COMP0120: Numerical Optimisation Project Report

Jiahao Li 19138867

May 6, 2020

1 Introduction

Support Vector Machine (SVM) is a classical methods on classification problem in machine learning and there are lots of well established theory on this field. Particularly, SVM is formulated as a constraint optimization problem, and it is a good test bench for assessing the performance of different kinds of optimization algorithm. In this report, the details of the mathematical derivation and how to formulate a SVM and a semi-supervised variants of SVM, One-Class classification SVM, as constraint optimization problems will be discussed. The constraint optimization problems are developed with linear and non-linear kernel tricks. Two optimization algorithms, Quasi-Newton method (BFGS) and Sequential Minimal Optimization (SMO), are used to find out the optimal solution of the formulated problem. At the experiment section, the developed SVM models are tested under a imbalanced dataset. Then the convergence of the optimization algorithms for the developed SVM models applied on the imbalanced dataset will be discussed.

2 Mathematical Derivation

Support Vector Machine is to find a hyper-plane which can separate the different class of data. Figure 1 illustrate the geometry relations of SVM where W is the normal vector of the Hyper-plane, the distance between point x and the hyper-plane is the projection of $X - X'$ orthogonal to the hyper-plane and the points, X', lying on the hyper-plane has an equality $WX' + b = 0$. W is the calculated weights of the developed SVM and b is the bias term of the models. The distance equation is expressed as

$$distance(x, b, w) = |\frac{W^T}{||W||}(X - X^{'})| = \frac{1}{||W||}|W^T X + b| = \frac{1}{||W||}y(W^T X + b)$$

where y is the corresponding label of data point X. In SVM, we are trying to find out variable W and b which can describe the feeding data as

$$Wx_i + b \geq +1 \text{ for } y_i = +1$$
$$Wx_i + b \leq +1 \text{ for } y_i = \text{-}1$$

From [1], the two equations inequality can be unified as $y_i(x_i w + b) - 1 \geq 0 \; \forall i$. Since the distance between the selected supported vector and the hyper-plane is trying to be maximized and the minimum of $y_i(x_i w + b)$ is 1, the constraint optimization problem is formulated as

$$max \frac{1}{||W||}$$
$$\text{s.t. } \mathrm{y}_i(x_i w + b) - 1 \geq 0$$

which is equivalent to

$$min \frac{1}{2}||W||^2$$
$$\text{s.t. } \mathrm{y}_i(x_i w + b) - 1 \geq 0$$

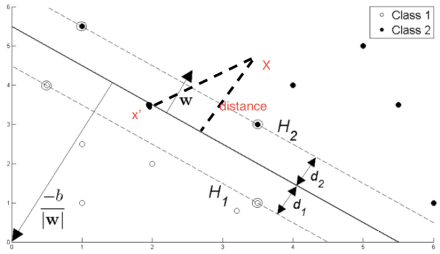

Figure 1: Hyper-Plane Separating a Binary Class from [1]

When it comes to the differences between hard margin and soft margin SVM, hard margin SVM forces every data points lying on the correct sides pf its own class. However, soft margin SVM gives tolerance a few data points which can show up at the sides belonging to others class by introduce a slack variable ξ_i. When the dataset can is perfectly separable, hard margin SVM is preferred. But, for most of the real world dataset, they cannot be perfectly sparable. Thus, soft margin is a good choices with controlling a free parameter C for the trade-off between accuracy and the size of margin. In this report, the mathematical derivation of soft margin SVM is discussed. By introducing a positive slack variable to the constraint optimization problem, the optimization problem is formulated as

$$min\frac{1}{2}||W||^2 + C\sum_{i=1}^{L}\xi_i$$
$$\text{s.t. } \mathrm{y}_i(x_i w + b) - 1 + \xi_i \geq 0,\ \xi_i \geq 0$$

