Assignment 1: Predict diabetes via Perceptron

Moaz Mohamed
The University of Adelaide
Adelaide SA 5005
a1779177@student.adelaide.edu.au

1. Introduction

Over the last few years, deep learning had a significant impact in the medical field. The application of numerous Deep learning algorithms have been implemented in different sector of the medical field. In this paper Perceptron algorithm will be explored and its performance will be investigated on classification task to predict diabetes on the provided dataset.In addition to a complete review of the Perceptron algorithm describing its strength and weakness. An alternative algorithm will be introduced and explored that can mitigate some of the inherit weakness of Perceptron.

2. Methodology and Background

2.1. Support Vector Machines

An alternative approach to pattern recognition is using Support Vector Machines(SVM). Which in hindsight is very similar to the perceptron algorithm but SVM ensures finding a hyperplane that can separate between the two classes with maximum margin. the decision for either positive or negative classification follows the eq 1.

$$y_i\left(wx_i+b\right)\geq 1 \tag{1}$$

for the data points that lie exactly on the support vector boundary.

$$y_i\left(wx_i+b\right)-1=0 \tag{2}$$

finding the width for the maximum margin will depend on finding the distance between the negative and positive points that exist on the boundary $(x_p - x_n)$ multiplying by $\frac{\bar{w}}{\|w\|}$ as unit vector provides the width.

$$(x_p - x_n)\times\frac{\bar{w}}{\|w\|} \tag{3}$$

using eq 2 the term that need to be maximized comes naturally as $\frac{2}{\|w\|}$ or for minimization as $\frac{1}{2}\times\|w\|^2$

$$\begin{array}{c}\frac{1}{\|w\|}\times\left(x_pw - x_nw\right)\\ \frac{1}{\|w\|}\times\left([1-b]+[1+b]\right)\end{array} \tag{4}$$

$$\begin{array}{c}\frac{2}{\|w\|}\\ \frac{1}{2}\times\|w\|^2\end{array} \tag{5}$$

thus the margin is

$$\frac{1}{2}\times\|w\|^2 \tag{6}$$

which can be constructed as a function to minimize

$$j\left(w\right)=\frac{1}{2}\times\|w\|^2 \tag{7}$$

subject to

$$y_i\left(w^Tx_i+b\right)\geq 1,\forall j \tag{8}$$

Such formulation does work on finding the optimal hyperplane and it can be solved with optimization tool such as CVXOPT or MOSER. a slack variable can also be added to relax the SVM with C as regularization parameter.

minimization of

$$\frac{1}{2}\times\|w\|^2+c\times\frac{1}{n}\sum_{i=1}^{n}\xi_i \tag{9}$$

subject to

$$\begin{array}{r}y_i\left(w^Tx_i+b\right)\geq 1-\xi i\\ \xi_i\geq 0,\forall n\end{array} \tag{10}$$

Kuhn-Tucker theorem and Lagrange multipliers can be utilized to drive the previous problem from primal problem to dual problem.

$$L_D\left(\alpha\right)=\sum_{i}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jy_iy_j < X_i^TX_j > \tag{11}$$

subject to for (non-separable case)

$$0\leq\alpha_i\leq C/n,\forall i,\sum_{i=1}^{n}\alpha_iy_i=0 \tag{12}$$

or subject to for (separable case)

$$\alpha_i \geq 0, \forall i, \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{13}$$

and the $\overrightarrow{w}$ and the bias term can be calculated as follows.

$$w = \sum_{i=1}^{n} \alpha_i y_j x_i \tag{14}$$

$$b = \frac{1}{y_j} - w^T x_j \tag{15}$$

It is interesting to note that in equation 14 the weight vector can be formulated as summation of alpha, label and the support vectors. Hence the number of the support vectors is depended on α_i as some values of α_i will be zero. instead of in the primal problem where the $\overrightarrow{w}$ is obtained directly from solving the primal optimization problem. Such difference allows for the usage for something called the ”Kernel Trick”.

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_j, x_i) \tag{16}$$

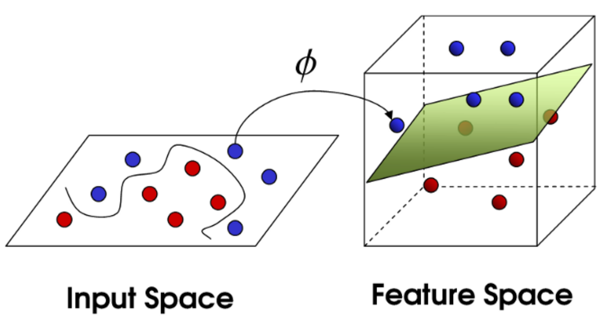

Figure 1: Mapping

$$K(x_j, x_i) = \left(x_i^T x_j + 1\right)^p \tag{17}$$

$$K(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2} \times \|x_j - x_i\|^2\right) \tag{18}$$

The optimization problems for both equations 16 and 11 except for the (x_j, x_i) in equation 16 is being applied to a kernel (Ex eq 17 and 18) and the data is being projected into higher dimensions. here the SVM is exploiting cover’s theorem.”A complex pattern-classification problem, cast in a high-dimensional space non-linearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated.” [4] Which increases the complexity of the SVM by applying the kernel and enabling it to classify classes even in non-separable cases as demonstrated in figure 1 and 4

3. Experimental Analysis.

3.1. SVM Primal and Dual

By utilizing the support vectors in eq 14 SVM is able to find the support vectors that can maximize the margin between the two classes as demonstrated in equation 11 but SVM without utilizing a kernel SVM has the same weakness as Perceptron in that regard. But when a kernel is utilized project the data into higher dimensions SVM is capable of finding a hyperplane between the two classes as demonstrated in figure 4.

