MN50750 – OPTIMISATION AND SPREADSHEET MODELLING

Coursework 1: Building a Support Vector Machine using Excel

Name: Ritwik Singh

Word Count: 1970 Words

Contents

1 | INTRODUCTION

Considering recent events at the local GP, a noticeable surge in patients seeking assistance for heart-related issues has been observed, presenting a significant challenge. This increase not only raises concerns about the capacity of healthcare resources but also questions the timely delivery of care to those in critical need. Recognizing the urgency of the situation, a call is made for a system capable of providing swift, early, and accurate diagnoses, supporting healthcare professionals in prompt decision-making.

Dr. Swede Hart, the head doctor at the GP, has personally observed the challenges faced by patients, particularly concerning heart-related issues. A solution is contemplated by Dr. Hart—an accurate diagnostic model that could potentially save lives through early identification and timely treatment of diseases.

In the realm of medical diagnostics, the Support Vector Machine (SVM) can be likened to a detective. It is seen as an intelligent assistant, aiding doctors in predicting health outcomes, much like a crystal ball guides an astrologer. This report focuses on the creation of three specialized models using SVM, with the objective of identifying the most effective one to assist healthcare professionals.

The medical team at the GP, led by Dr. Hart, has identified four crucial factors for early heart disease checks: age, resting blood pressure, cholesterol level, and maximum heart rate. The SVM models will utilize this checklist, drawing insights from real data collected by the cardiology department over the past month. This approach allows the models to learn about the individual's physical condition and predict the likelihood of a heart condition.

The report unfolds by elucidating the workings of all three models in the subsequent section. Following that, the intention is to share intriguing discoveries from testing the models in the third section. Section 4 delves into a detailed discussion on why the models exhibit varied behaviour with different data. Finally, Section 5 guides in drawing conclusions and determining the most suitable model for the task.

2 | MODEL GENERATION

Fig. 1: Flowchart for the Diagnosis

Fig. 1 depicts the need to convert random data into information, the four factors associated with the risk of a heart issue need to be converted and put into a mathematical formula. This would help the computers to understand and do any required calculations. The data in multiple chunks needs to go through two phases to predict a future event. The model converts the data to information and classifies it to predict the occurrence of a forthcoming occurrence.

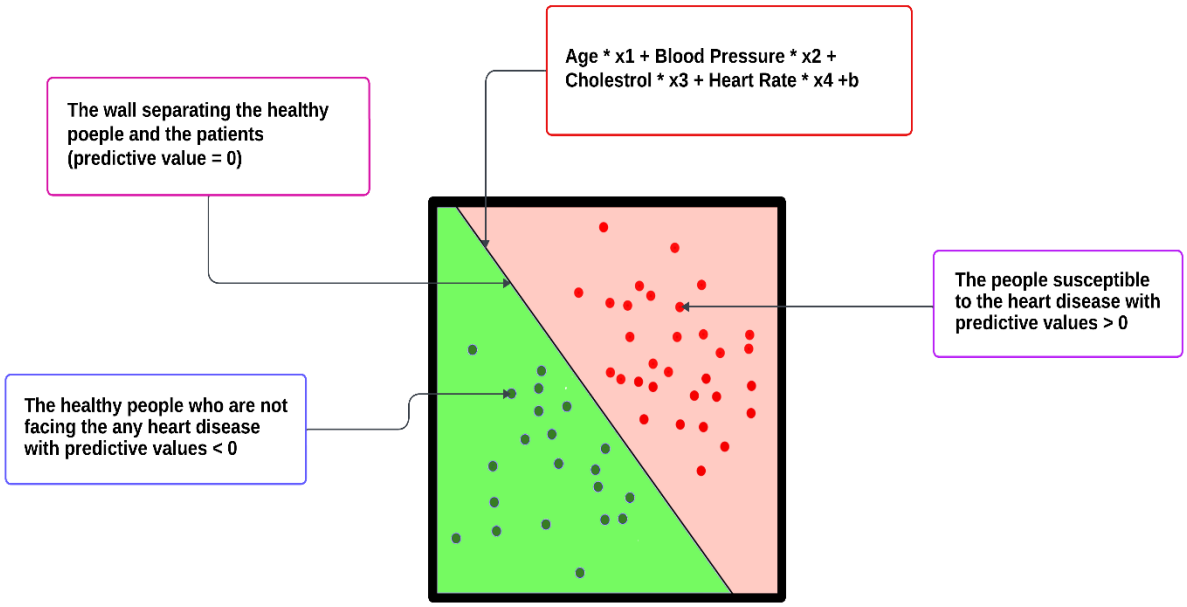


Fig. 2 The Generation Model of SVM

Fig. 2 shows how predicted values of each patient can be obtained (SVM1) by the multiplication of the four factors with x_1, x_2, x_3 and x_4 and then adding them to b. Think of a picture of a lush garden which is divided into two sections – one side having the flourishing plants (or the people who are not facing the cardiovascular disease with predictive values < 0), while the other side consisting of the plants which require special care and attention to thrive (or the people susceptible to the heart disease with predictive values > 0). The main motive of the models is to improve these parameters by training and making them more accurate and effective in understanding the needs of the plant in need.

