# COMP3009 Machine Learning

## Assignment 1: Report

November 11, 2021

Group 18

| Name | Student ID | Username |
|---|---|---|
| Teo Shi Bin | 20183717 | hcyst2 |
| How Khai Chuin | 20210654 | hcykh1 |
| Loh Qian Kai | 20194664 | hcyql1 |
| Mohamad Arif Bin Mohamed Abu Baker | 20116042 | hfymm5 |
| Muhammad Fikri Bin Mohd Roslee | 20116065 | hfymm4 |

# Contents

# Introduction

In this coursework we have chosen two data set where one is classification for death event and another one is regression for concrete strength. Both datasets were shuffled and normalised using Min-Max Normalisation with 5%/95% percentile.

## Data Set for Classification

For the classification, we have chosen the heart failure clinical records data set. It consists of 12 features as well as a total of 299 instances inside the data set. The goal is to classified the death event based on the 12 features.

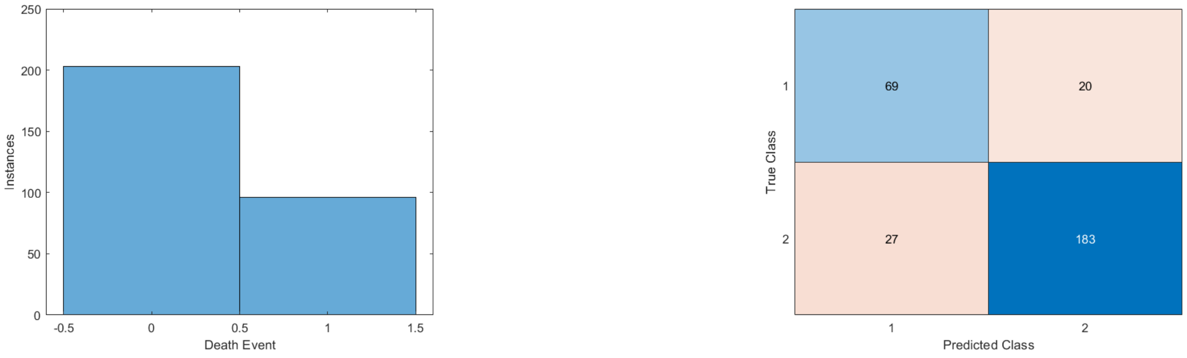
## Data Set for Regression

For the regression, Concrete Compressive Strength data set is chosen. In this data set it consists of 8 features and a total of 1030 instances inside the data set. The goal is to predict the Concrete Compressive Strength based on the 8 features.

# Task 1: Train SVMs with the Linear Kernel

## Classification

The trained SVM performance is shown here (Figure 1b). The confusion matrix is provided because accuracy might not be the best metric to measure the performance of the model, this is caused by the unbalance classes shown here (Figure 1a).



(a) Unbalance Classes

(b) Confusion Matrix

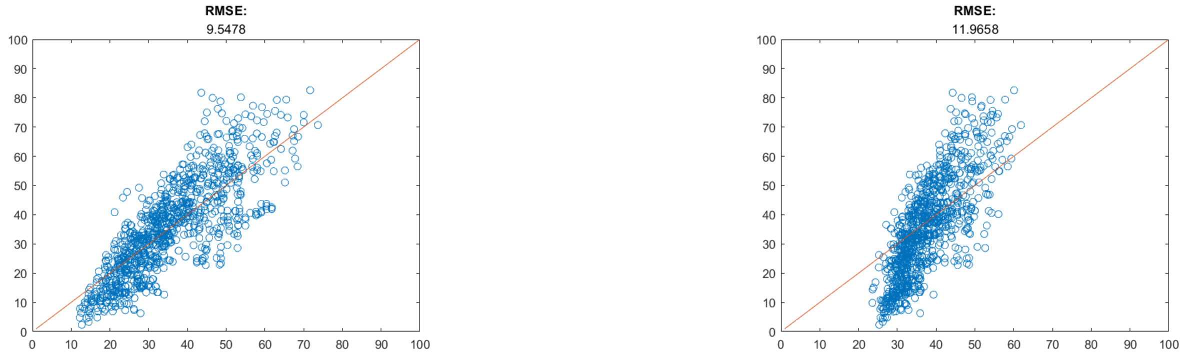
**Accuracy: 0.842809**

Figure 1: Linear Classification Result

The linear model able to fit the train data set quite well with accuracy 84% but we can seek for alternative model to have a better fit.

