Semi-supervised Learning

Lectured by Shangsong Liang
Sun Yat-sen University

Produced by Gholamreza Haffari
Simon Fraser University,
School of Computing Science

Outline

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for Structured Prediction
- Conclusion

Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

~~Semi-~~Supervised Learning

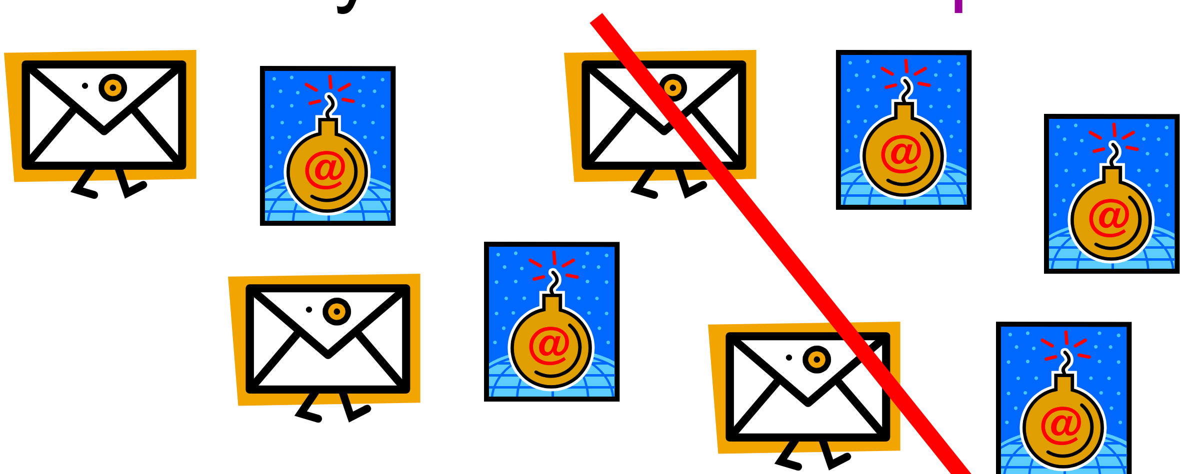
Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

- E.g, say you want to train an email classifier to distinguish spam from important messages

~~Semi-~~Supervised Learning

Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

- E.g, say you want to train an email classifier to distinguish spam from important messages
- Take sample S of data, labeled according to whether they were/weren't spam.



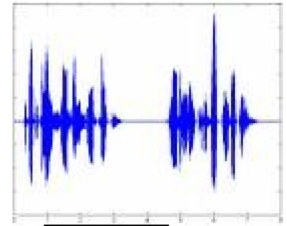
~~Semi~~-Supervised Learning

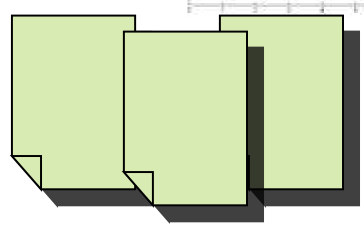
Supervised Learning = learning from labeled data. Dominant paradigm in Machine Learning.

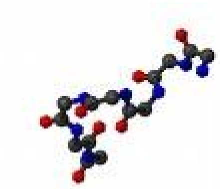
- E.g, say you want to train an email classifier to distinguish spam from important messages
- Take sample S of data, labeled according to whether they were/weren't spam.
- Train a classifier (like SVM, decision tree, etc) on S. Make sure it's not overfitting.
- Use to classify new emails.

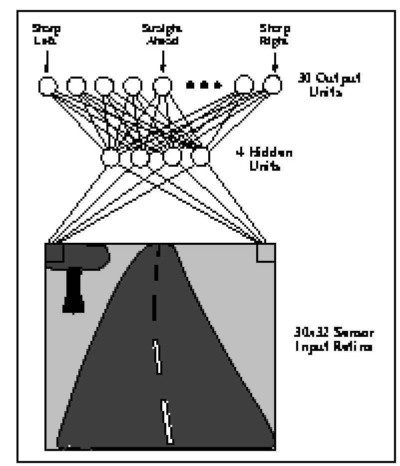
Basic paradigm has many successes

- recognize speech,
- steer a car,
- classify documents
- classify proteins
- recognizing faces, objects in images
- ...









However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,...

Unlabeled data is much cheaper.

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Speech

Customer modeling

Images

Protein sequences

Medical outcomes

Web pages

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Task: speech analysis

[From Jerry Zhu]

- Switchboard dataset
- telephone conversation transcription
- 400 hours annotation time for each hour of speech

film ⇒ `f ih_n uh_gl_n m`

be all ⇒ `bcl b iy iy_tr ao_tr ao l_dl`

However, for many problems, labeled data can be rare or expensive.

Need to pay someone to do it, requires special testing,…

Unlabeled data is much cheaper.

Can we make use of cheap unlabeled data?

Learning Problems

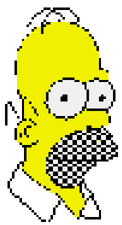
- Supervised learning:
  - Given a sample consisting of object-label pairs (x_i, y_i), find the predictive relationship between objects and labels.

- Un-supervised learning:
  - Given a sample consisting of only objects, look for interesting structures in the data, and group similar objects.

- What is Semi-supervised learning?
  - Supervised learning + Additional unlabeled data
  - Unsupervised learning + Additional labeled data

Semi-Supervised Learning

Can we use unlabeled data to augment a small labeled sample to improve learning?



But unlabeled data is missing the most important info!!

But maybe still has useful regularities that we can use.

Bu

Bu

But…

Semi-Supervised Learning

Substantial recent work in ML. A number of interesting methods have been developed.

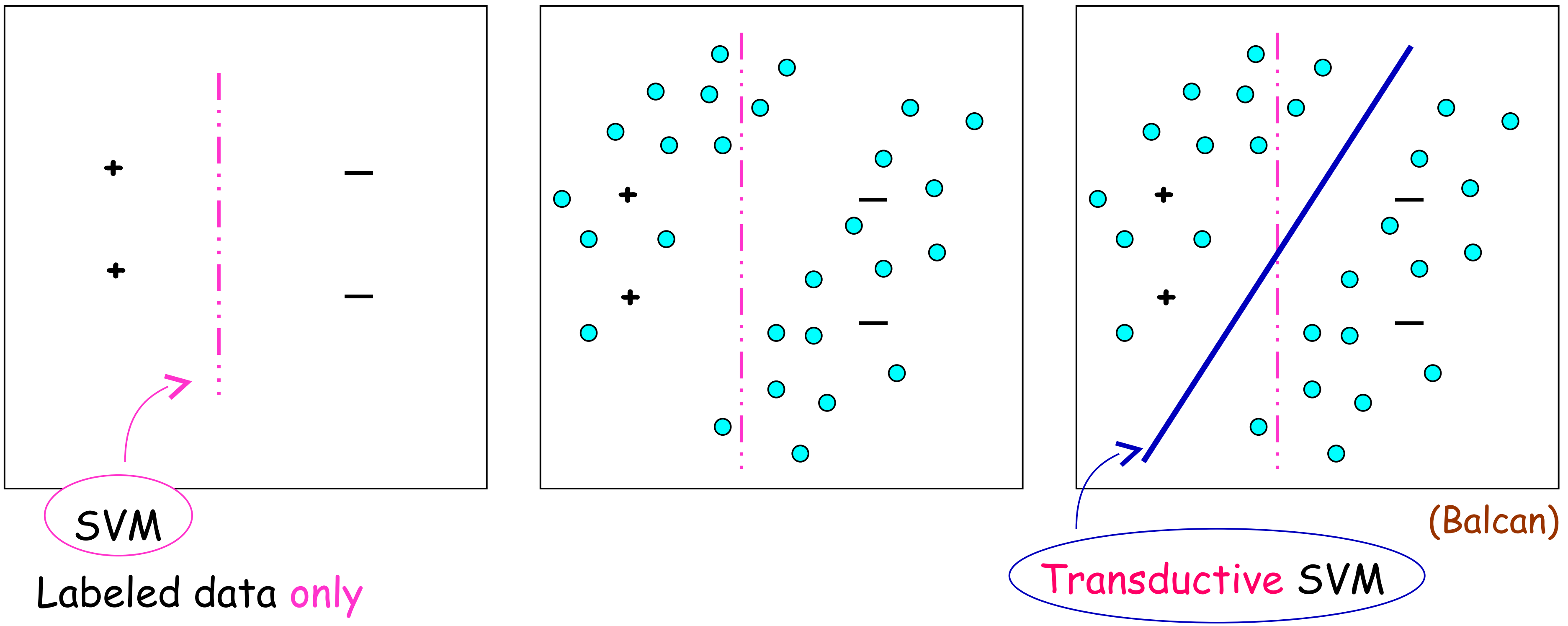
This lecture:

- Discuss several diverse methods for taking advantage of unlabeled data.
- General framework to understand when unlabeled data can help, and make sense of what’s going on.

Motivation for SSL (Belkin & Niyogi)

- Pragmatic:
  - Unlabeled data is cheap to collect.
  - Example: Classifying web pages,
    - There are some annotated web pages.
    - A huge amount of un-annotated pages is easily available by crawling the web.
- Philosophical:
  - The brain can exploit unlabeled data.

Intuition



Inductive vs.Transductive

- Transductive: Produce label only for the available unlabeled data.
  - The output of the method is not a classifier.

- Inductive: Not only produce label for unlabeled data, but also produce a classifier.

- SSL: Given labeled training data $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{L}$, unlabeled data $\mathcal{U} = \{\mathbf{x}_j\}_{j=L+1}^{L+U}$, learn a function f
- In SSL, f is used to predict labels for the future test data
- This is called Inductive Learning (learning a function to be applied on test data). Semi-supervised learning is therefore inductive.
- Transductive Learning: Given labeled training data $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{L}$, unlabeled data $\mathcal{U} = \{\mathbf{x}_j\}_{j=L+1}^{L+U}$
- Transductive Learning: No explicit function is learned. We don't get some "future" test data. All we care about is the predictions for $\mathcal{U}$
- Transductive Learning: The set $\mathcal{U}$ is the test data and is available at the training time

Two Algorithmic Approaches

- Classifier based methods:
  – Start from initial classifier(s), and iteratively enhance it (them)

- Data based methods:
  – Discover an inherent geometry in the data, and exploit it in finding a good classifier.

Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

EM

(Dempster et al 1977)

- Use EM to maximize the joint log-likelihood of labeled and unlabeled data:

$$\sum_i \log \left(P(y_i|\pi) P(x_i|y_i, \theta) \right) +$$

L_l : Log-likelihood of labeled data

$$\sum_j \log \left(\sum_y P(y|\pi) P(x_j|y, \theta) \right)$$

L_u : Log-likelihood of unlabeled data

Stable Mixing of Information

(Corduneanu 2002)

- Use **λ** to combine the log-likelihood of labeled and unlabeled data in an optimal way:

$$(1-\lambda)L_l + \lambda L_u$$

- EM can be adapted to optimize it.

- Additional step for determining the best value for **λ.**

EM_λ Operator

- E and M steps update the value of the parameters for an objective function with particular value of **λ**.

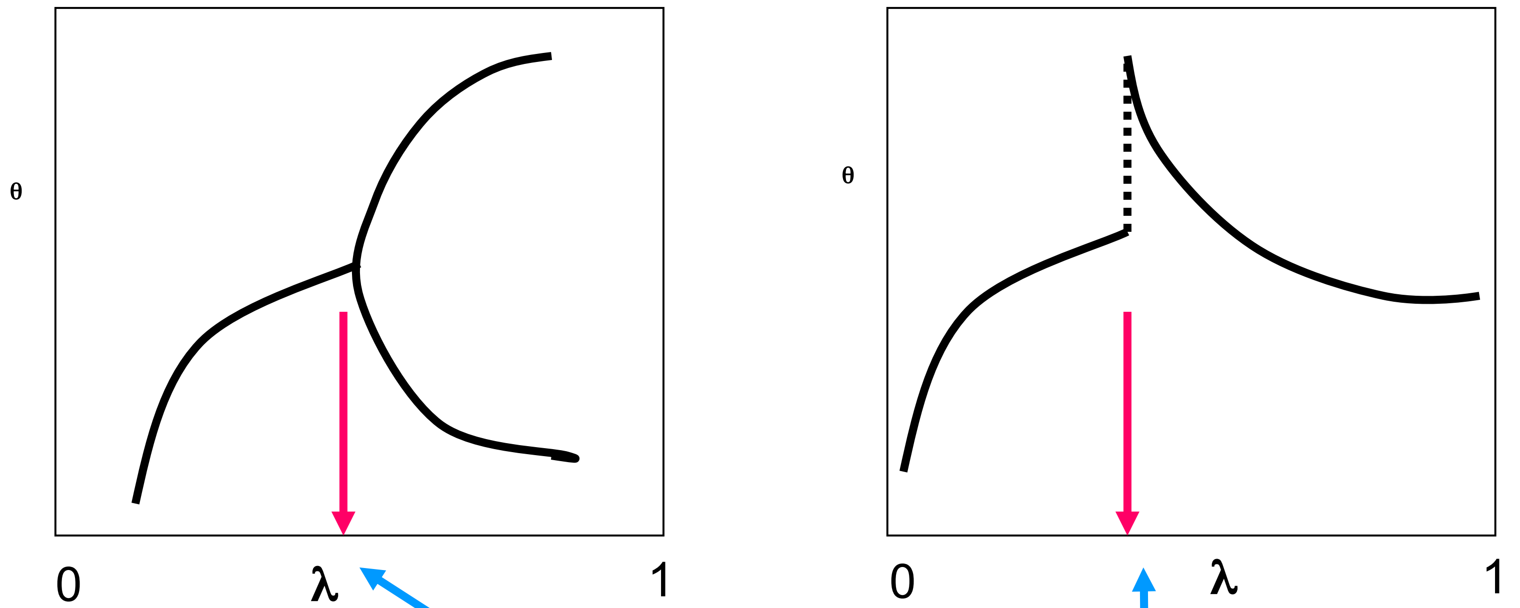
- Name these two steps together as EM_λ operator:

$$\theta^{new} = EM_\lambda(\theta)$$

- The optimal value of the parameters is a fixed point of the EM_λ operator:

$$\theta = EM_\lambda(\theta)$$

Path of solutions



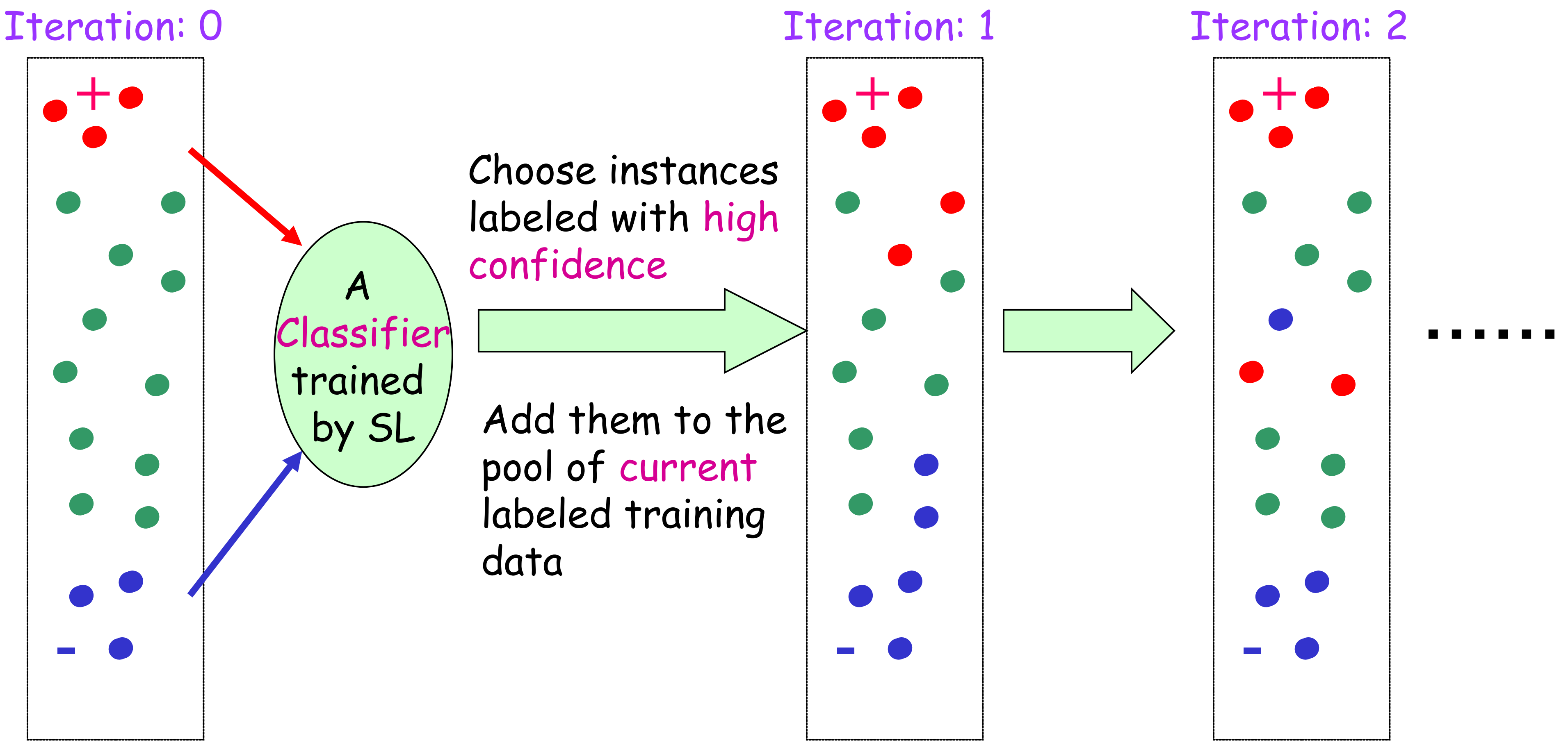
- How to choose the best **λ** ? $(1-\lambda)L_l + \lambda L_u$
  - By finding the path of optimal solutions as a function of **λ**
    - Choosing the first λ where a bifurcation or discontinuity occurs; after such points labeled data may not have an influence on the solution.
  - By cross-validation on a held out set. (Nigam et al 2000)

Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

The Yarowsky Algorithm

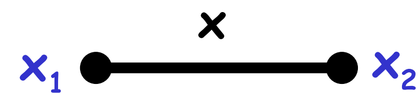
(Yarowsky 1995)



Co-Training

(Blum and Mitchell 1998)

- Instances contain two sufficient sets of features
  - i.e. an instance is $x=(x_1,x_2)$
  - Each set of features is called a View



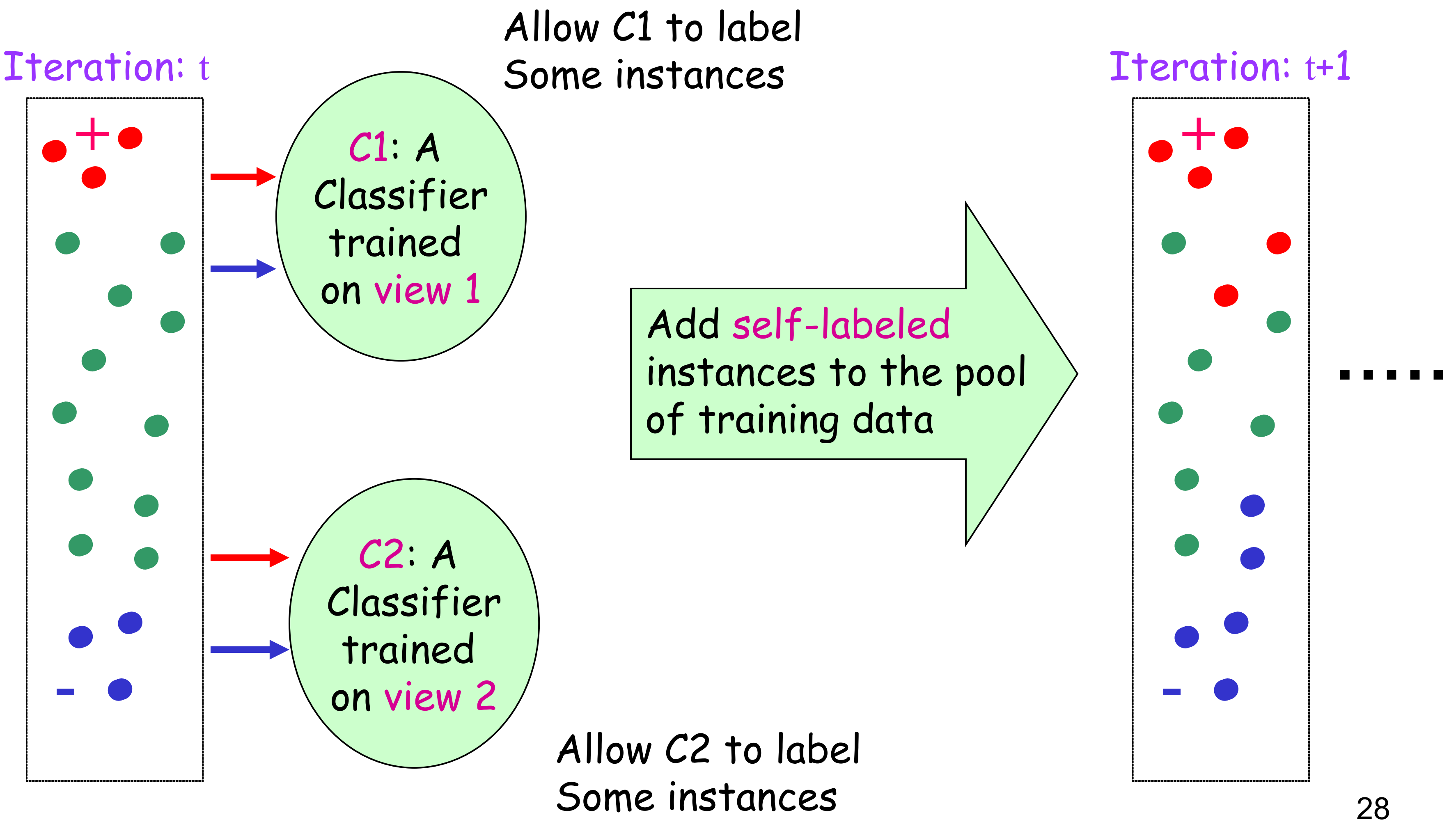
- Two views are independent given the label:

$$P(x_1|x_2, y) = P(x_1|y)$$
$$P(x_2|x_1, y) = P(x_2|y)$$

- Two views are consistent:

$$\exists c_1, c_2 : c^{opt}(x) = c_1(x_1) = c_2(x_2)$$

Co-Training



Allow C1 to label
Some instances
Iteration: t
Iteration: t+1
+
-
C1: A Classifier trained on view 1
C2: A Classifier trained on view 2
Add self-labeled instances to the pool of training data
Allow C2 to label
Some instances

- Given labeled data L and unlabeled data U
- Create two labeled datasets L_1 and L_2 from L using views 1 and 2
- Learn classifier $f^{(1)}$ using L_1 and classifier $f^{(2)}$ using L_2
- Apply $f^{(1)}$ and $f^{(2)}$ on unlabeled data pool U to predict labels
    - Predictions are made only using their own set (view) of features
- Add K most confident predictions $((\mathbf{x}, f^{(1)}(\mathbf{x}))$ of f_1 to L_2
- Add K most confident predictions $((\mathbf{x}, f^{(2)}(\mathbf{x}))$ of f_2 to L_1
- Note: Absolute margin could be used to measure confidence
- Remove these examples from the unlabeled pool
- Re-train $f^{(1)}$ using L_1, $f^{(1)}$ using L_2

Agreement Maximization

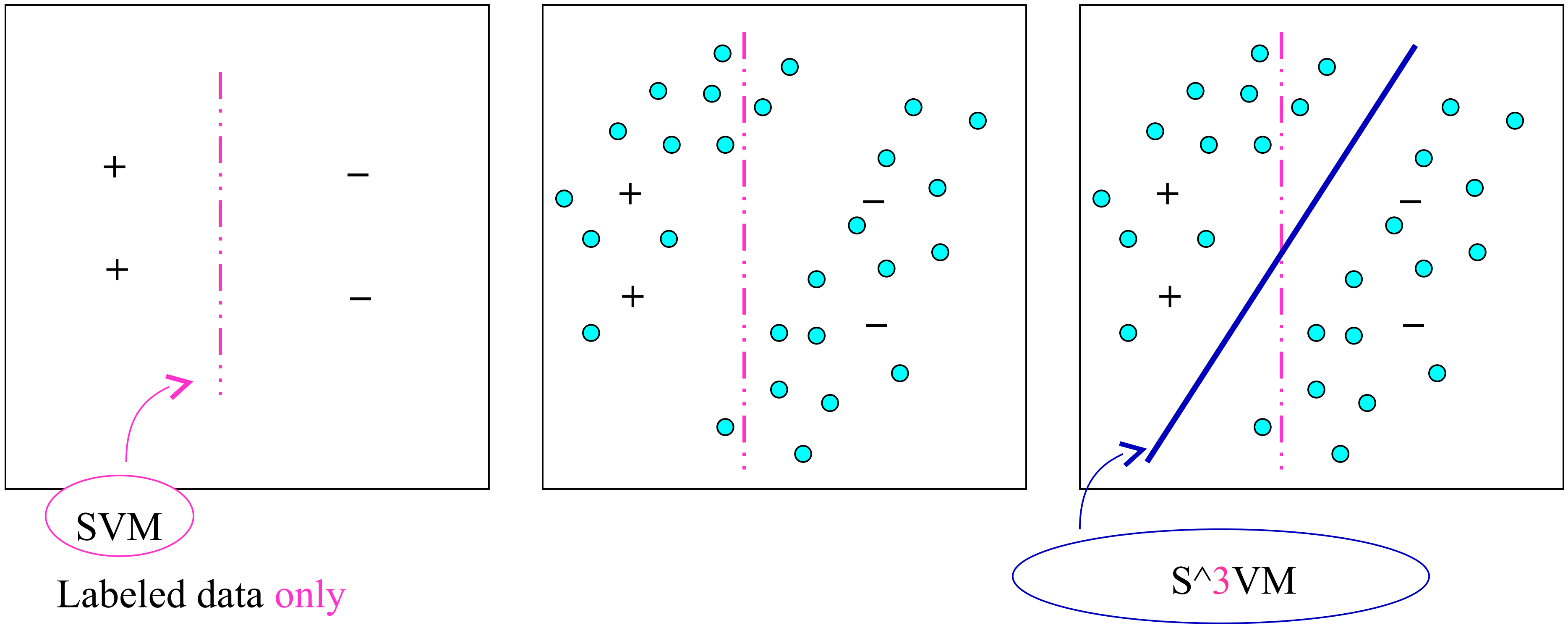
(Leskes 2005)

- A side effect of the Co-Training: Agreement between two views.

- Is it possible to pose agreement as the explicit goal?
  - Yes. The resulting algorithm: Agreement Boost

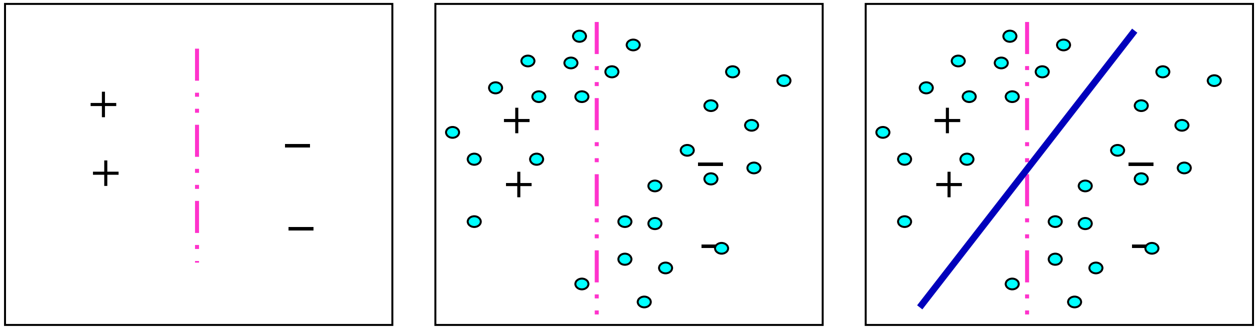
S^{3}VM [Joachims98]

- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)



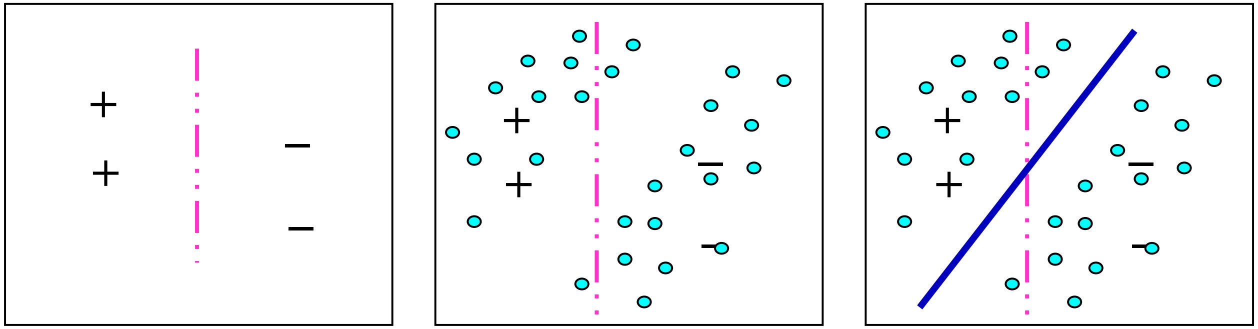
S^{3}VM [Joachims98]

- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)
- Unfortunately, optimization problem is now NP-hard. Algorithm instead does local optimization.
  - Start with large margin over labeled data. Induces labels on U.
  - Then try flipping labels in greedy fashion.



