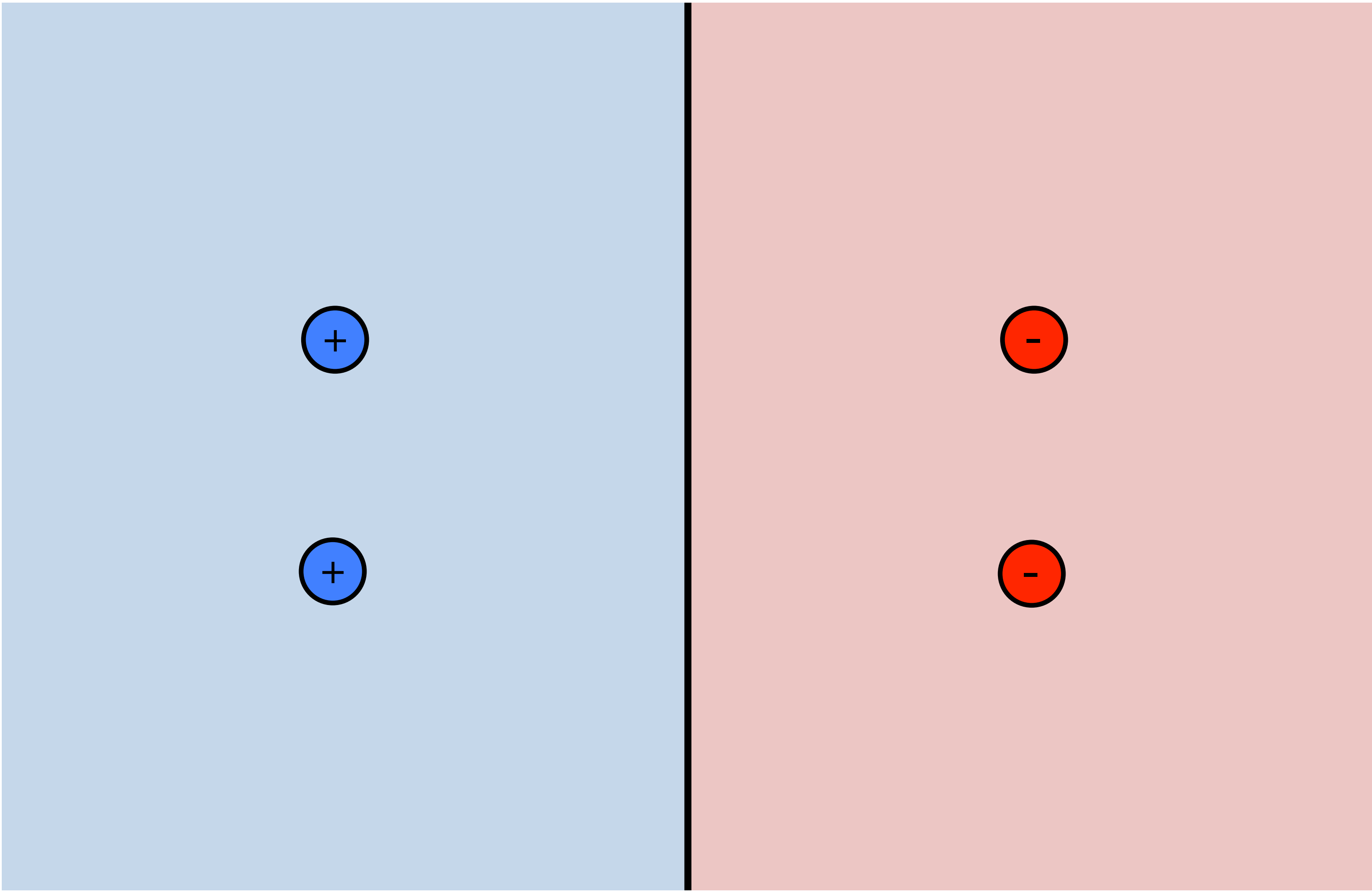
- SSL assumes that the structure of the data $p(x)$ is related to the conditional probability $p(y|x)$.

- Smoothness Assumption:
  - Two nearly points are likely to have the same output.

- Cluster assumption:
  1. The data form distinct clusters.
  2. Two points in the same cluster are expected to be of the same class.

- Manifold assumption:
  1. The data lie (roughly) on a low-dimensional manifold.
  2. Points on the same submanifold are likely have the same output.