2.1 | Improvements in the Model

Consider a healthy person, there may occur a similar value of the four factors for any of the individual. Hence, the introduction of λ (Lambda) in the second (SVM2) and third (SVM3) model. This represents the degree of misclassification which is acceptable. The ideal value of λ is calculated based on the performance of the model when tested with training and test data. In addition to this, the observation diagnosis of the last five patients is not accurate. Hence, the third model (SVM3) will be able to check them in the process of training.

3 | ASSESSMENT OF THE MODELS

This section discusses the performance of all the three models using the data in training and test dataset. The Solver in MS Excel is used for the checking the working and training of the SVM models with the starting value of all the variables set to 1. Further, there is a discussion about the following metrics to compare the models and conclude the best of the three models to carry out the predictive values of the patients.

3.1 | Description of the Metrics

➢ Accuracy

This metric is used to check the accuracy of the predicted values of the patients and their susceptibility to have the heart disease. But, when the number of patients with the heart disease is meagre, this metric would give an inaccurate account of the performance of the model.

➢ Recall

Due to the high amount of money invested for the diagnosis, this metric is majorly important. The calculation of the number of correct identifications among all the patients is done using the recall metric.

➢ Precision

This metric helps the doctors to identify the proportion of healthy people among all the people who have been classified as healthy. This metric is crucial to screen out healthy people from the lot.

➢ Specificity

This metric helps us to understand the rate of misdiagnosis of the model. If the rate of specificity is very low, the conclusion is that the misdiagnosis rate of the model is high.

3.2 | Comparison of the Models

The first level of comparison is done for the above four metrics to calculate the performance of the three SVM models using the training dataset.

Metrics	SVM1	SVM2(λ=0)	SVM3(λ=0)
Accuracy	**Infeasible**	**84%**	**84%**
Recall	**Infeasible**	**88%**	**78%**
Precision	**Infeasible**	**70%**	**70%**
Specificity	**Infeasible**	**93%**	**93%**

Table 1 The Value of Each Metric for all the SVM Models

SVM1 vs SVM2: SVM1 being a hard margin model, faces difficulty in finding feasible solution within the training dataset due to the non-linear separability of the data. This means that it struggles to draw a straight line to divide the data cleanly in to two groups i.e., healthy people and the patients.

On the other hand, SVM2, being a soft margin model, accommodates a more flexible approach, which allows certain degree of misclassification. This makes SVM2 capable of handling the data which is not perfectly separable, which enhances its ability to separate patients and healthy individuals. Thus, SVM2 is the choice of model among SVM1 and SVM2.

Furthermore, there are only two models which can be selected as the optimal model for the analysis i.e., SVM2 and SVM3.

Any value of λ between 0 to 1000 gives the same accuracy rate as shown in the table below. This shows that there is no significant difference in the accuracy rate even when the value of λ is changed in SVM2.

SVM2	Training	Test
λ	Accuracy	Accuracy
0-10	84%	64%
50	84%	64%
100	84%	65%
500	84%	65%
1000	84%	67%
10000	82%	67%
50000	76%	64%

Table 2 Performance of SVM2 on the given two Datasets

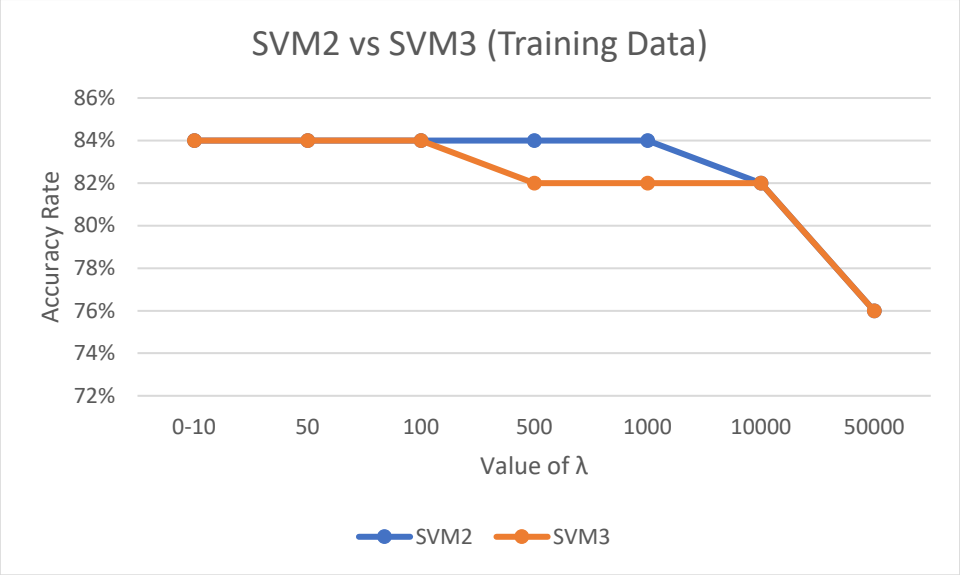
The classification rate for SVM2 is 0.84 and the accuracy for the model is 84% for the values of λ between 0 to 1000.