S^3VM [Joachims98]

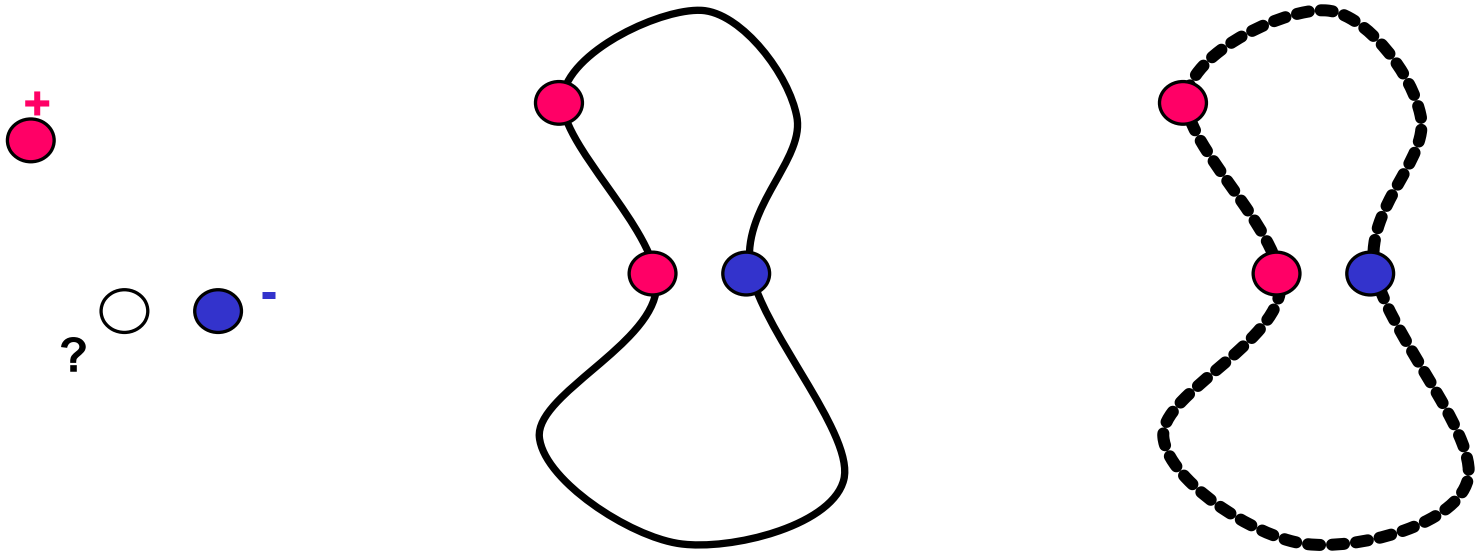
- Suppose we believe target separator goes through low density regions of the space/large margin.
- Aim for separator with large margin wrt labeled and unlabeled data. (L+U)
- Unfortunately, optimization problem is now NP-hard. Algorithm instead does local optimization.
  - Or, branch-and-bound, other methods (Chapelle etal06)
- Quite successful on text data.



Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

Data Manifold



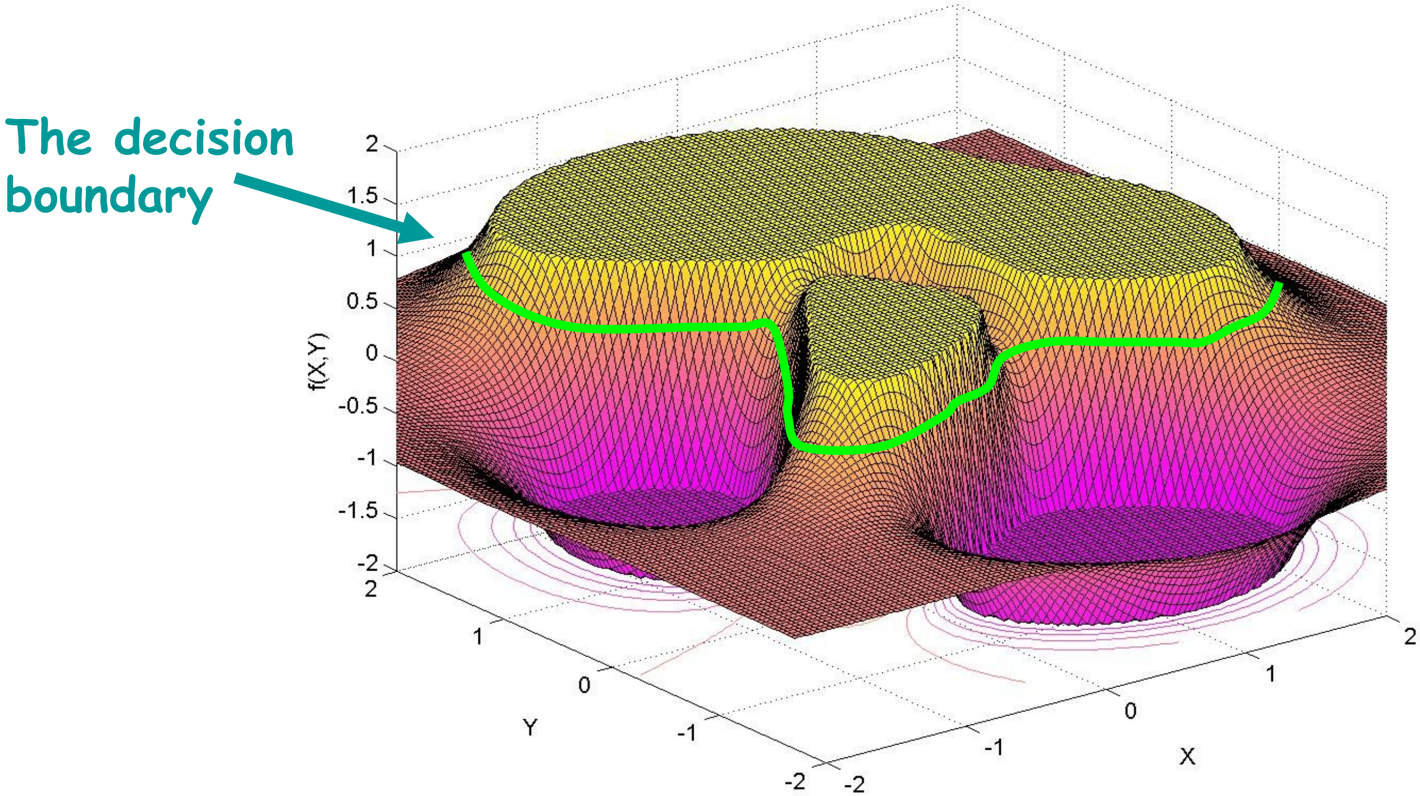
- What is the label?
- Knowing the geometry affects the answer.
  - Geometry changes the notion of similarity.
  - Assumption: Data is distributed on some low dimensional manifold.
- Unlabeled data is used to estimate the geometry.

Smoothness assumption

- Desired functions are smooth with respect to the underlying geometry.
  - Functions of interest do not vary much in high density regions or clusters.
    - Example: The constant function is very smooth, however it has to respect the labeled data.

- The probabilistic version:
  - Conditional distributions $P(y|x)$ should be smooth with respect to the marginal $P(x)$.
    - Example: In a two class problem $P(y=1|x)$ and $P(y=2|x)$ do not vary much in clusters.

A Smooth Function



- Cluster assumption: Put the decision boundary in low density area.
  - A consequence of the smoothness assumption.

What is smooth? (Belkin&Niyogi)

- Let $f : \mathcal{M} \to R$. Penalty at $x \in \mathcal{M}$:

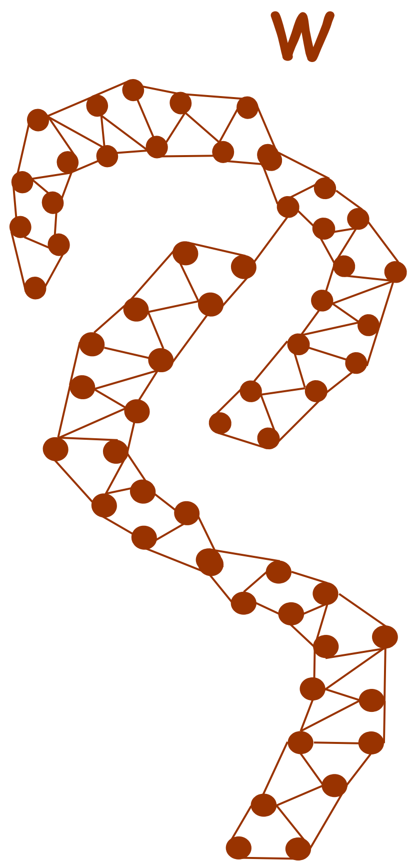
$$\frac{1}{\delta^k} \int_\Delta (f(x) - f(x+\delta))^2 p(x)\, d\delta \approx ||\nabla f||^2 p(x)$$

- Total penalty:

$$\int_\mathcal{M} ||\nabla f||^2 p(x)\, dx$$

- p(x) is unknown, so the above quantity is estimated by the help of unlabeled data:

$$\sum_{i,j} (f(x_i) - f(x_j))^2 W_{ij}$$



(Krishnapuram)

Manifold Regularization

(Belkin et al 2004)

Data dependent regularization

$$f^{opt} = \arg\min_{f \in H} \lambda_I ||f||_I^2 + \lambda_k ||f||_k^2 + \frac{1}{l}\sum_i^l (f(x_i) - y_i)^2$$

Smoothness term: Unlabeled data

Function complexity: Prior belief

Fitness to Labeled data

- Where:
  - H is the RKHS (Reproducing Kernal Hilbert Space) associated with kernel k(.,.)
  - Combinatorial laplacian can be used for smoothness term:

$$||f||_I^2 = f^T \cdot (D - W) \cdot f = \sum_{i,j} (f(x_i) - f(x_j))^2 W_{ij}$$

The Representer Theorem

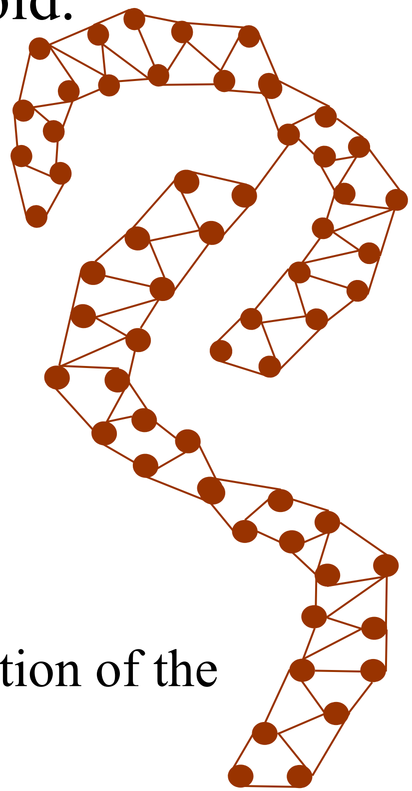
- The Representer theorem guarantees the following form for the solution of the optimization problem:

$$f^{opt}(.) = \sum_{i=1}^{l+u} \alpha_i k(x_i, .)$$

Harmonic Mixtures

(Zhu and Lafferty 2005)

- Data is modeled by a mixture of Gaussians.
  - Assumption: Look at the mean of Gaussian components, they are distributed on a low dimensional manifold.



- Maximize the objective function:

$$\lambda L(\theta) - (1 - \lambda)\varepsilon(\theta)$$

  - θ includes mean of the Gaussians and more.
  - $L(\theta)$ is the likelihood of the data.
  - $\varepsilon(\theta)$ is taken to be the combinatorial laplacian.
    - Its interpretation is the energy of the current configuration of the graph.

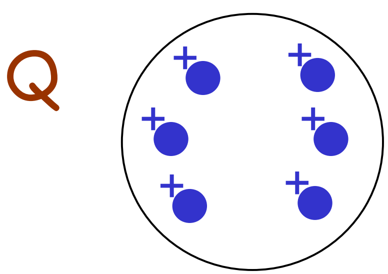
Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

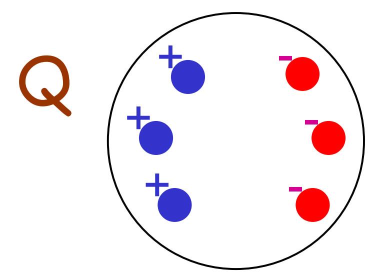
Mutual Information

- Gives the amount of variation of y in a local region Q:

$$I_Q(x,y) = \sum_y \int_Q p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$



- I(x,y) = 0
- Given the label is +, we cannot guess which (x,+) has been chosen (independent).



- I(x,y) = 1
- Given the label is +, we can somehow guess which (x,+) has been chosen.