Thus, the Lagrangian, primal equation, of the constraint optimization problem is

$$L_p = \frac{1}{2}||w||^2 + C\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}a_i(y_i(x_i w + b) - 1 + \xi_i) - \sum_{i}^{L}\lambda_i\xi_i$$

where x_i is the data to be study, y_i is the label of corresponding data, the positive parameter, a_i, is the Lagrangian multiplier of $y_i(x_i w + b) - 1 + \xi_i$ and positive parameter, λ_i, is the Lagrangian parameter of ξ_i. In order to find its corresponding dual equation, the primal equation is taken derivative with respect to w, b and ξ_i. See below

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{L}a_i y_i x_i \rightarrow w = \sum_{i=1}^{L}a_i y_i x_i,\ \frac{\partial L_p}{\partial b} = \sum_{i=1}^{L}a_i y_i = 0,\ \frac{\partial L_p}{\partial \xi} = C - a_i - \lambda_i = 0 \rightarrow C = a_i + \lambda_i$$

The dual problem is obtained by subsituting the equality above into the primal equation. Then, the dual problem is

$$\begin{aligned}
L_D &= \frac{1}{2}||a_i y_i x_i||^2 + (a_i + \lambda_i)\sum_{i=1}^{L}\xi_i - \sum_{i=1}^{L}a_i(y_i(x_i w + b) - 1 + \xi_i) - \sum_{i=1}^{L}\lambda_i\xi_i \\
&= \frac{1}{2}||a_i y_i x_i||^2 - \sum_{i=1}^{L}a_i(y_i(x_i w + b) - 1) \\
&= \frac{1}{2}||a_i y_i x_i||^2 + \sum_{i=1}^{L}a_i - \sum_{i=1}^{L}a_i y_i x_i w - \sum_{i=1}^{L}a_i y_i b \\
&= -\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}a_i a_j y_i y_j x_i x_j + \sum_{i=1}^{L}a_i
\end{aligned}$$

Since the dual provides a lower bound for the primal equation, the maximum of dual equation is to be found at this constraint problem. In order to adapt this problem into gradient descent style algorithm, the new constraint optimization problem is formulated as

$$min \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} a_i a_j y_i y_j x_i x_j - \sum_{i=1}^{L} a_i$$
$$\text{s.t. } -a_i \leq 0,\ a_i - C \leq 0,\ \sum_{i=1}^{L} y_i a_i = 0$$

The prediction of a data point, x_i, can be obtained by $\hat{y}_i = sign(wx_i + b) = sign(\sum_{n=1}^{L} a_n y_n x_n x_i + b)$. And $b = \frac{1}{L}(\sum_{i=1}^{L} y_i - wx_i) = \frac{1}{L}(\sum_{i=1}^{L} y_i - \sum_{n=1}^{L} a_n y_n x_n x_i)$. And C is controlling the the extent of margin violation. Large C want a less margin violation, i.e. every data point is supposed to be at its corresponding sides of Hyper-plane. And small C allows a large margin, which indicates that more data points would be at an incorrect sides of the Hyper-plane. Interestingly, the soft margin problem can be transformed into hard margin problem by taking C as 0.

Additionally, a semi-supervised variant of SVM, One-Class SVM, is developed to study the anomaly detection problems, outlier, which can be set up by using an imbalanced binary classification dataset. In the one-class SVM, one decision is make, which is determine if a data point belongs to the class being assessed or not. Therefore, the label of data i, y_i, is not necessary anymore in the objectives function. Based on [2], the one-class svm, with uniform expression as soft-margin svm derivation above, is expressed as

$$min \frac{1}{2}||W||^2 - b + C \sum_{i=1}^{L} \xi_i$$
$$\text{s.t. } \mathrm{x}_i w - b + \xi_i \geq 0,\ \xi_i \geq 0$$

where b is the learned bias term, $C = \frac{1}{Lv}$ and v, free parameter, is a proposed upper bound of the outlier fraction at the testing data. [2] The primal equation and first derivative are

$$L_p = \frac{1}{2}||W||^2 - b + C \sum_{i=1}^{L} \xi_i - \sum_{i=1}^{L} a_i (wx_i - b + \xi_i) - \sum_{i=1}^{L} \lambda_i \xi_i$$
$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1} a_i x_i = 0 \rightarrow w = \sum_{i=1} a_i x_i,\ \frac{\partial L_p}{\partial b} = (\sum_{i=1}^{L} a_i) - 1 = 0,\ \frac{\partial L_p}{\partial \xi_i} = C - a_i - \lambda_i = 0$$