SVM2 vs SVM3: Table 2 gives us a clearer picture after applying the model on the training and test data. The high accuracy of 84% in the training dataset till the value of λ reaches 1000. Once the value of λ increases from 1000, the accuracy decreases to 82%. This suggests that the increase in the value of λ, the accuracy of the SVM2 model gradually decreases for training dataset. Also, when the same model is applied on the test dataset, a variable increase in accuracy rate is seen as the value of λ is increased till 10000. The accuracy comes back to 64% as the value of λ is increased to 50000.

SVM3	Training	Test
λ	Accuracy	Accuracy
0-10	84%	68%
50	84%	68%
100	84%	67%
500	82%	65%
1000	82%	65%
10000	82%	67%
50000	76%	65%

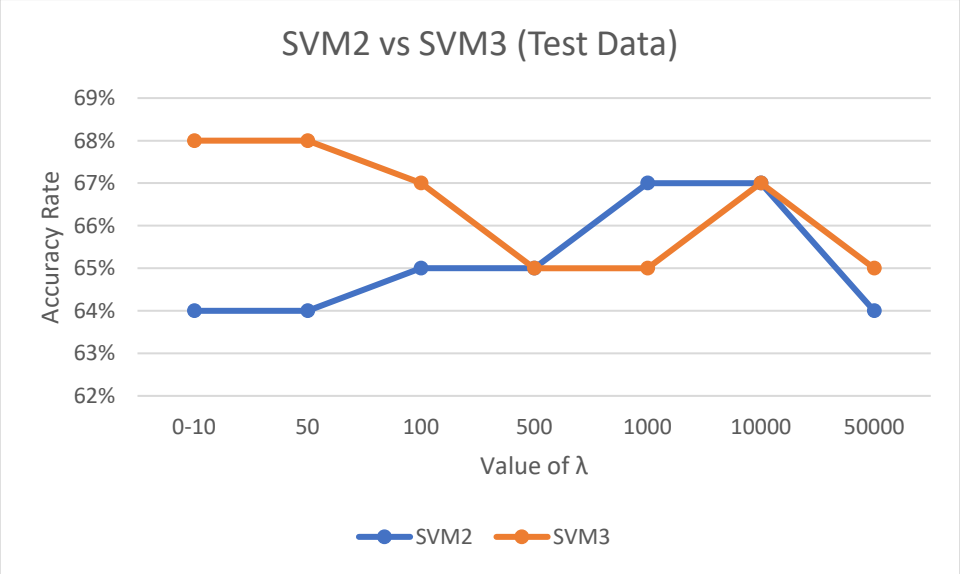
Table 3 Performance of SVM3 on the given two Datasets

The performance of SVM3 is shown in the table 3. This table depicts that the accuracy rate gradually decreases from 84% to 76% as the value of λ is increased from 500 to 50000 in case of the training data. When the same model is run for test data, the accuracy of the model varies from 68% to 65% as the value of λ is increased.



Graph 1 Comparison of Accuracy Rate based on λ Values for Training Data

The performance of both the models SVM2 and SVM3 shows a remarkable similarity, but SVM3 has an upper hand. This advantage becomes particularly significant in scenarios where the data might be incomplete or partially unavailable. SVM3, being a semi-supervised model, shows robustness by effectively leveraging both labelled and unlabelled data. SVM3's ability to manage the scenarios involving partial data makes it more adaptable and a sensible option, proving that the model is dependable even when the full data is unavailable.