Information Regularization

(Szummer and Jaakkola 2002)

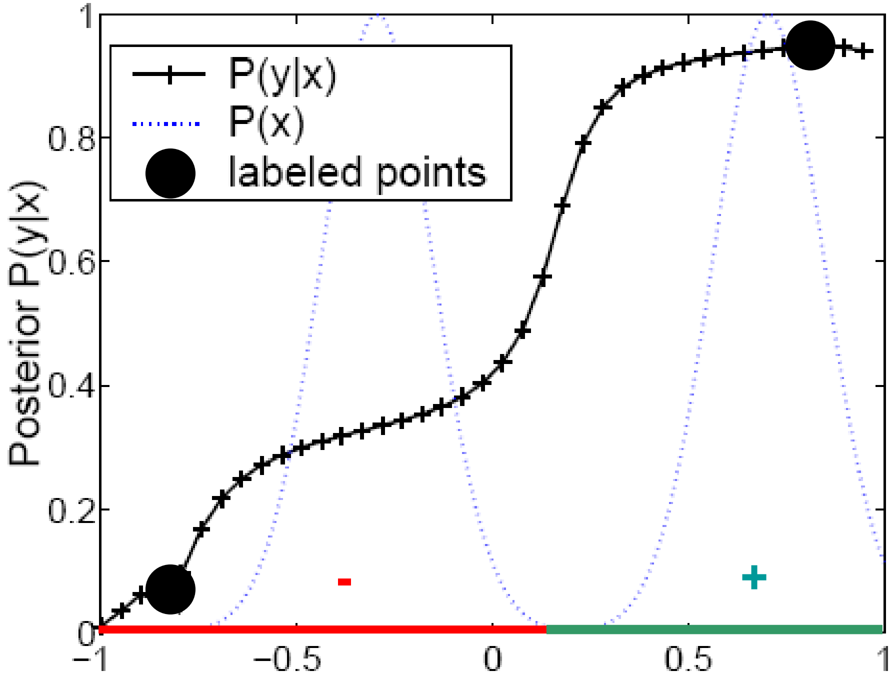
- We are after a good conditional P(y|x).
  - Belief: Decision boundary lays in low density area.
  - P(y|x) must not vary so much in high density area.

- Cover the domain with local regions, the resulting maximization problem is:

$$p^{opt}(y|x) = \arg\max_{p(y|x)}$$

$$\sum_{i=1}^{l} \log p(y_i|x_i) - \lambda \int_{\mathcal{X}} p(x) Tr[F(x)]\, dx$$

Example



- A two class problem (Szummer&Jaakkola)

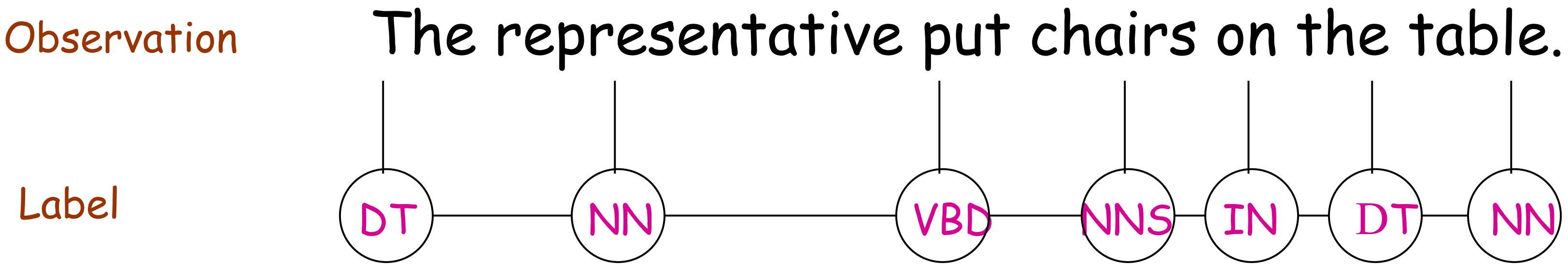
Return to smoothness

Outline of the talk

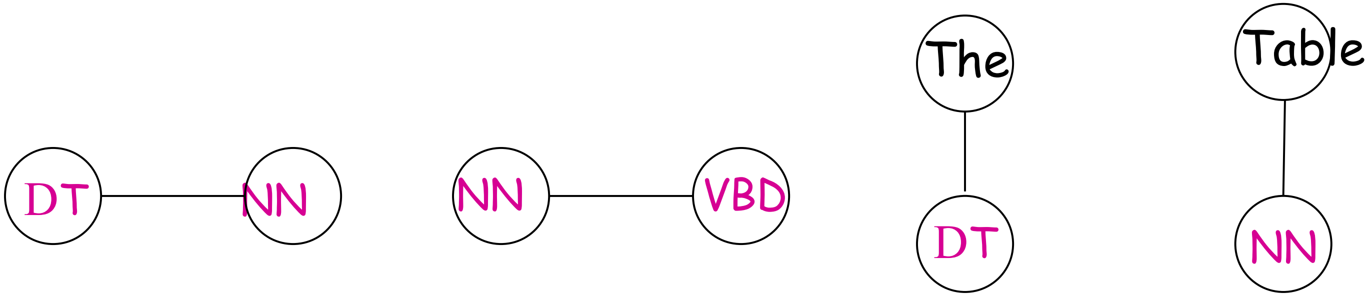
- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for Structured Prediction
- Conclusion

Structured Prediction

- Example: Part-of-speech tagging:



- The input is a complex object as well as its label.
  - Input-Output pair (x,y) is composed of simple parts.
  - Example: Label-Label and Obs-Label edges:



Scoring Function

- For a given x, consider the set of all its candidate labelings as Y_x.
  - How to choose the best label from Y_x?
- By the help of a scoring function S(x,y):

$$y = \arg\max_{y' \in Y_x} S(x, y')$$

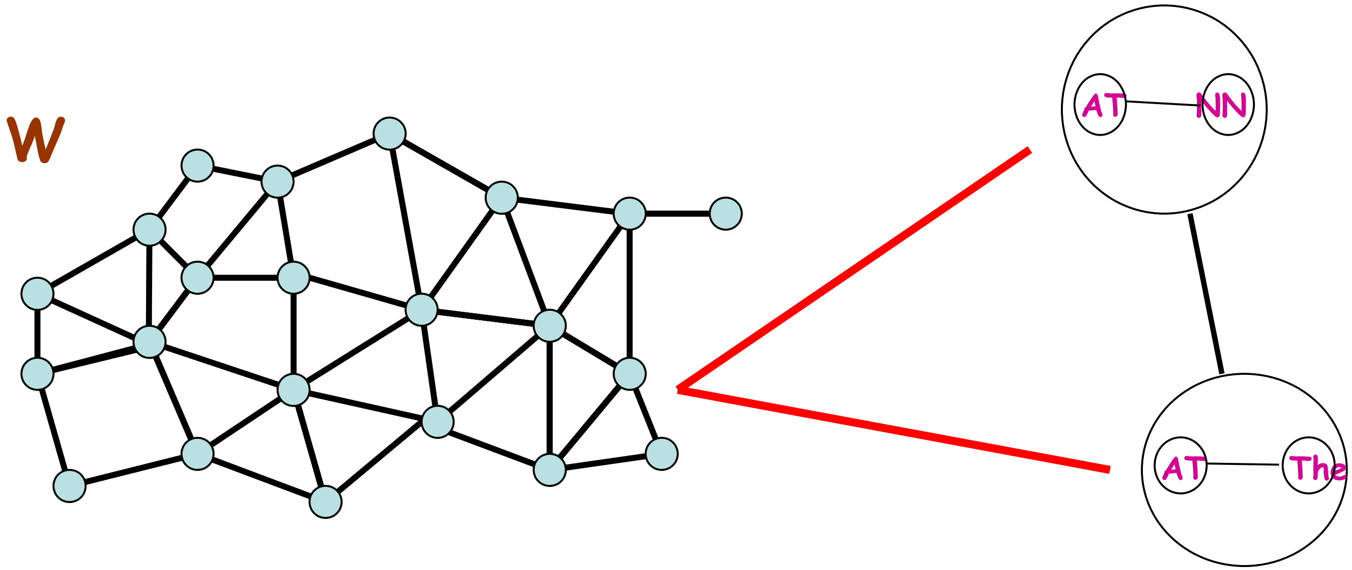
  - Assume S(x,y) can be written as the sum of scores for each simple part:

$$S(x, y) = \sum_{r \in R(x,y)} f(r)$$

    - R(x,y) the set of simple parts for (x,y).
  - How to find f(.)?