## Regression

Epsilon defines maximum distance between a data point and the function that can be neglected in the cost function. Smaller epsilon making the function trying to fit more data point, while larger epsilon allow the function to be simpler but still capture the data point in the range of epsilon.



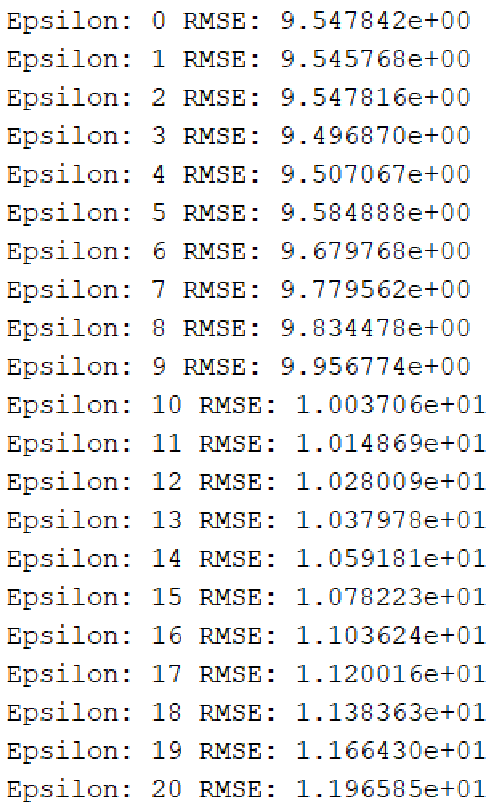
(a) Epsilon = 0

(b) Epsilon = 20

Figure 2: True Value against Predicted Value

Figure 2 and Figure 3 shows the impact of changing epsilon value. In Figure 2, a perfect model will always predict the right value which lies on the red line. We can see lower value of epsilon will have lower Root Mean Square Error (RMSE) than the higher value of epsilon. A full list of result is shown in Figure 3, the RMSE increase as the value of epsilon increases.

This is reasonable as the penalty is high for low epsilon to have an unfit data point. However, it doesn't mean that epsilon should always be low to have better fit because overfitting to training data set will bring bad performance on unseen data (test data set). Therefore, we need hyper-parameter tuning and data splitting to find a appropriate value of epsilon.



```
Epsilon: 0 RMSE: 9.547842e+00
Epsilon: 1 RMSE: 9.545768e+00
Epsilon: 2 RMSE: 9.547816e+00
Epsilon: 3 RMSE: 9.496870e+00
Epsilon: 4 RMSE: 9.507067e+00
Epsilon: 5 RMSE: 9.584888e+00
Epsilon: 6 RMSE: 9.679768e+00
Epsilon: 7 RMSE: 9.779562e+00
Epsilon: 8 RMSE: 9.834478e+00
Epsilon: 9 RMSE: 9.956774e+00
Epsilon: 10 RMSE: 1.003706e+01
Epsilon: 11 RMSE: 1.014869e+01
Epsilon: 12 RMSE: 1.028009e+01
Epsilon: 13 RMSE: 1.037978e+01
Epsilon: 14 RMSE: 1.059181e+01
Epsilon: 15 RMSE: 1.078223e+01
Epsilon: 16 RMSE: 1.103624e+01
Epsilon: 17 RMSE: 1.120016e+01
Epsilon: 18 RMSE: 1.138363e+01
Epsilon: 19 RMSE: 1.166430e+01
Epsilon: 20 RMSE: 1.196585e+01
```

Figure 3: The result of changing value of epsilon

## Task 2: Train SVMs with Gaussian RBF and Polynomial Kernels

The number of support vectors below are all retrieved from all 6 of the scripts using nested cross validation and trained using the best hyper-parameter for each cases, denoted as **support vectors / vectors (percentage)** .