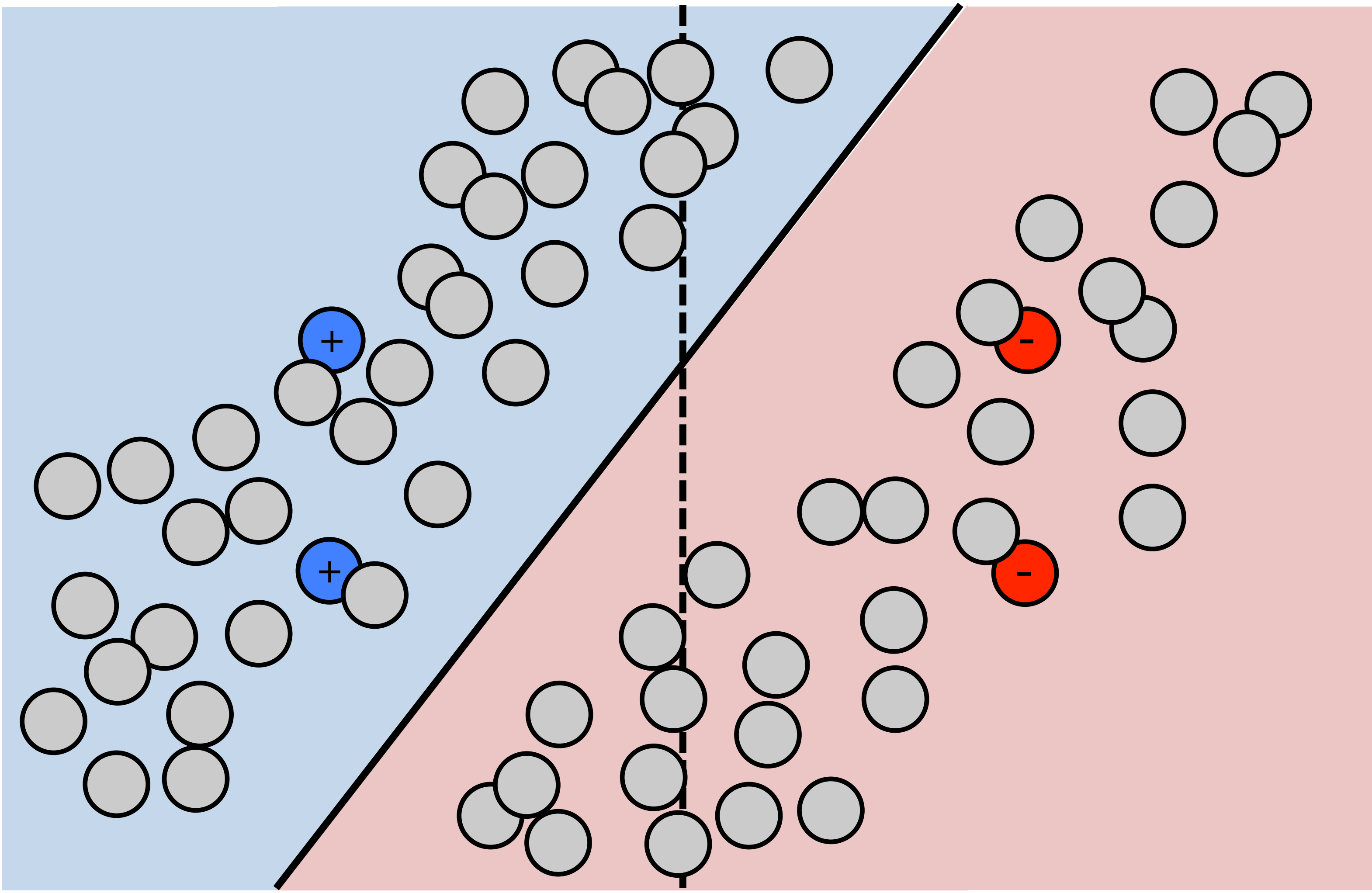
Supervised learning



Max. margin

adapted from material by H. Bischof

Semi-supervised learning



Low density around decision boundary

adapted from material by H. Bischof

Semi-supervised learning approaches

- Self-training
- Generative Models
- Semi-supervised SVMs
- Graph-based approaches