Manifold of "simple parts"

(Altun et al 2005)



- Construct d-nearest neighbor graph on all parts seen in the sample.
  - For unlabeled data, put all parts for each candidate.
- Belief: f(.) is smooth on this graph (manifold).

SSL for Structured Labels

- The final maximization problem:

Data dependent regularization

$$\arg\min_{f\in\mathcal{H}} \sum_i loss(f(x_i), y_i) + \lambda_k ||f||_k + \lambda_I \sum_{r,r'} W_{r,r'}(f(r) - f(r'))^2$$

Fitness to Labeled data

Function complexity: Prior belief

Smoothness term: Unlabeled data

- The Representer theorem:

$$f(.) = \sum_{r\in R(S)} \alpha_r k(r, .)$$

  - R(S) is all the simple parts of labeled and unlabeled instances in the sample.
  - Note that f(.) is related to $\alpha = (\alpha_1, ..., \alpha_{R(S)})$.

Modified problem

- Plugging the form of the best function in the optimization problem gives:

$$\arg\min_\alpha \sum_i loss(f_\alpha(x_i), y_i) + \alpha^T \cdot Q \cdot \alpha$$

  - Where Q is a constant matrix.
  - By introducing slack variables ε_i:

$$\arg\min_\alpha \sum_i \varepsilon_i + \alpha^T \cdot Q \cdot \alpha$$

Subject to

$$\forall i, loss(f_\alpha(x_i), y_i) \leq \varepsilon_i$$

Modified problem(cont'd)

- Loss function:

$$\arg\min_\alpha \sum_i \varepsilon_i + \alpha^T \cdot Q \cdot \alpha$$

Subject to

$$\forall i, loss(f_\alpha(x_i), y_i) \leq \varepsilon_i$$

  - SVM: $loss(f_\alpha(x), y) = \max_{y' \in Y_x} \Delta(x, y, y') + S_\alpha(x, y') - S_\alpha(x, y)$

Hamming distance

  - CRF: $loss(f_\alpha(x), y) = -S_\alpha(x, y) + \log \sum_{y' \in Y_x} \left(\exp S_\alpha(x, y') \right)$

  - Note that an α vector gives the f(.) which in turn gives the scoring function S(x,y). We may write $S_\alpha(x,y)$.

Outline of the talk

- Introduction to Semi-Supervised Learning (SSL)
- Classifier based methods
  - EM
  - Stable mixing of Complete and Incomplete Information
  - Co-Training, Yarowsky
- Data based methods
  - Manifold Regularization
  - Harmonic Mixtures
  - Information Regularization
- SSL for structured Prediction
- Conclusion

Conclusions

- We reviewed some important recent works on SSL.
- Different learning methods for SSL are based on different assumptions.
  - Fulfilling these assumptions is crucial for the success of the methods.
- SSL for structured domains is an exciting area for future research.

Thank You

References

- Adrian Corduneanu, Stable Mixing of Complete and Incomplete Information, Masters of Science thesis, MIT, 2002.

- Kamal Nigam, Andrew McCallum, Sebastian Thrun and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3), 2000.

- A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39* (1), 1977.

- D. Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In Proceedings of the 33rd Annual Meeting of the ACL*, 1*995.

- A. Blum, and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In Proceedings of the of the COLT*, 1*998.

References

- B. Leskes. The Value of Agreement, A New Boosting Algorithm. In Proceedings of the of the COLT*, 2005*.

- M. Belkin, P. Niyogi, V. Sindhwani. Manifold Regularization: a Geometric Framework for Learning from Examples. University of Chicago CS Technical Report TR-2004-06, 2004..

- M. Szummer, and T. Jaakkola. Information regularization with partially labeled data. Proceedings of the NIPS, 2002.

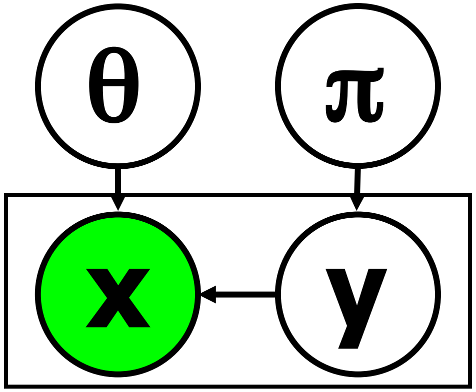
- Y. Altun, D. McAllester, and M. Belkin. Maximum Margin Semi-Supervised Learning for Structured Variables. Proceedings of the NIPS, 2005.

Further slides for questions…

Generative models for SSL

- Class distributions P(x|y,θ) and class prior P(y|π) are parameterized by θ and π, and used to derive:

$$P(y|x,\theta,\pi) \propto P(x|y,\theta) \cdot P(y|\pi)$$



(Seeger)

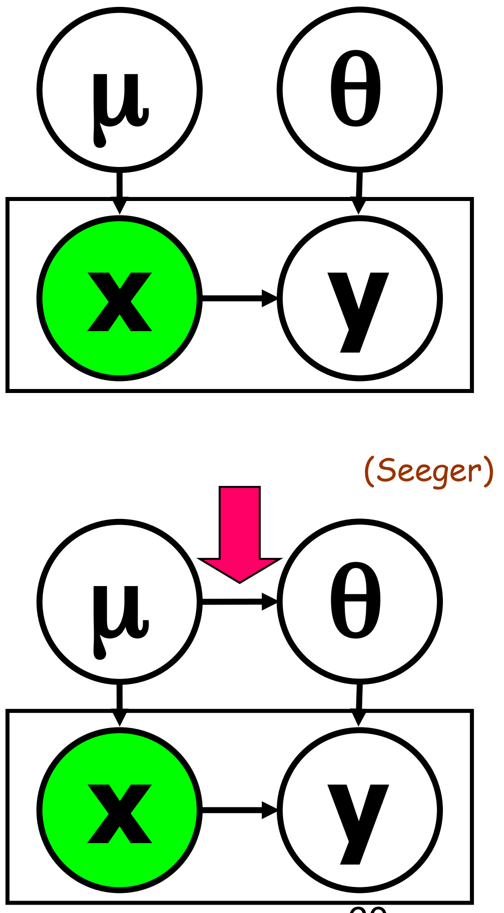
- Unlabeled data gives information about the marginal P(x|θ,π) which is:

$$P(x|\theta,\pi) = \sum_y P(x|y,\theta) \cdot P(y|\pi)$$

- Unlabeled data can be incorporated naturally!

Discriminative models for SSL

- In Discriminative approach $P(y|\mathbf{x},\theta)$ and $P(x|\mu)$ are directly modeled.
- Unlabeled data gives information about μ, and $P(y|x)$ is parameterized by θ.
- If μ affects θ then we are done!
  - Impossible: θ and μ are independent given unlabeled data.
- What is the cure?
  - Make μ and θ a priori dependent.
  - **Input Dependent Regularization**



Fisher Information

Fisher Information matrix:

$$F(x) = E_{p(y|x)}[\nabla_x \log p(y|x) \cdot \nabla_x \log p(y|x)^T]$$