Practical Conditions for Effectiveness of the Universum Learning

Sauptik Dhar (dharx007@umn.edu), Student ID: 3909558

INTRODUCTION

Sparse high-dimensional data is common in modern machine learning applications. In micro-array data analysis, technologies have been designed to measure the gene expression levels of tens of thousands of genes in a single experiment. However, the sample size in each data set is typically small ranging from tens to low hundreds due to the high cost of measurements. Most approaches to learning with high-dimensional data focus on improving existing inductive methods that try to incorporate a priori knowledge about the optimal model. One such approach is the non-standard learning setting known as Learning through Contradiction, or learning in the Universum environment[1]. For example, consider the medical diagnosis of 'hypo-thyroid' disease ; where the goal is to estimate a binary decision rule for discriminating between 'hypo-thyroid' and 'normal' patients based on their demographic characteristics (age/sex), clinical test results etc. This decision rule is estimated using past medical records of correctly diagnosed patients, i.e., labeled training set. Suppose that we also have available data for the 'hyper-thyroid' patients. Although these 'hyper-thyroid' patients cannot be assigned to any of the two classes ('hypo-thyroid' or 'normal'), they contain certain information about the thyroid disease. This additional information can be used to extract better decision rules for diagnosis (classification) of 'hypo-thyroid' disease. Note that the a priori information about learning is provided through data samples which are known *not* to belong to either one of the classes. Such samples are called *universum* samples as these samples contradict the labeled samples and the learning is called *learning through contradiction*. Such non-standard learning settings reflect properties of real-life applications, and can result in improved generalization, relative to standard inductive learning. However, these new methodologies are more complex, and their advantages and limitations are not well understood. This project aims at understanding the conditions when the Universum setting is suggested over the standard inductive setting for the SVM based learning formulations.

UNIVARIATE HISTOGRAMS OF PROJECTIONS

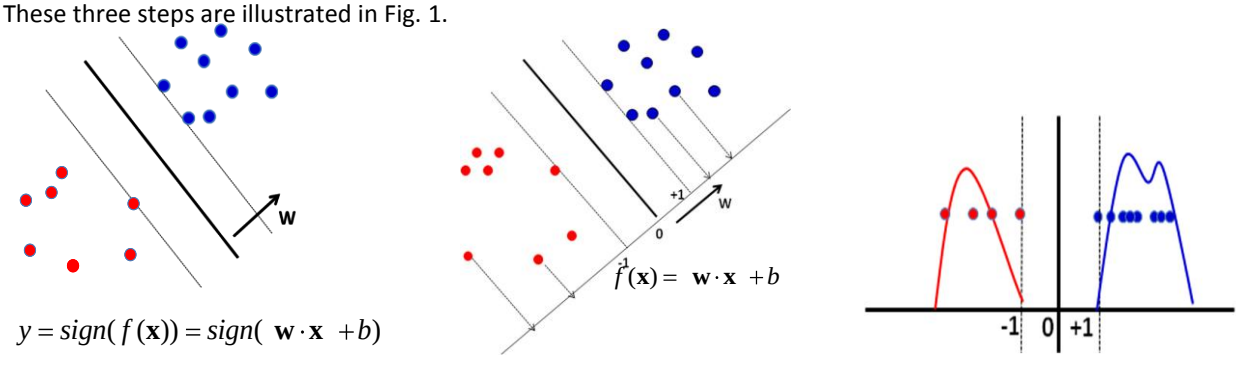
In order to have a better understanding of the SVM based models we need a simple visualization technique which helps us to understand how the SVM based model visualizes the data. For this we present a very simple visualization called the histogram of projections.

Univariate Histogram of Projections *~ is the histogram of the projection values of the data samples onto the normal weight vector of the SVM decision boundary.*

This method has been proved useful for a variety of problems for SVM based models[2].The idea itself is very simple, and it has been widely used, under different contexts, in statistics and machine learning. Such a histogram is obtained via the following three steps:-

a. Estimate standard SVM classifier for a given (labeled) training data set. Note that this step includes optimal model selection, i.e. tuning of SVM parameters (regularization parameter, kernel).
b. Generate low-dimensional representation of training data by projecting it onto the normal direction vector of the SVM hyperplane estimated in (a).
c. Generate the histogram of the projected values obtained in (b).

These three steps are illustrated in Fig. 1.



(a) (b) (c)

Fig.1. Illustration of the steps to generate the univariate histogram of projections.(a) The estimated SVM model and training data.(b) Projection of the training data onto the normal weight vector (**w**) of the SVM hyper plane.(c) Univariate histogram of projections. i.e. histogram of f(x).