The dual equation can be obtained by substitute the zero first derivatives into the primal equation.

$$L_D = \frac{1}{2}||\sum_{i=1}^{L} a_i x_i||^2 - b + (a_i + \xi_i) \sum_{i=1}^{L} \xi_i - \sum_{i=1}^{L} a_i (wx_i - b + \xi_i) - \sum_{i=1}^{L} \lambda_i \xi_i$$

With using similar mathematical derivation as the soft-margin SVM, the optimization problem of one-class SVM is as following

$$min \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} a_i a_j x_i x_j$$
$$\text{s.t. } 0 \leq a_i \leq C,\ (\sum_{i=1}^{L} a_i) - 1 = 0$$

The prediction of one-class svm can be obtained by $\hat{y}_i = sign(wx_i - b) = sign(\sum_{n=1}^{L} a_n x_n x_i - b)$ and $b = \frac{1}{L} \sum_{i=1}^{L} wx_i = \frac{1}{L} \sum_{i=1}^{L} \sum_{n=1}^{L} a_n x_n x_i$.

Although soft-margin can help to classify those datasets which care perfectly separable, kernel tricks can be used to greatly improve the classification accuracy. Kernel tricks is achieved by using non-linear feature mapping $x \rightarrow \psi(x)$. In the mathematical derivation above, $x_i x_j$ can be regarded as a linear kernel where $\psi(x_i)\psi(x_j) = x_i x_j$. In this project, the nonlinear kernel, RBF kernel, is used to learn a non-linear hyper-plane for classification problem. And RBF kernel is expressed as $\psi(x_i)\psi x_j = \exp -\frac{||x_i - x_j||^2}{2\sigma^2}$. σ is the spread of RBF kernel, and the selection of σ can greatly affect the performance of models. Evangelista et.al [3] formulate the selection of σ as an optimization problem, but it is not the scope of this project. The suggested parameter from this paper is used. Kernel tricks can be embedded into the constraint optimization above by simply change x_i to $\psi(x_i)$ and change the dot product of two data points to kernel tricks expression.

3 Problem Formulation and Optimization

The SVM optimization problem is to find a set of Lagrangian Multiplier $A = a_1, a_2, ..,_{a\ m}$ to maximize the objectives function. Quadratic Penalty methods is used to transform the constraint optimization problem into an unconstraint optimization problem. The constraints in the original constraint optimization problem are added as a term at the reformulated unconstraint problem with increasing penalty parameter μ. And the problem formulation of soft margin SVM and one-Class SVM is as following

$$Q(a_i, \mu) = \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} a_i a_j y_i y_j x_i x_j - \sum_{i=1}^{L} a_i + \frac{\mu}{2}((\sum_{i=1}^{L} y_i a_i)^2 + ([-a_i]^+)^2 + ([a_i - C]^+)^2)$$

$$Q(a_i, \mu) = \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} a_i a_j x_i x_j + \frac{\mu}{2}(((\sum_{i=1}^{L} a_i) - 1)^2 + \sum_{i=1}^{L}([-a_i]^+)^2 + \sum_{i=1}^{L}([a_i - C]^+)^2)$$

where $[y]^+ = max\{y, 0\}$ and y is increased by 1.5 times at every iteration. Fromm Nocedal Wright [4], large μ is always a bad approximation of the quadratic penalty function. The upper bound of the penalty parameter μ is empirically set as 100. Then, the quadratic penalty equation can be solved by any unconstraint optimization algorithm.

