Semi-Supervised Learning and Spectral Method

CAP6676: Knowledge Representation
Instructor: Guo-Jun Qi
Semi-supervised learning VS. supervised learning

• Supervised learning
  • Building a classifier with only labeled examples (training set)
  • Drawback
    • The number of training examples might be limited
    • Does not utilize a large volume of unlabeled training examples

• Semi-supervised learning
  • Still accessing the labeled training examples
  • Leveraging unlabeled examples as well, which specifies the distribution of testing examples
  • Leveraging unlabeled examples to compensate for the limited training set
Running example

• Two labeled examples of different classes
  • Supervised learning – SVM
    • The optimal decision boundary between two classes is the plan maximizing the margin
Running example

• Two labeled examples of different classes
  • Supervised learning – SVM
    • The optimal decision boundary between two classes is the plan maximizing the margin
  • Unsupervised learning
    • When we have access to additional unlabeled examples
    • Intuitively, the optimal model should make classification by the distribution formed by the available examples (labeled and unlabeled)
Manifold Assumption

• The population of all the examples do not fill up the whole feature space (ambient space).

• Instead, in many cases, they reside on hyper-surfaces of lower dimensions embedded in an ambient space.

• For example, changing the poses of a face will result in a 1D manifold parameterized by the angles.
Graph Representation for Manifold

• Manifold Graph
  • Vertex: labeled and unlabeled examples
  • Edge: similar examples

• Regularizing the classifier $f(x)$ by the graph
  • If $X_1$ and $X_2$ are connected in the graph
  • Then $f(X_1)$ and $f(X_2)$ are very likely to be the same
  • Graph is the cut into two parts below
Label Propagation

- Connect the data points that are close to each other
Label Propagation

• Connect the data points that are close to each other
• Propagate the class labels over the connected graph
Matrix Representation

• Similarity matrix \((W)\)
  • \(N \times N\) matrix, where \(N\) is the number of examples
  • Each entry \(W_{ij}\) is the similarity between \(X_i\) and \(X_j\)
Matrix Representations

- **Degree matrix (D)**
  - N x N diagonal matrix
  
  \[ D(i, i) = \sum_j w_{ij} \]

- Total weight of edges incident to vertex \( X_i \)
Matrix Representations

• Normalized Similarity Matrix (S)

\[ S = D^{-0.5} W D^{-0.5} \]

• The similarity of two examples is less significant if both examples are similar to many other examples
  • For example, the similarity between two common
Normalized Similarity Matrix

\[ S = D^{-0.5} WD^{-0.5} \]
Initial Labels and Predictions

• Let Y denote the initial assignment of labels
  • Positive label: $Y_j = +1$
  • Negative label: $Y_j = +1$
  • Unlabeled: $Y_j = +0$

• Let F denote the predicted labels
  • $F_j > 0$ denotes positive label
  • $F_j < 0$ denotes negative label
Initial Label and Prediction

Initial Label

\[ Y = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \]

Prediction

\[ F = \begin{bmatrix} f_1 \\ \vdots \\ f_i \\ \vdots \\ f_n \end{bmatrix} \]
Label Propagation

• One Iteration
  • $F = Y + \alpha SY = (1 + \alpha S)Y$
  • $\alpha$ weights the importance of propagation against keeping the original labels

After one iteration, all the vertices connected to the labeled examples are labeled.
Label Propagation

• Two iterations
  • $F = Y + \alpha SY + \alpha^2 S^2 Y = (1 + \alpha S + \alpha^2 S^2)Y$

The examples labeled in the first iteration propagate the labels to their neighbors
Label Propagation

• More iterations
  • $F = \sum_{n=1}^{+\infty} (1 + \alpha^n S^n) Y = (1 - \alpha S)^{-1} Y$ by power-series expansion
Alternative Approach: Graph Partition

- Classification as graph partitioning
- Search for a classification boundary
  - Consistent with labeled examples
  - Partition with small graph cut
    - Graph cut is the number of edges across two partitions
    - Or the sum of similarities between the vertices belonging to two different partitions
Spectral Clustering

• Express a bi-partition ($C_1$ and $C_2$) as a vector

$$f_i = \begin{cases} 
1 & \text{if } x_i \in C_1 \\
-1 & \text{if } x_i \in C_2
\end{cases}$$

• Spectral clustering minimizes the graph-cut of two partitions by finding a non-trivial vector $f$ that minimizes

$$g(f) = \sum_{i,j \in \mathcal{V}} w_{ij} (f_i - f_j)^2 = f^T L f$$
Spectral Bi-Partitioning with Minimal Graph-Cut

• Pre-preprocessing
  • Building Laplacian matrix $L$ of the graph from the similarity matrix

• Decomposition
  • Finding the eigenvector of the Laplacian matrix $L$ corresponding to the second smallest eigenvalue
    • The eigenvector corresponding to the smallest eigenvalue is always a vector with the same element, thus resulting in a trivial graph cut with a single partition.

• Split all the vertices by the resultant eigenvector
  • Choosing a proper threshold
Semi-supervised learning

• Criterion 1: Small graph cut

\[ g(f) = \sum_{i,j \in V} w_{ij}(f_i - f_j)^2 = f^T L f \]

• Criterion 2: consistent with the labeled examples

\[ f = \begin{bmatrix} y_l \\ f_u \end{bmatrix}, \quad L = \begin{bmatrix} L_{ll} & L_{lu} \\ L_{ul} & L_{uu} \end{bmatrix} \]

\[ \min_{f_u} f^T L f \]

• Only minimizing over the unlabeled part
Semi-supervised learning

• Alternatively, introducing a penalty for the violation on the labeled examples

\[
\min_f \, f^T L f + (f - y)^T C (f - y)
\]

\[C_{ii} = 1 \quad \text{if } x_i \text{ is labeled}\]

• Both methods have quadratic objective function
  • Having closed-form solution by setting the derivatives to the decision variables to zero.
Summary

• Motivation: Semi-supervised VS. supervised
  • Exploring the distribution on the manifold formed by the unlabeled examples to classify them

• Method I: label propagation

• Method II: Semi-supervised learning with small graph-cut criterion

• Next Lecture: Transductive SVM