UNIVERSUM LEARNING

The SVM based formulation for the Universum learning setting is provided by Vapnik[1] and the optimization formulation/ implementation is discussed in[3]. I have implemented the MATLAB interface and it shall be provided. Here we provide the Universum SVM (U-SVM) formulation in the primal form (1). For the Universum samples $(\mathbf{x}_j^*)$, we need to penalize the real-valued outputs of our classifier that are 'large'. This is accomplished using ε – insensitive loss (as in standard support vector regression). Let ξ_j^* denote slack variables for samples from the Universum. Then the Universum SVM formulation can be stated as:

$$\min_{\mathbf{w},b} R(\mathbf{w},b) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{w}) + C\sum_{i=1}^{n}\xi_i + C^*\sum_{j=1}^{m}\xi_j^* \quad (1)$$

subject to constraints

for labeled data: $y_i[(\mathbf{w}\bullet\mathbf{x}_i)+b] \geq 1-\xi_i \qquad \xi_i, \geq 0, i=1,\ldots,n$

for Universum data: $\left|(\mathbf{w}\cdot\mathbf{x}_j^*)+b\right| \leq \varepsilon+\xi_j^* \qquad \xi_j^* \geq 0,\ j=1,\ldots,m \qquad$ where $C, C^* \geq 0; \varepsilon \geq 0$

Parameters C and C^* control the trade-off between minimization of errors and the maximization of the number of contradictions. Selecting 'good' values for these parameters constitutes model selection (usually performed via resampling). When C^*=0, this U-SVM formulation is reduced to standard soft-margin SVM. Based on our findings the practical conditions for the effectiveness of universum learning is provided below[4],

PRACTICAL CONDITIONS

(C1) The histogram of projection of training samples lies outside the soft-margins. (i.e. the training samples are well separable)

(C2) The projection of Universum samples is symmetric relative to the (standard) SVM decision boundary, and

(C3) The projection of Universum samples has wide distribution between SVM margin borders denoted as points -1/+1 in the projection space.

We provide some results to confirm our point (more results are available on reference). For this we use 2 different datasets

- *Real-life MNIST handwritten digit data set*, where data samples represent the handwritten digits 5 and 8. Each sample is represented as a real-valued vector of size 28*28=784. On average, approximately 22% of the input features are non-zero which makes this data very sparse. The training set size is 1,000, validation set size is 1,000, and test set size is 1,866 samples. For U-SVM, 1,000 Universum samples are the digits '1' and '3'.
- *Real-life ABCDETC data set,* where data samples represent the handwritten lower case letters 'a' and 'b'. Each sample is represented as a real-valued vector of size 100*100=10000. The training set size is 150 (75 per class), validation set size is 150 (75 per class). The remaining 209 samples are used as test samples (105 from class 'a' and 104 from class 'b'). For U-SVM, 1,500 Universum samples '*All digits from 0 to 9*' and *'Random Averaging'*.

For both SVM and U-SVM the classifier is estimated using the training data, its model complexity is optimally tuned using validation data and finally the test error is estimated using test data. The experiment is repeated 10 times, using different random realizations of the data, and the average test error rates are reported for comparison. For the MNIST data we use a RBF kernel of the form $K(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)$ and for the ABCDETC data we use the polynomial kernel $K(\mathbf{x},\mathbf{x}') = (\mathbf{x}\bullet\mathbf{x}')^d$ of degree d=3 following[3]. The range of parameters used during model selection: C~[10^{-11},10^{-9},…,1,100], C/C*~[10^{-4} ,3x10^{-4},…,3x10^{-1},1,3] and ε= [0, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4]. The results for the comparison of SVM with U-SVM are provided in Table.1. From the results (in Table 1 and Fig. 1) for MNIST data we observe that the U-SVM outperforms SVM specifically for the digits '3'. For digit '1' we can observe that the universum samples are not symmetric about the decision boundary and hence do not satisfy our condition (C2). A similar analysis can be seen for the ABCDETC data (Table 1 and Fig. 2) where we see that the RA universum samples have a smaller distribution in comparison to the other two universum types and hence based on our condition (C3) is outperformed by the "all digits" universum samples.

MNIST DATA	SVM	U-SVM (digit 1)	U-SVM(digit 3)
Test error	1.47% (0.32%)	1.31% (0.31%)	1.01% (0.28%)
ABCDETC DATA	**SVM**	**U-SVM (all digits)**	**U-SVM (RA)**
Test error	20.47 %(2.60%)	18.37 % (3.47%)	18.85 % (2.81%)

Table.1 Comparison of average test error (with standard deviation in parenthesis) for different methods.