Since it is costly to compute an exact Hessian matrix for big data set or high dimensional data. In this project, quasi-newton methods (BFGS), which approximates a Hessian matrix for the objectives function, is used to optimized the Quadratic Penalty objectives function. For Quasi-newton methods, BFGS algorithm is used to directly approximate the inverse of Hessian with using first order gradient information and Sherman-Morrison-Woodbury formula.[4]

Sequential Minimal Optimization (SMO) proposed by John Platt [5], is an efficient optimization algorithm to solve quadratic programming problem. SMO is to find out the maximum values of the dual problem of SVM optimization problem

$$maxL(a_1, .., a_L) = max - \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} a_i a_j y_i y_j x_i x_j + \sum_{i=1}^{L} a_i$$

$$\text{s.t. } -a_i \leq 0,\ a_i - C \leq 0,\ \sum_{i=1}^{L} y_i a_i = 0$$

The basic idea of SMO is to simultaneously do gradient ascent with respect to two of Lagrangian Multipliers iteratively. In every iteration of update, two Lagrangian multipliers are selected based on some heuristic and keep the others Lagrangian Multipliers fixed. With using the equality constraint in SVM and chosing a_i and

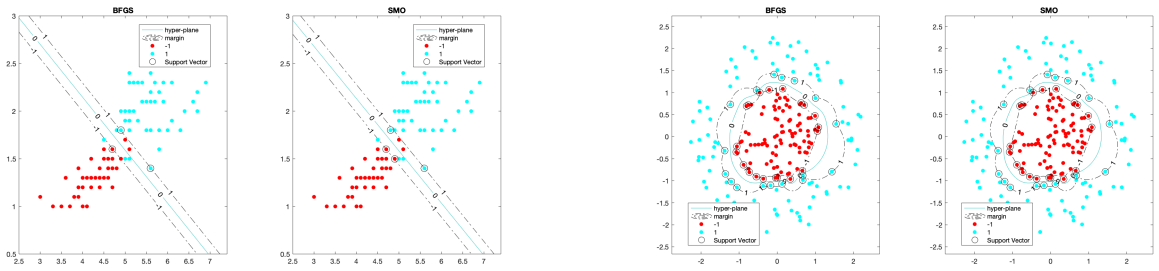

Figure 2: Left: Linear Kernel Right: RBF kernel

a_2 as update variable, referred from [6] [7], the binary classification SVM update can be

$$\sum_{i=1}^{L} = a_i y_i = 0$$

$$a_1 y_1 + a_2 y_2 = -\sum_{i=3}^{L} a_i y_i$$

$$a_1 y_1 + a_2 y_2 = K$$

$$a_1 = (K - a_2 y_2) y_1$$

where K is a constant and $y_i^2 = 1$ in binary classification problem. Then, the above equality can be substituted back to the dual equation optimization problem, and the multipliers, a_1 and a_2, can be solved analytically. The tolerance parameter is set as 0.001 suggested by Ng [6], which is used to set a stop condition as well as check the convergence conditions. The convergence Quadratic Penalty method and SMO algorithms is discussed in the following section using real-world dataset.

4 Experiment

This project is developed under Matlab envirnment and reusing the optimization codes developed in the previous assignments. And the experiment of this project is composed of two stages. In the first stages, the experiments from the Matlab Support Vector Machine documentation[1] are re-implemented to visualize the validity of the developed model. Then, the developed model is tested on Ionosphere Data Set from UCI machine learning repository. The Ionosphere Data Set records the radar data and labels the recorded data as Good or Bad. This is an imbalanced dataset, which contains 225 good radar data and 126 bad radar data.

4.1 Binary Classification SVM

The experiment of linear and non-linear, RBF kernel, with variance 1, soft margin SVM at Matlab documentation are re-implemented using both quadratic penalty and SMO methods. The penalty parameter C of soft-margin is empirically set as 100. The results is of soft-margin SVM with linear kernel and RBF kernel is at Figure 2.

For linear SVM, the experiment uses the first 100 rows of Iris dataset, which is built-in at the Matlab Machine Learning Toolbox. And, for nonlinear SVM, nonlinear separable synthetic data is generted manually. Based on the experiment results above, the developed model with linear kernel and RBF kernel can correctly address the binary classification problem.

