CSE5820: Machine Learning
Final Project: Crowd Learning [1]

Nayeff Najjar and Salem Alqahtani

December 2, 2014

Date Performed: Nov 30, 2014
Instructor: Prof. Jinbo Bi

1 Objective

To design an SVM based algorithm that classifies data that are labeled by multiple *teachers* where some of the teachers are not reliable.
Note: the fidelities of the teachers are not known a priori.

1.1 Problem Formulation

The Primal SVM Optimization problem can be written as:

$$\begin{aligned} \min_{\mathbf{w}\in\mathcal{R}^n} \quad & \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{m}\sum_{i=1}^{m}\max(1 - y_i\mathbf{x}_i\cdot\mathbf{w}, 0) \\ s.t. \quad & \lambda > 0 \end{aligned} \tag{1}$$

where $\mathbf{w} \in \mathcal{R}^n$ is a vector that contains the weights of the SVM hyperplane, $(\mathbf{x}_i, y_i)$ is the training sample-label pair $\forall\, i = 1, 2, \ldots m$, λ is the regularization factor, m is the number of training samples and n is the dimension of the data. Formulating the Lagrangian, one can derive the Dual form for the SVM Optimization problem, which is as the following:

$$\begin{aligned} \max_{\boldsymbol{\alpha}\in\mathcal{R}^m} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2\lambda}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j \\ s.t \quad & 0 \le \alpha_i \le \frac{1}{m}\,\forall i = 1, 2, \ldots m \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2 \ldots \alpha_m]^T$ is a vector that holds dual variables. The following can be observed about the SVM Dual problem:

- $\mathbf{x}_i$ is a support vector if:

$$\begin{aligned} \alpha_i > 0 \text{ if } \quad & y_i \mathbf{x} \cdot \mathbf{w} \leq 1 \\ \alpha_i = \tfrac{1}{m} \text{ if } \quad & y_i \mathrm{x}_i \cdot \mathbf{w} < 1 \end{aligned} \tag{3}$$

- For a bad teacher, there are many labels that are wrong
  $\Rightarrow$ most of α_i's for a bad teacher are $\frac{1}{m}$
  $\Rightarrow$ average of α_i's for a bad teacher is larger than the average of α_i's for all the data. i.e.,
  $\frac{1}{|\mathcal{S}^e|} \sum_{\mathbf{x}_i \in \mathcal{S}^e} \alpha_i > \frac{1}{m} \sum_{i=1}^{m} \alpha_i$ ($\mathcal{S}^e$ is the training data set for the evil teacher)

- For a good teacher; however,
  $\frac{1}{|\mathcal{S}^g|} \sum_{\mathbf{x}_i \in \mathcal{S}^g} \alpha_i \approx \frac{1}{m} \sum_{i=1}^{m} \alpha_i$ ($\mathcal{S}^g$ is training data set for good teachers)

As a result, we can add the following constraint to the Dual problem in Eq. (2) as the following:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathcal{R}^m} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2\lambda} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ s.t \quad & 0 \leq \alpha_i \leq \frac{1}{m} \, \forall i = 1, 2, \ldots m \\ & \frac{1}{|\mathcal{S}_t|} \sum_{\mathbf{x}_i \in \mathcal{S}_t} \alpha_i \leq \frac{1}{m} \sum_{i=1}^{m} \alpha_i + \frac{\epsilon}{m\sqrt{|\mathcal{S}_t|}} \, \forall t = 1, 2, \ldots k \end{aligned} \tag{4}$$

where $\mathcal{S}_t$ are the training data that belong to a specific teacher t, k is the number of teachers and ϵ is a tuning parameter [1].

2 Procedure

The procedure is of four steps: tuning, training, testing and evaluation.

2.1 Tuning

First of all, a grid search was performed for different values of λ and ϵ to maximize the accuracy. In particular, two sets are defined: $\boldsymbol{\lambda} = \{1, 1.5, \ldots 100\}$ and $\boldsymbol{\epsilon} = \{0.01, 0.02, \ldots 1\}$. After that, the SVM model suggested in Eq. (4) is trained and tested for each of the parametric combinations $(\lambda, \epsilon) \in \boldsymbol{\lambda} \times \boldsymbol{\epsilon}$ and then evaluated using the function 'rocplot'. Details for training, testing and evaluation steps are explained in Sections 2.2, 2.3 and 2.4.
Finally, the accuracy is calculated for each pair $(\lambda, \epsilon) \in \boldsymbol{\lambda} \times \boldsymbol{\epsilon}$ using the function 'rocplot', and the pair $(\lambda^*, \epsilon^*) \in \boldsymbol{\lambda} \times \boldsymbol{\epsilon}$ that maximizes the accuracy is selected. Additionally, the pair (λ^*, ϵ^*) is selected such that the corresponding classifier

is not leading to uniform decisions; i.e., the decision of the classifier is d_0 for all the testing samples (where $d_0 \in \{-1, +1\}$).
Next, the model corresponding to the pair of parameters (λ^*, ϵ^*) is evaluated using the function 'rocplot' as in following sections.

2.2 Training Phase

To solve the problem in Matlab using 'quadprog' command, the new Dual optimization problem in 4 needs to be rewritten in matrix format as in the following:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} + \boldsymbol{f}^T \boldsymbol{\alpha} \tag{5}$$

$$s.t. \quad \mathbf{A}\boldsymbol{\alpha} \leq \mathbf{b} \tag{6}$$

$$\mathbf{lb} \leq \boldsymbol{\alpha} \leq \mathbf{ub} \tag{7}$$

For which we $\mathbf{H}$, $\boldsymbol{f}$, $\mathbf{A}$, $\mathbf{b}$, $\mathbf{lb}$ and $\mathbf{ub}$ are defined as in the following:

- $\mathbf{H} = \frac{1}{\lambda}(\mathbf{Y}\mathbf{Y}^T) .* (\mathbf{X}\mathbf{X}^T)$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}_{m \times 1} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}_{m \times n}$$

  where $.*$ is the element by element multiplication of the two elements.

- $\boldsymbol{f}$, $\mathbf{lb}$ and $\mathbf{ub}$ are as follows

$$\boldsymbol{f} = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}_{m \times 1} \quad \mathbf{lb} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}_{m \times 1} \quad \mathbf{ub} = \begin{bmatrix} 1/m \\ \vdots \\ 1/m \end{bmatrix}_{m \times 1}$$

- Assuming the first $|\mathcal{S}_1$ samples are assigned to teacher 1, second $|mathcalS_2$ sampels are assigned to teacher 2 and so forth, $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{bmatrix} 1/|\mathcal{S}_1| & \dots & 1/|\mathcal{S}_1| & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & 1/|\mathcal{S}_2| & \dots & 1/|\mathcal{S}_2| & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & 1/|\mathcal{S}_k| & \dots & 1/|\mathcal{S}_k| \end{bmatrix}_{k \times m}$$

$$- \begin{bmatrix} 1/m & \dots & 1/m \\ \vdots & \ddots & \vdots \\ 1/m & \dots & 1/m \end{bmatrix}_{k \times m}$$

and

$$\mathbf{b} = - \begin{bmatrix} \frac{\epsilon}{m\sqrt{|S_1|}} \\ \vdots \\ \frac{\epsilon}{m\sqrt{|S_k|}} \end{bmatrix}_{k \times 1}$$

- Furthermore, if *unbiased* classifier design is desired, the following constraint is to be added

$$\mathbf{Q}\boldsymbol{\alpha} = 0 \tag{8}$$

where $\mathbf{Q} = [y_1, y_2, \ldots y_m]$

Please note that the trainig is perforemd on the provided randomly selected trainig sampels. The output of this step is the model defined by $\mathbf{w}^*$ and b^*, where $\mathbf{w}^*$ and b^* can be calculated as follows:

$$\begin{aligned} \mathbf{w}^* &= \sum_{i=1}^{m} \boldsymbol{\alpha}_i y_i \mathbf{x}_i \\ b^* &= \sum_{i=1}^{t} \mathbf{w}^* \cdot \mathbf{s}_i - y_i \end{aligned} \tag{9}$$

where $\mathbf{s}_i$ and t are respectively the support vectors and and the number of support vectors. Support vectors are selected based on the following rule:

$$\{\mathbf{s}_i\} = \{\mathbf{x}_i | \alpha_i \geq 0.5 \cdot \alpha_{max}\} \; \forall i = 1, 2, \ldots m \tag{10}$$

where $\alpha_{max} = \max \alpha_i \forall i = 1, 2, \ldots m$.

2.3 Testing Phase

The decision of the classifier (be it $\hat{c}_i$) is made on testing sample $\mathbf{x}_i$ based on the position of $\mathbf{x}_i$ with respect to the hyperplane $\mathbf{w}^* \cdot \mathbf{x}_i + b^* = 0$. Mathematically, this can be expressed as:

$$\hat{c}_i = \text{sign}(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) \tag{11}$$

2.4 Evaluation

Using the output of equation 11, a confusion matrix can be made. The confusion matrix is a standard method of showing the results of a classifier. The rows of the confusion matrix represents the actual labels of the data and the columns of the confusion matrix represents the labels that are predicted by the classifier (i.e., the decisions made by Eq. (11)). Table **??** shows an illustration of the confusion matrix. where TP, TN, FP and FN are the true positives, true negatives, false

Table 1: An Illustration of the Confusion Matrix

		Classifier Decision	
		$+1$	-1
Actual	$+1$	TP	FN
	-1	FP	TN

positives, and false negatives; respectively. From the confusion matrix, the true positive rate (TPR) and the false positive rate (FPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

One visualization method of the performance of the SVM classifier is the *Receiver Operating Characteristic* (ROC) curve. The ROC curve can be obtained in the following method. First, a threshold γ^0 is set such that:

$$\begin{aligned} \hat{c}_i &= +1 \text{ if } \mathbf{w}^* \cdot \mathbf{x}_i + b^* \geq \gamma^0 \\ \hat{c}_i &= -1 \text{ if } \mathbf{w}^* \cdot \mathbf{x}_i + b^* < \gamma^0 \ \forall i \end{aligned} \tag{14}$$

Then the TPR^0 and the FPR^0 are calculated. After that, the threshold γ^0 is changed to γ^1, and TPR^1 and FPR^1 are calculated. The ROC curve is then the plot of FPR vs TPR. A straight line ROC curve indicates that the classifier is making random decisions. The ideal ROC curve reaches $TPR = 1$ and $FPR = 0$. The accuracy is defined to be the area under the curve.

3 Results

$\lambda^* = 2.5, \epsilon^* = 0.94$
Accuracy = 68.39%
Note: we needed to flip the decisions

$$\text{Confusion Matrix} = \begin{bmatrix} 47 & 60 \\ 13 & 37 \end{bmatrix}$$

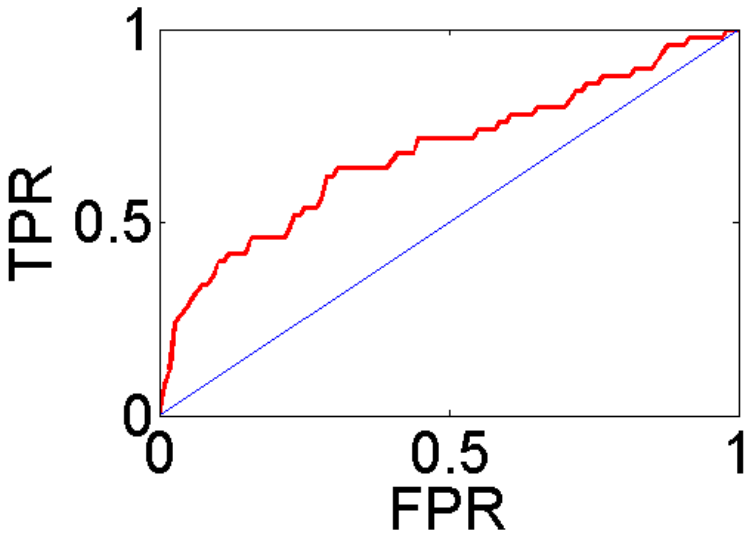

Figure 1: ROC Curve

References

[1] O. Dekel and O. Shamir, "Good learners for evil teachers," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 233–240.