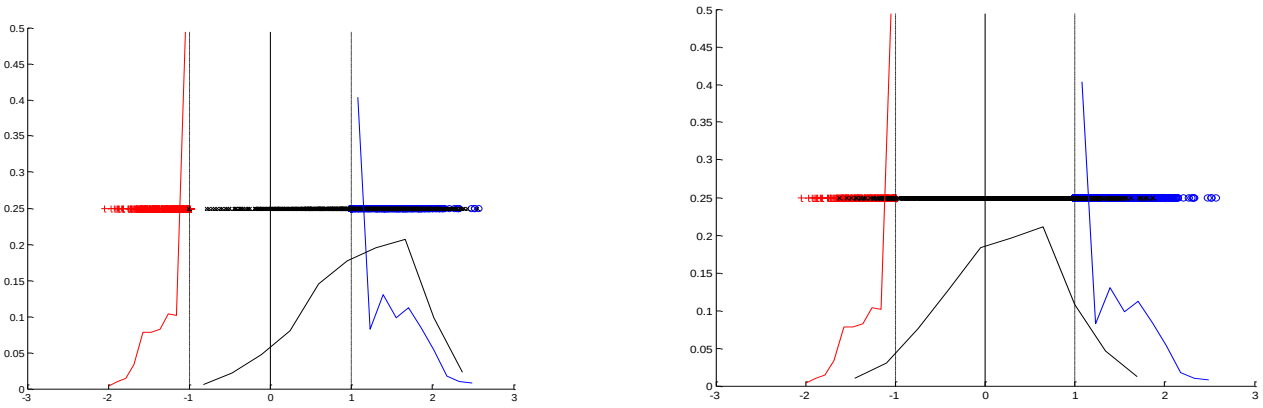


Fig. 2. Univariate histogram of projections for 3 different types of Universa. Training set size ~ 1,000 samples. Universum set size ~ 1,000 samples. (a) digit 1 Universum. (b) digit 3 Universum.

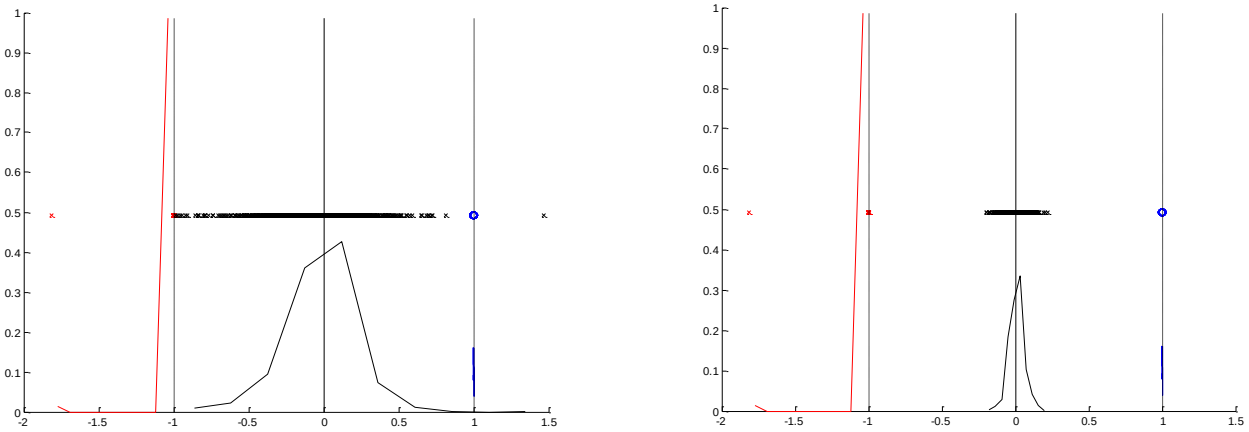


Fig. 3. Univariate histogram of projections for 3 different types of Universa for ABCDETC data Training set size ~ 150 samples. Universum set size ~ 1,500 samples. (a) 'digits 0-9' Universum. (b) RA Universum.

CONCLUSION

Thus now we have some simple conditions based on a simple graphical tool; to say when it may be useful to apply U-SVM over SVM. These simple conditions are really helpful for people who would like to apply U-SVM without having a much deeper understanding of the underlying optimization. These conditions have been derived based on some mathematical derivations which have not been shown in this report owing to space constraints.

REFERENCE

[1] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data. Empirical Inference Science:* Afterword of 2006. New York: Springer, 2006.

[2] V.Cherkassky, S.Dhar, "Simple Method for Interpretation of High-Dimensional Nonlinear SVM Classification Models", DMIN, July 2010.

[3] J. Weston, R. Collobert, F. Sinz, L. Bottou, and V. Vapnik, "Inference with the Universum," *Proc. ICML*, 2006, pp. 1009-1016.

[4] V. Cherkassky, S. Dhar and W. Dai, "Practical Conditions for Effectiveness of the Universum Learning," IEEE Trans. on Neural Networks, Aug 2010, submitted for publication.