For the experiment on the Ionosphere dataset, the good records and bad records are splited into two folds

[1] https://uk.mathworks.com/help/stats/fitcsvm.htmlbt7oo83-5

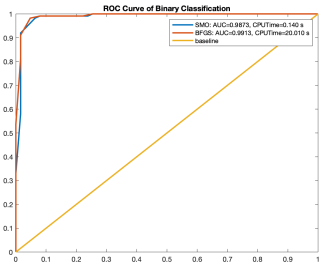

Figure 3: Ionosphere: Binary Classification

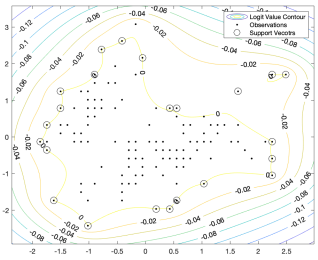

Figure 4: One-Class SVM for Anomaly Detection

evenly as training set and test set. And RBF kernel is used in the binary classification task. The ROC curve for the testing data and CPU time of training model is in Figure 3. The convergence analysis plot for BFGS and SMO is in Figure 6. Based on the results, the performance of BFGS and SMO are indifferent in the performance of classification problem. The convergence time of SMO, 0.14 seconds, is much faster than BFGS methods, 20.01 seconds. But BFGS requires less optimization iterations. Since the stop condition of the two algorithms are different, the stop criterion are set empirically to optimize the performance and not comparable. The convergence analysis of the two algorithms are detailed at next section.

4.2 One-Class SVM

Anomalies or outlier are the instances in the dataset which are deviate from the majority data. Although unsupervised nearest neighbors methods achieve the best performance in anomaly detection, Amer et.al [8] argues that SVM methods can achieve comparable results as the nearest neighbor methods, computationally efficient and effectly address the 'curse of dimensionality' problem in the nearest neighbor. In this project, a semi-supervised variants of SVM proposed by Mennatallab et.al [2] is developed address the anomaly detection in a real-world dataset. The experiment of outlier detection at the Matlab documentation is also re-implemented. The experiment uses RBF kernel with outlier fraction 0.05. The experiment results is at Figure 4. The outliers can be classified from the outsides of contour values 0. The developed model can achive similar results as the experiment at the matlab documentation.

One-class SVM, therefore, can help to address the classification in a imbalanced dataset. The training and testing splits of Ionosphere data is same as binary classification SVM setting. The outlier fraction in the one-class SVM model is set as 0.5. The whole dataset is visualized using TSNE [9] at Fgure 5. The bad records have a different behaviour from the good records, thus are regarded as outliers in this problem setting. At the prediction stage, bad records is classified as class 1 and the good records is classified as class -1. The ROC curve recording the accuracy of identifying outlier data point is also in Figure 5. The optimized spread parameter, $\sigma = 1$, in [2] is used directly in this experiment. Based on the results, One-Class SVM can achieve meaningful results in classification problem. Based on the visualized plot, the bad records are not really deviated very much from the good records. Thus, One-Class SVM performance is still worse than

binary classification SVM. But One-Class SVM does not need labels of data during training.

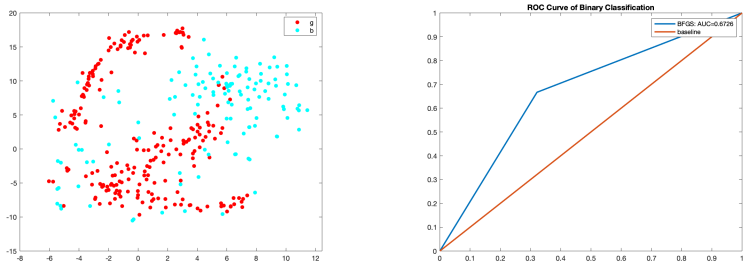

Figure 5: Left: Visulization of Dataset Right: ROC curve of Anomaly Detection

5 Discussion and Conclusion

In this section, the convergence of Quadratic Penalty method and SMO algorithm in the binary classification problem is discussed. Both of the two optimization methods are checked using Q-convergence in Nocedal & Wright Nocedal Wright [4] $\frac{||x_k - x^*||}{||x_{k-1} - x^*||}$. From [4], the convergence of quadratic penalty can be checked from $||\nabla Q(a_i, u)|| \leq \tau_k$ and $\tau_k \to 0$ as iteration k increasing. The convergence of SMO algorithms is to the KKT dual-complementary conditions in [6]