Graph 2 Comparison of Accuracy Rate based on λ Values for Test Data

4 | THE OPTIMAL SOLUTION

The varying selection of the values of λ, gives us the following solution with the decision variables:

SVM3	λ=0	λ=70	λ=1000
x_1 (Age)	0.0029723	0.00271459	0.00139809
x_2 (Blood Pressure)	-0.030535	-0.02961986	0.009335712
x_3 (Cholesterol)	0.0013027	0.00129607	0.000540323
x_4 (Max Heart Rate)	-0.038954	-0.03831737	-0.019809172
x_5 (Intercept)	9.3050254	9.10439872	3.875361293

The SVM3 calculations suggest that older people are more likely to face heart problems. Knowing a person's age and cholesterol levels, things doctors can easily check, can help prevent heart issues when addressed early. Using **λ=70** gives values that make SVM3 a good choice for predicting heart diseases. When we use λ=70, the predictions match closely with what doctors diagnose, showing that the model works well.

5 | ADVANTAGES AND LIMITATIONS

4.1 | Advantages

1. **Robust Nature of Model in Noisy Data:**
   The semi supervised nature of the SVM3 model enhances the robustness in handling noisy or uncertain data. The model becomes tackles outliers and variations in the dataset, resulting in better overall performance of the model.
2. **Cost Effective Model Training:**
   Due to its decreased need on laborious manual labelling, SVM3 training is frequently more affordable than fully supervised models. Reducing the number of resources needed for intensive labelling efforts, the model may learn from the easily accessible unlabelled data.

4.2 | Limitations

1. **Complex Challenges in Solving:**
   While SVM3 has shows great performance potential, achieving the optimal outcome can be a tedious task. It usually ends with the solution which works well, but it might not be the best solution every time.
2. **Chance of Misdiagnosis:**
   In comparison with the fully supervised model, SVM3 might mistakenly label healthy people as patients more often. This might not make a huge difference to the doctors but will create an issue for the non-medical audience.
3. **Correcting Mistakes is Difficult:**
   There comes a need to double check the diagnosis as SVM3 may disagree with the doctors sometimes. This happens mostly for the semi-supervised data as for the last five patients in the training dataset.

6 | CONCLUSION

As more and more people face the heart disease, there is a growing need for accurate medical prediction methods. The report investigated three distinct Support Vector Machine models using the provided datasets. To determine the optimal solution, the testing of the three models under varied conditions was conducted and tabulated to compare them. The research gained credibility by examining the effects of various settings in the models.

The conclusion was reached by considering the models' advantages and limitations. With the value of λ lying between 0 and 1000, SVM3 performs exceptionally well. It had a high accuracy and low mistake rate for identifying heart issues with the patients. Based on the findings, it is advisable to use SVM3 to make a diagnosis of the heart disease.

APPENDIX

1. Mathematical model of SVM1, SVM2, SVM3:

Symbol	Description
$a\%$	Age (decision variables)
$b\%$	Resting blood pressure (variables)
$c\%$	Cholesterol level (variables)
$m\%$	Maximum heart rate (variables)
$t\%$	A label to mark whether people are sick or not (-1-healthy,1-sick)
$z\%$	Whether people are diagnosed with disease or not(0-healthy,1-sick)
$y\%$	The classification error (variables)
K	Visiting patients (50)
J	The last fifth patient (45)

(SVM1) $min \quad x_1^2 + x_2^2 + x_3^2 + x_4^2$

$$s.t \quad x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b} \geq 1 \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b} \leq -1 \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$

(SVM2) $min \quad \lambda(x_1^2 + x_2^2 + x_3^2 + x_4^2) + \sum_{k=1}^{K} y_k^2$

$$s.t \quad y_k \geq 1 - (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$y_k \geq 1 + (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$
$$y_k \geq 0 \quad \forall k \in \{1, \dots, K\}$$

(SVM3) $min \quad \lambda(x_1^2 + x_2^2 + x_3^2 + x_4^2) + \sum_{k=1}^{K} y_k^2$

$$s.t \quad y_k \geq 1 - (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = 1(patient)$$
$$y_k \geq 1 + (x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) \quad \forall k \in \{1, \dots, K\}: t_k = -1(healthy)$$
$$y_k \geq (1 - 2z_k)(x_1 a_k + x_2 b_k + x_3 c_k + x_4 m_k + \mathrm{b}) + 1 \quad \forall k \in \{1, \dots, J\}: t_k = 0(undiagnosed)$$
$$x_1, x_2, x_3, x_4, b \ unrestricted$$
$$y_k \geq 0 \quad \forall k \in \{1, \dots, K\}$$
$$z_k \in \{0,1\} \quad \forall k \in \{J+1, \dots, K\}$$

2. The Algorithms of Accuracy, Precision, Recall and Specificity:

$$Accuracy: \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$Precision: \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall: \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specificity: \frac{True\ Negative}{True\ Negative + False\ Positive}$$