Kernel Methods in Pattern Analysis Course Project(CS6011)

Laplacian Twin SVM for Semi Supervised Classification

Ramnandan Krishnamurthy[1] Aritra Ghosh[2]

[1]Department of Computer Science and Engineering
IIT Madras

[2]Department of Computer Science and Engineering
IIT Madras

14^{th} May ,2015

Outline

Motivation

Semi-supervised learning

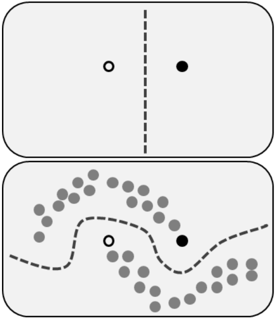

Figure: Two different classifiers with and without labelled points

- Most of the data we have today is unlabelled.
- Getting Labelled data is expensive.
- Can we use the unlabelled data "effectively" to build better classifiers?

- Learns two non-parallel hyperplanes, one for each of the two classes such that each hyperplane is closer to points of one class and are at least unit distance from the other class points (with slacks allowed).
- Consider a binary **supervised** classification problem of m_1 positive class and m_2 negative class points. Let $A \in \mathbb{R}^{m_1 \times n}$ and $B \in \mathbb{R}^{m_2 \times n}$ denote points of each class.
- For linear case, the TWSVM determines two nonparallel hyperplanes:

$$f_+(x) = w_+^T x + b_+$$

$$f_-(x) = w_-^T x + b_-$$

- The TWSVM solves two quadratic optimization problems corresponding to each class:

$$min_{w_+,b_+,\xi}\frac{1}{2}||Aw_+ + e_+b_+||_2^2 + c_1e_-^T\xi$$

$$s.t - (Bw_+ + e_-b_+) + \xi \geq e_-, \xi \geq 0$$

and

$$min_{w_-,b_-,\eta}\frac{1}{2}||Bw_+ + e_-b_-||_2^2 + c_2e_+^T\eta$$

$$s.t(Aw_- + e_+b_-) + \eta \geq e_+, \eta \geq 0$$

- First term in the objective function captures the sum of squares of distance of points of that particular class to the hyperplane.
- Second term takes into account the sum of errors, thus reducing the misclassification due to points belonging to the other class.
- Constraint requires the hyperplane to be at least one distance away from other class points, allowing for slacks.

- TWSVMs give rise to two smaller sized QPPs. Approximately 4 times faster than usual SVM. This is because complexity of usual SVM is $O(m^3)$ where m is the number of constraints (number of data points).
- Since constraints of each QPP in TWSVMs are points only in one class, ratio of runtimes is approximately:

$$\frac{m^3}{2 * (\frac{m}{2})^3} = 4$$

- TWSVMs can also handle preferential classification problems by solving only one smaller sized SVM.

- Regularization- Obtain smooth decision functions to avoid over-fitting.
- Fundamental Assumption(Smoothness)- The labels of two points that are close are same or similar.
- Laplacian SVM - Also considers the smoothness of the function(Well-posedness)

Laplacian SVM
Details

- The decision function is obtained by minimizing:

$$f^* = argmin_f \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma_H ||f||_H^2 + \gamma_M ||f||_M^2$$

- V is a loss function on labelled data - Square loss.
- $||f||_H^2$ controls the smoothness of f
- $||f||_M^2$ represents complexity of function in the intrinsic geometry of marginal distribution

$$||f||_M^2 = \sum_{i=1}^{l+u} \sum_{j=1}^{l+u} w_{ij}(f(x_i) - f(x_j))^2 = f^T L f$$

- γ_H, γ_M are the associated weights controlling importance - hyperparameters.

Similar to Semi Supervised SVM

$$\psi(w, w_o, t^U) = \frac{1}{2}||w||^2 + C_1 \sum_{n=1}^{N_L} \xi_n^L + C_2 \sum_{n=1}^{N_U} \xi_n^U$$

subject to the following constraints

$$t_n^L(w^T x_n^L + w_o) \geq 1 - \xi_n^L, n = 1, 2, ...n_L$$
$$\xi_n^L \geq 0, n = 1, 2, ..N_L$$
$$t_n^U(w^T x_n^U + w_o) \geq 1 - \xi_n^U, n = 1, 2, ...n_U$$
$$\xi_n^U \geq 0, n = 1, 2, ..N_U$$
$$t_n^U \in \{+1, -1\}$$

$$\psi^*(t^U) = min_{w,w_o} \psi(w, w_o, t_u)$$

and then find $t^{U*} = argmin_{t^U} \psi^*(t^U)$

Lap-TWSVM

- Extending the Lap-SVM to learn two non-parallel hyperplanes corresponding to each class (as in Twin-SVM)
- The decision functions learnt are:

$$f_+(x) = w_+^T x + b_+$$

$$f_-(x) = w_-^T x + b_-$$

- Following Lap-SVM, the regularization terms $||f_{+,-}||_H^2$ are defined as follows:

$$||f_{+,-}||_H^2 = \frac{1}{2}(||w_{+,-}||_2^2 + b_{+,-}^2)$$

- The Manifold regularization terms are given by:

$$||f_{+,-}||_M^2 = f_{+,-}^T L f_{+,-} = \frac{1}{(l+u)^2} \sum_{i,j=1}^{l+u} W_{i,j}(f_{+,-}(x_i) - f_{+,-}(x_j))^2$$