$$\begin{aligned} a_i = 0 &\to y_i(wx_i + b) \geq 1 \\ a_i = C &\to y_i(wx_i + b) \leq 1 \\ 0 \leq a_i \leq C &\to y_i(wx_i + b) = 1 \end{aligned}$$

For quadratic penalty methods, the norm of gradient for the objectives function is approaching to 0 as k increasing, which validate the convergence of quadratic penalty methods. And Q-convergence of quadratic penalty method is bounded below 1. From Q-Convergence in Nocedal & Wright [4], the developed quadratic penalty method has a linear convergence rate. For SMO algorithm, the number of data point satisfying the KKT dual-complementary conditions is counted. And the fraction of data points in the Ionosphere dataset satisfying the conditions are calculated and present at the convergence plot Figure 6. The tolerance of condition checking is set as 0.001. From the convergence plots, The fraction of data points satisfying the KKT dual-complementary conditions is increasing to 1 as k becoming larger, showing SMO achieving convergence. Also, the Q-convergence of SMO is bounded below 1, hence linear convergence rate in SMO algorithm can be observed.

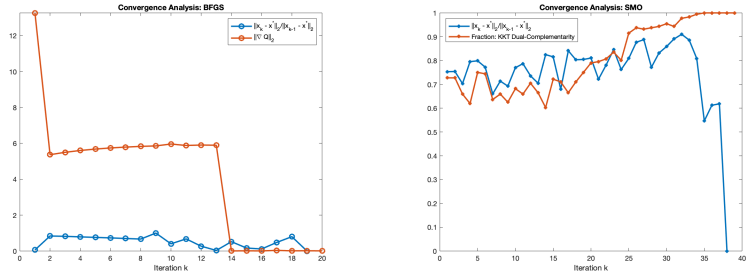

Figure 6: Convergence of BFGS and SMO of Ionosphere Binary Classification

In this project, two SVM models, soft-margin binary classification SVM and one-class SVM, are developed. Since SVM is a constraint optimization problem, quadratic penalty methods and Sequential minimal

optimization are used to solved the constraint optimization problem. Quadratic penalty method transform the constraint problem into an equivalent unconstraint optimization and solve the transformed by using Quasi-Newton method, BFGS, which approximate a Hessian for the objectives function for computational efficiency. And SMO algorithm update two Lagrangian Multiplier iteratively with gradient ascent to maximize the dual equation in SVM optimization. Kernel tricks is used to address those data which are not linearly separable. And two kernel, linear kernel and RBF kernel, are developed in this project. In conclusion, the developed SVM model and the optimization algorithms are working properly in both sythetic data and real world dataset.

References

[1] Tristan Fletcher. Support vector machines explained, 2008.

[2] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 582–588, Cambridge, MA, USA, 1999. MIT Press.

[3] Paul F. Evangelista, Mark J. Embrechts, and Boleslaw K. Szymanski. Some properties of the gaussian kernel for one class learning. In *Proceedings of the 17th International Conference on Artificial Neural Networks*, ICANN'07, page 269–278, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, second edition, 2006.

[5] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, April 1998.

[6] Andrew Ng. Cs229 lecture notes , support vector machines, 2018.

[7] Cs229 simplified smo algorithm. http://cs229.stanford.edu/materials/smo.pdf. Accessed: 2020-05-03.

[8] Mennatallah Amer, Markus Goldstein, and Slim Abdennadher. Enhancing one-class support vector machines for unsupervised anomaly detection. In *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description*, ODD '13, page 8–15, New York, NY, USA, 2013. Association for Computing Machinery.

[9] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.