Support Vectors Count and Percentage

| **Dataset \Kernel** | **Linear** | **Polynomial** | **Gaussian** |
|---|---|---|---|
| **Classification** | 158/299 (0.528428) | 141/299 (0.471572) | 169/299 (0.565217) |
| **Regression** | 826/1030 (0.801942) | 546/1030 (0.530097) | 695/1030 (0.674757) |

## Task 3: Method Evaluation

The F1 score below are obtained from from all 3 classification scripts with nested cross validation and 3 of them were trained with grid search in finding each of the most optimal hyper-parameter. F1 score is being used is because it can balance between the Precision and Recall so with this we would not have accuracy that contain a large number of True Negatives. In this case here, F1 score is better to be used as it can be balanced between the Precision and Recall rather than finding the most optimal hyper-parameter that is just determine by the accuracy.

Performance for Classification

| **Kernel** | **F1 score** |
|---|---|
| Linear | 0.654426 |
| Gaussian | 0.621967 |
| Polynomial | 0.579288 |

The RMSE below are obtained from the 3 regression scripts with the implementation of nested cross validation and 3 of them were trained with the most optimal hyper-parameter that is found using grid search. RMSE is used as it indicate on how closed is the data point from the predicted value by the model.The lower the value obtained by RMSE means that it is fitting the linear regression line closer and is better. As a result, it is good in showing how accurate the model predicts the value given.

Performance for Regression

| **Kernel** | **RMSE** |
|---|---|
| Linear | 12.942871 |
| Gaussian | 6.462563 |
| Polynomial | 3.769308 |

## Task 4: Additional Questions

1. **What does the kernel parameter of the Gaussian RBF kernel signify (sigma)? What happens when you increase its value?**

   **Answer:**
   In Gaussian Distribution, sigma represent the standard deviation which will determine the width for the distribution.The Gaussian RBF kernel is mathematically represented with:

   $$K\left(X_1, X_2\right) = \exp\left(-\frac{\left\|X_1 - X_2\right\|^2}{2\sigma^2}\right)$$

   The sigma in the kernel controls how much separating hyperplane bends around the training examples. Let $d_{12}$ be the distance between the two points $X_1$ and $X_2$ when plotting the graph.

   $$d_{12} = \left\|X_1 - X_2\right\|^2$$

If the sigma value is small($\sigma = 0.1$), the width for the region of similarity is minimal. Thus, the points are considered similar only if the points are very close.

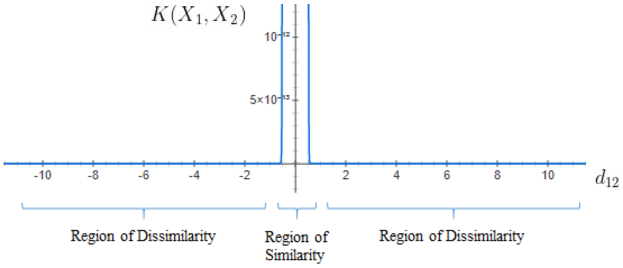


Figure 4: RBF Kernel $\sigma = 0.1$

When sigma value is increased to a larger value, $\sigma = 10$, the width of the curve is larger and the region of similarity increases. The points farther away can now be considered similar.