- Zhu, X., Semi-supervised learning literature survey. *TR* 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html
- Chapelle, O., Zien, A., & Schölkopf, B. (Eds.), *Semi-supervised learning*. MIT Press, 2006. http://olivier.chapelle.cc/ssl-book/

Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Graph Transduction

- Given a set of data points grouped into
  - labeled data: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$
  - unlabeled data: $\{\mathbf{x}_{\ell+1}, \ldots, \mathbf{x}_n\}$

  $\ell \ll n$
- Express data as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
  - $\mathcal{V}$: nodes representing labeled and unlabeled points
  - $\mathcal{E}$: pairwise edges between nodes weighted by the similarity between the corresponding pairs of points
- Assumption: Nodes connected with edges of high weights tend to have the same label.
- Idea: Propagate the information available at the labeled nodes to unlabeled ones in a consistent way
  - A regularized function estimation problem

Constructing input graphs

- The input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ characterizes data geometry.
  - $\mathcal{V}$: labeled and unlabeled points
  - $\mathcal{E}$: pairwise similarities

    Gaussian kernel is the most straightforward choice
    $w_{ij} = \exp(-dist(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma^2)$
- A crucial step before deciding which algorithm to use
- Common types of graphs:
  - Fully connected graph
  - *k* NN : $w_{ij} = 0$ if $\mathbf{x}_j \notin k\mathrm{NN}(\mathbf{x}_i)$
  - ϵ NN : $w_{ij} = 0$ if $dist(\mathbf{x}_i, \mathbf{x}_j) > \epsilon$
- More complicated methods:
  - Wang, F. and Zhang, C., Label propagation through linear neighborhoods. In *ICML* 2006.
  - Jebara, T., *et al.*, Graph Construction and b-Matching for Semi-Supervised Learning. In *ICML* 2009.

Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Basic graph-based algorithms

- Mincut (*st*-cut) (Blum and Chawla, 2001)
- Gaussian fields and Harmonic Functions (Zhu *et al.*, 2003)
- Local and global consistency (Zhou *et al.*, 2004)
- Manifold regularization (Belkin *et al.*, 2005)

• All these methods solves a different regularized function estimation problem.

• The classification function f should simultaneously satisfy two conditions:
  ① f should be close to the given labels y_l on the labeled nodes,
  ② f should be smooth on the whole graph.

Mincut (*st*-cut)

- Binary labels $f(\mathbf{x}_i) \in \{-1, 1\}$
- Positive labels act as sources, negative ones act as sinks.
  - Known labels are fixed, *i.e.* $f(\mathbf{x}_i) = y_i, \ \forall \mathbf{x}_i \in \mathcal{X}_\ell$
- Solves the problem:

$$\min_{f \in \{-1,1\}^n} \infty \sum_{i=1}^{\ell} (f(\mathbf{x}_i) - y_i)^2 + \sum_{i,j=1}^{n} w_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$$

- Performs hard classification – *no confidence estimation*
- Extensions to multi-label case is hard

Blum, A., Chawla, S., Learning from Labeled and Unlabeled Data using Graph Mincuts. In *ICML* 2001.

Gaussian Fields and Harmonic Functions (GHFH)

- The classification function f is relaxed to have real values.
- Solves the problem:

$$\min_{f \in \mathbb{R}^n} \underline{\infty \sum_{i=1}^{\ell} (f(\mathbf{x}_i) - y_i)^2} + \underbrace{\frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f(\mathbf{x}_i) - f(\mathbf{x}_j))^2}_{f^T \Delta f}$$

where

- regularization term is based on graph Laplacian $\Delta = D - W$
- known labels are again clamped

- The harmonic property ensures f is smooth.

$$f(\mathbf{x}_i) = \frac{\sum_j w_{ij} f(\mathbf{x}_j)}{\sum_j w_{ij}} \quad \forall \mathbf{x}_i \in \mathcal{X}_u$$

Zhu, X., Ghahramani, Z. & Lafferty, J., Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML* 2003.

Gaussian Fields and Harmonic Functions (GHFH) (cont'd.)

