Semi-Supervised Learning

Unsupervised and Supervised
• Unsupervised learning: no labels available
  • Clustering, manifold learning
  • Data exploration / discovering of structure
• Supervised learning: labeling provided
  • Construct a mapping from feature vectors $x$ to labels (or other outputs) $y$
  • Requires enough labeled training data

Semi-Supervised Learning [SSL]
• Class of machine learning techniques that exploits both labeled and unlabeled in training
• Somewhere between unsupervised learning and supervised learning
• Typical situation: little labeled data, large amount of unlabeled data

Graphically Speaking
• Unlabeled data cheap / easy to get
  • Data annotation is boring
    • Annotation of speech data
  • Labeling require experts
    • Medical images
  • Special, expensive device or method
    • Bioinformatics
  • Graduate student is on vacation [Zhu 2007]

Graphically Speaking
• [Sindhwani et al, 2005]
Graphically Speaking

Two Views

- SSL is
  - Unsupervised learning with additional constraints
  - Considered supervised learning with additional information
- Latter view is most common
- So why not call it hyper-supervised learning [HSL]?

What we need is Assumptions

- "Smoothness" assumption / local consistency
  - Nearby point have similar outputs
- Cluster assumption / global consistency
  - Points on same cluster have same output
- Low-density separation
  - Decision boundaries are in low-density regions
- Manifold assumption
  - Data lies on lower-dimensional manifold

E.g.: Manifold Assumption

- Very high dimensional data indeed on a lower-dimensional manifold?
- Unlabeled data can be used to reduce the dimensionality / change representation
- In lower-dimensional space, estimation etc. much more accurate
- Result: better classifiers

What we need is Assumptions

- Some assumptions correspond to ones also exploited in supervised or unsupervised learning
- Many methods make no explicit use of them
- SLL is not only in availability of [un]labeled data; additional assumptions also essential
  - Often leads to specific algorithms
  - Not necessarily "in between" supervised and unsupervised methods

Self-Learning / Self-Training

- Simple wrapper approach to exploit unlabeled data
  - Train on labeled data
  - Classify unlabeled data
  - Pick points classified e.g. above certain confidence
  - Include in training set and retrain, etc.
Self-Learning / Self-Training

- Local consistency assumption implicit
- Supervised learning with additional constraints
- Compare to EM-like clustering
- Possible problems with convergence?
- Errors can get amplified easily
- Generic approach

Mixture of Gaussians

- Use EM to find parameters, however known labels are not hidden
  \[
  \sum_{y} \log(P(x|y, \theta)P(y|\beta)) + \sum_{y} \log(P(x|y, \theta)P(y|\beta))
  \]
- Directly related to self-learning
- More of the form: unsupervised learning with additional constraints
- Generative model
  May be preferable as it possible enables incorporating additional knowledge more easily

Transductive SVMs

- Implements assumption on low-density separation
- Original objective function altered
  - Include unlabeled data
  - Force decision boundary away from this data, i.e., into low-density
  - Retain large margin

Transductive SVMs

- Adapted objective function
  \[
  \min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{labeled} \max(0, 1 - y_i(w^T x_i + b)) + C \sum_{i=1}^{unlabeled} \max(0, 1 + y_i(w^T x_i + b))
  \]
- Problem: \( L' \) is nonconvex
  - Various optimization methods to find some solution
  - Solution generally not globally optimal
  - Different optimization, different solution
Other Problem for TSVMs

Graph-based Methods [et al]
• Graph is constructed on all data
• Instances strongly connected presumably have same label
• Employ algorithms like, or related to, graph cuts
  • Cf. LLE, Laplacian and Hessian embedding, etc.
• Provides labeling for all data, but transductive
• Typically: performance good, when graph fits data, if not...

Graph-based Methods [et al]
• Possible approach:
  • Labels $f(x)$ on all nodes $x$ should be such that they minimize
    \[ E(f) = \frac{1}{2} \sum_{i<j} w_{ij}[f(i) - f(j)]^2 + \frac{1}{4} \sum_i \| f(i) - f(j) \|^2 \]
    • Weights indicate similarity
    • Some $f(x)$ are prefixed
    • Related to finding minimum cut in graph

Graph-based Methods [et al]
• Other possibility, employing underlying structure better
• Information is “diffused” according to structure of underlying data

Final Remarks
• Small training set “assumption”
• Specialized algorithms
  • Significantly increased training time
  • Potentially very powerful
• What if more classes than indicated?
• Should study the theoretical setting in which $p(x)$ is really known?
• One should ask what SSL can do for supervised and unsupervised methods

References
• Chapelle et al, Semi-Supervised Learning, 2007
• Chapelle, A Taxonomy of Semi-Supervised Learning Algorithms, 2005
• Seeger, Learning with labeled and unlabeled data, 2002
• Blum et al, Semi-Supervised Learning Using Randomized Mincuts, 2004
• Sindhwani et al, Beyond the Point Cloud: from Transductive to Semi-supervised Learning, 2005
• Zhu, Semi-Supervised Learning Tutorial, 2007