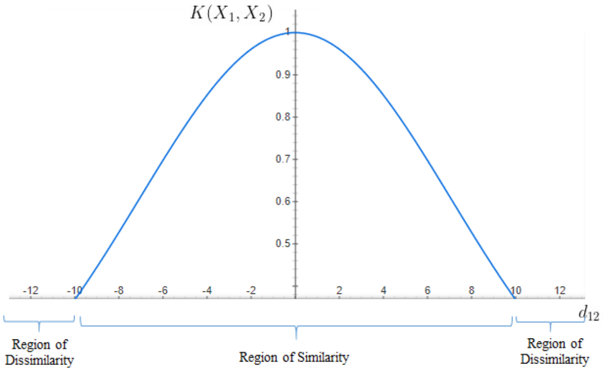
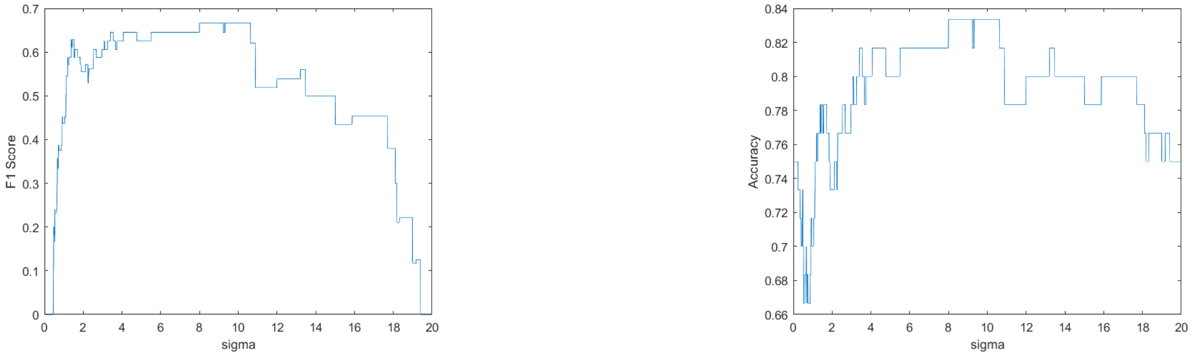


Figure 5: RBF Kernel $\sigma = 10$

From the example above, it can be observed that the width of the region of similarity changes as sigma changes. Sigma acts like an amplifier of the distance between $X_1$ and $X_2$. When the distance is larger than sigma, the kernel function tends to be 0. Therefore, with a small sigma value, the decision boundary tends to be strict and sharp and tends to overfit . Generally, smaller sigma tends to produce a local classifier whereas larger sigma produce a more general classifier. Larger sigma tends to have less variance and more bias. Smaller sigma tends to have more variance but less bias.

To observe the effect of the value of sigma on the heart failure dataset(refer task4.m for matlab code), the graphs of the f1 score against sigma (Figure 6a) and accuracy against sigma (Figure 6b) were plotted. From the graph, we can observe that a sigma value that is too low or too high will contribute to poor F1 score result.



(a) Graph of sigma value and f1 score

(b) Graph of sigma value and accuracy

Figure 6: Graph for heart failure dataset

2. **Explain what happens when a hard-margin SVM is fit to a dataset of two classes with overlapping features. What value do you need to set C (the slack-variable hyper-parameter) to attain a hard-margin SVM?**

   **Answer:**

   Hard-margin SVM will not works on overlapping features. This is because SVM is trying to minimize the cost function subject to the constraint. However, overlapping features will violate the constraint and make SVM failed to satisfy the constraint in quadratic programming.

   The C determine the weight of slack variable in the cost function. In order to mimic the behaviour of hard-margin, the C have to be infinity or a very big value so the slack variables have to be small when optimizing the cost function. Hard-margin behaviour will happens when slack variables are zeros due to the large C.

3. **Explain why do you need to use both inner cross-validation and an outer k-fold cross-validation (known as nested cross-validation) in the machine learning process?**

   **Answer:**

   The nested cross-validation procedure is used to overcome the problem of overfitting the training dataset, retrieving generalisation error of unseen data and optimizing hyper-parameters. The idea of nested cross validation is simply optimizing hyper-parameter using the inner cross validation and validate it against the inner test set, eventually once the tuning is done, use the optimized model to test it against the outer test set to retrieve generalisation error or other performance measures. For parameter tuning we can use grid search, random search or other complex methods, in our case we're using grid search to optimize the hyper-parameters. Grid search evaluates each configuration of hyperparameters and choose the best configuration by comparing chosen performance metric. Cross validation reduces the risk of overfitting the dataset and generalizing a much robust model that performs better on unseen data.

# Conclusion

The best classification kernel for our death event data set is linear kernel. Gaussian and polynomial kernel are non-linear thus capable of handling dataset that are not linear-separable and usually it performs better than linear kernel. In this case, the linear outperformed both kernels might be the reason of lacking data instances. In our case most of the model bias towards ”no death event”, however due to the lack of instances we preserve current results.

The best regression kernel for our concrete strength data set will be the polynomial kernel because the RMSE of it is the lowest. This result contrast to the result of classification as the larger dataset and less attributes might have introduced more certainty to polynomial and gaussian kernel.