- The harmonic property

$$f(\mathbf{x}_i) = \frac{\sum_j w_{ij} f(\mathbf{x}_j)}{\sum_j w_{ij}} \quad \forall \mathbf{x}_i \in \mathcal{X}_u \qquad \equiv \qquad f = Pf$$

where $P = D^{-1}W$

- Label propagation:

1. $f^{t+1} = Pf^t$
2. $f^{t+1}(\mathbf{x_i}) = y_i \;\; \forall \mathbf{x}_i \in \mathcal{X}_\ell$

until convergence

Zhu, X., Ghahramani, Z. & Lafferty, J., Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML* 2003.

Local and Global Consistency (LGC)

- Does not fix $f(\mathbf{x}_i) = y_i$ for labeled points
- Solves the problem:

$$\min_{f \in \mathbb{R}^n} \mu \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \left(\frac{f(\mathbf{x}_i)}{\sqrt{d_{ii}}} - \frac{f(\mathbf{x}_i)}{\sqrt{d_{jj}}} \right)^2$$

  - It employs a local fitness term which allows $f(\mathbf{x}_i)$ to be different from y_i for the labeled points $\mathbf{x}_i \in \mathcal{X}_\ell$
  - The regularization term is based on normalized Laplacian $L = D^{-1/2} \Delta D^{-1/2}$

- The parameter μ balances the trade-off between global smoothness and local fitness.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B., Learning with local and global consistency. In *NIPS* 2004.

Manifold Regularization

- An inductive SSL approach
- Also does not fix $f(\mathbf{x}_i) = y_i$ for labeled points
- Assumes the data really lies on a lower-dimensional manifold
  - A second (extrinsic) regularization term
- Solves the problem:

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{\ell} (f(\mathbf{x}_i) - y_i)^2 + \underbrace{\lambda_1 \|f\|_{\mathcal{H}}^2}_{\text{extrinsic smoothness}} + \underbrace{\lambda_2 f^T \Delta f}_{\text{intrinsic smoothness}}$$

Belkin, M., Niyogi, P., & Sindhwani, V., On Manifold Regularization. In *AISTATS* 2005

Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Some advanced topics

- SSL on large scale graphs
  - Tsang, I and Kwok, J., Large-scale sparsified manifold regularization. In *NIPS* 2006
  - Fergus, R., Weiss, Y., & Torralba, A., Semi-supervised Learning in Gigantic Image Collections, in *NIPS* 2009.
- SSL on directed graphs
  - Zhou, D., Huang, J., & Schölkopf, B., Learning from labeled and unlabeled data on a directed graph. In *ICML* 2005.
- SSL with dissimilarities
  - Goldberg, A., Zhu, X., & Wright, S., Dissimilarity in graph-based semi-supervised classification. In *AISTATS* 2007.
  - Tong, W., & Jin, R., Semi-supervised learning by mixed label propagation. In *AAAI* 2007.

Some advanced topics (cont'd.)

- SSL with hypergraphs
  - Agarwal, S., Branson, K. & Belongie, S., Higher order learning with graphs. In *ICML* 2006.
  - Zhou, D., Huang, J., & Schölkopf, Learning with Hypergraphs: Clustering, Classification, and Embedding, In *NIPS* 2007.
- Online SSL on graphs
  - Goldberg, A., Li, M., & Zhu, X., Online manifold regularization: A new learning setting and empirical study. In *ECML*, 2008
  - Valko, M., Kveton, B. Huang, L., & Ting, D., Online Semi-Supervised Learning on Quantized Graphs. In *Conference on Uncertainty in Artificial Intelligence*, 2010.
- Multiple Instance Learning
  - Zhou, Z., Xu, J-M., On the relation between multi-instance learning and semi-supervised learning, In *ICML* 2007

Outline

- What is semi-supervised learning?
- What is graph transduction?
- Basic algorithms
- Some advanced topics
- Conclusions

Conclusions

- Advantages
  - Intuitive formulations
  - Empirically provide good results

- Disadvantages
  - Performance is greatly sensitive to the input graph
  - Do not scale well

Main References

- Zhu, X., Semi-supervised learning literature survey. *TR* 1530, Computer Sciences, University of Wisconsin-Madison, 2005. http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html

- Chapelle, O., Zien, A., & Schölkopf, B. (Eds.), *Semi-supervised learning*. MIT Press, 2006. http://olivier.chapelle.cc/ssl-book/

- Zhu, X., Semi-Supervised Learning, ICML 2007 Tutorial. http://pages.cs.wisc.edu/~jerryzhu/icml07tutorial.html

Thanks for your attention..
Any questions?