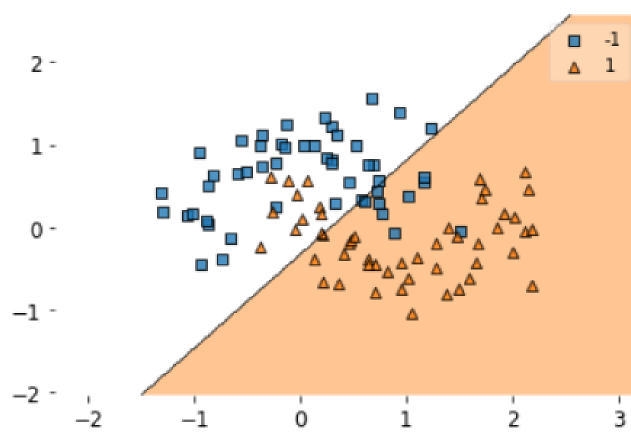

Figure 2: SVM decision boundary in non-linearly separable data

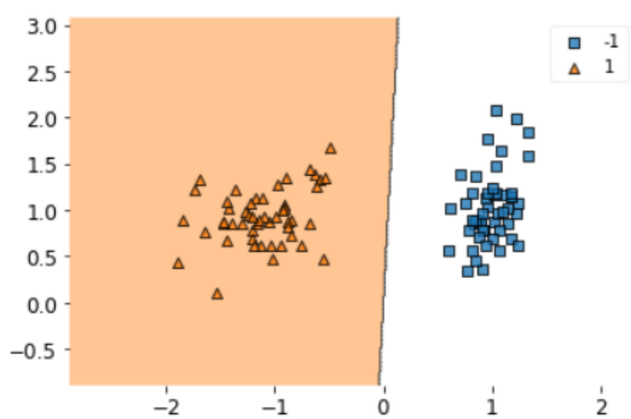

Figure 3: SVM decision boundary in linearly separable data

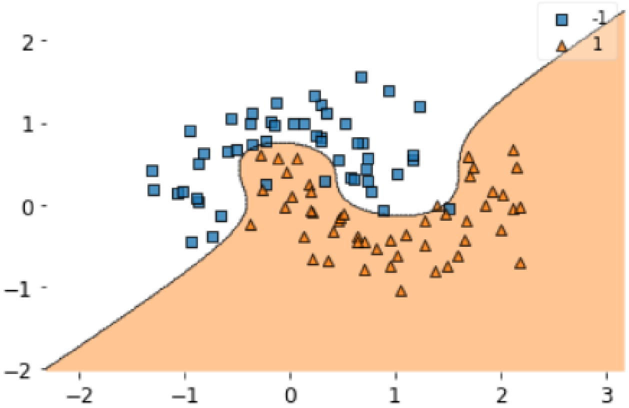

Figure 4: SVM with polynomial kernel applied.Decision boundary in non-linearly separable data

4. Performance on diabetes data-set

In table 1 are the results. with aggressive hyper-parameter tuning all algorithms did perform with no apparent close contender. the diabetes dataset is full of missing values and noise with careful data cleaning steps i expect better results.

Algorithm	W	bias	accuracy
Dual Soft	[0.7589323, -0.65157069]	0.29015581	59%
Primal Soft	[0.7588472, -0.65150936]	0.29023341	59%
sklearn	[0.7588628, -0.65153027]	0.29022996	59%

Table 1: SVM implementation for Soft margin

Algorithm	W	bias	accuracy
Dual Hard	[-1.754555, 0.0777976]	0.0037710	100%
Primal Hard	[-1.754516, 0.077783]	0.0038148	100%
sklearn	[-1.740545, 0.077163]	0.0117781	100%

Table 2: SVM implementation for Hard margin

Algorithm	training set	testing set	C
Dual Soft	97.74%	96.8%	1000
Primal Soft	97.71%	96.8%	1000
sklearn	97.75%	96.7%	1000

Table 3: SVM implementation for Soft margin on provided dataset

5. Conclusion

Perceptron and different variations of Support vector machines algorithm have been explored and analyzed. mathematical formulation of all the algorithms have been studied. decision boundaries for have been made for a careful study of the behavior of various algorithms in different data situations. Advantageous and weakness have been identified for all the explored algorithms. Perceptron and its creator have created and paved the way for the emergence of the field of machine learning. all subsequent algorithms builds on top of the weakness of the previous work as seen in perceptron, SVM and later kernalised SVM.

6. Code

all the code for this assignment is provided in this GitHub repo `https://github.com/RoastedKernel/uni_Machine_Learning`

References

[1] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[2] C. Murphy, P. Gray, and G. Stewart, "Verified perceptron convergence theorem," *MAPL 2017 - Proceedings of the 1st ACM SIGPLAN International Workshop on Machine Learning and Programming Languages, co-located with PLDI 2017*, pp. 43–50, 2017.

[3] M. Minsky and S. Papert, "Perceptrons: expanded edition," *MIT Press Cambridge MA*, vol. 522, p. 20, 1969.

[4] M. Cover, "Inequalities Applications Pattern," pp. 326–334, 1965.