- Thus, the primal optimization problem consists of 2 QPPs:

$$min_{w_+,b_+,\xi}\frac{1}{2}||Aw_+ + e_+b_+||_2^2 + c_1e_-^T\xi + c_2(||w_+||_2^2 + b_+^2)$$
$$+c_3(w_+^TM^T + e^Tb_+)L(Mw_+ + eb_+)$$
$$\text{s.t} - (Bw_+ + e_-b_+) + \xi \geq e_-, \xi \geq 0$$

and

$$min_{w_-,b_-,\eta}\frac{1}{2}||Bw_+ + e_-b_-||_2^2 + c_2e_+^T\eta + c_2(||w_-||_2^2 + b_-^2)$$
$$+c_3(w_-^TM^T + e^Tb_-)L(Mw_- + eb_-)$$
$$\text{s.t}1(Aw_- + e_+b_-) + \eta \geq e_+, \eta \geq 0$$

Label Propagation

- Strategy is to compute a real-valued function $f : V \to R$ on graph G with the smoothness property - unlabelled data points that are nearby in the graph to have similar labels.
- Formulate it as a graph mincut problem where the objective function to be minimized is:

$$\sum_{i=1}^{l} L_{\infty}(y_i, f(i)) + \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(f(i) - f(j))^2$$

- The function that minimizes the above objective function is *harmonic*. The harmonic property means that f at each unlabelled data point is the weighted average of f at neighboring points.
- To compute the harmonic solution in a single shot instead of an iterative fashion, the matrices W, D are split into 4 blocks after the l^{th} row and column and f_u (function value for unlabelled points) is defined as:

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l$$

Experiments

- Experiments have been conducted on synthetic datasets (2moons and clock) and real world (ionosphere and credit from UCI).
- Metrics of interest:
  - Accuracy of the Classifier
  - Effect of the Manifold Regularization
  - Effect of size of unlabelled data (keeping total dataset size and labelled dataset size fixed respectively)
  - Running times
- Classification accuracy and running times compared against - Transductive SVM, Lap-SVM

Experiments
Clock Dataset

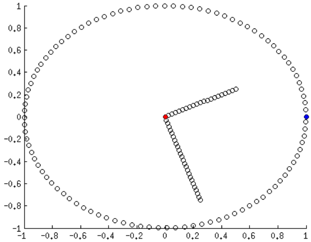

Figure: 1 labelled point from each class

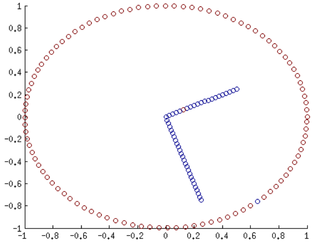

Figure: Predicted Output

Figure: Performance of Lap-TSVM on the Clock Dataset

Clock Dataset - $\sim$ 99.35%, 2moons dataset- $\sim$ 98.5%

Experiments
Accuracy

Dataset	Lap -TSVM	Lap-SVM	T-SVM
Two moons	98.35%	96.65%	89%
Clock	99.35%	97.2%	82.31%
Ionosphere	78.33%	76.2%	70.2%
Credit	75.6%	75.4%	71.4%

- Lap-TSVM outperforms others.
- Performance of Lap-TSVM very similar to Lap-SVM.

Effect of unlabelled data- Ionosphere

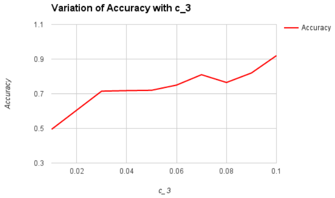

Figure: Variation of Accuracy with c_3

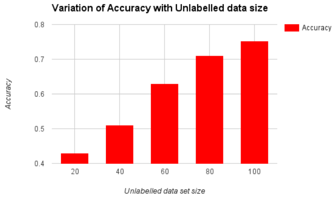

Figure: Variation of Accuracy with size of Unlabelled data

Dataset	Lap -TSVM	Lap-SVM	T-SVM
Ionosphere	17.32	53.16	-

Table: Running times(s) of the various Algorithms

- Graph based Label Propagation - 76.6% on Ionosphere and 72.4% on Credit dataset.
- Very poor results on 2moons and Clock dataset.

Conclusion

- Importance of Unlabelled data.
- Lap TSVM - exploits geometry information of the marginal distribution.
- Extends the Lap-SVM and TWSVM framework.
- Performs better than both the baselines.
- Running times order of magnitude lesser than TW-SVM.
- Compared to non-kernel Method- Graph Based Label Propagation.

References I

- Zhiquan Qi, Yingjie Tian, Yong Shi
  Laplacian twin Support Vector machine for semi supervised classification.
- Jayadeva, R. Khemchandani, Suresh Chandra
  Twin Support Vector Machines for Pattern Classification
- X. Zhu, Z. Ghahramani, J. Lafferty
  Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions