CSE5ML Assessment 1 Report

Michael Le 21689299

Method Procedure Task 1A. to Task 1E.

Before pre-processing the datasets, from the originally dataset contains 2928 rows featuring 19 columns gathered from the WHO (World Health Organization). Were interested in the 18 columns used to predict to generate our design matrix X, to find the life expectancy, the true label for the variable y. During the pre-processing process, we first want to identify any missing values in the dataset. We identified that there are three columns that has missing values which are Hepatitis B, Polio, and Diphtheria each missing 19 values, we decide to drop the following rows with null values and duplicate to maintain consistency and keeping relevant and unique data. Leaving 2909 rows remaining in the dataset, in addition we must remove any unit strings and transform any columns into float types. But luckily, none of the columns have unit strings. Furthermore, after pre-processing I observe there are 18 columns that are numeric (all are continuous) expect for only 1 column which is categorical (nominal) for the column Status (one to determine if the Status is Developed or Developing for that country). In terms of correlation analysis, we first read off the first column, interested in Life Expectancy when applying efficient understanding of probability and statistics and further mathematics, there are three different types of correlations (Positive, Weak or no correlation and Negative relationships). We conclude, by observation that there is positive strong relationship in BMI and Schooling. In addition, the following columns has a negative relationship in Adult Mortality and HIV/AIDS. It is important to understand this because this can impact predictions based on the patterns in data. Overall, it can help guide business processes, direction and performance when overcoming the relationship between multiple parameters in regression models.

BACKGROUND KNOWLEDGE

In supervised learning, Linear regression according from the lectures, involving labelled data are continuous, ordered and are not expressive. In a linear model, contains a prediction which is the sum of the product of features and weights. Our goal is to find the coefficient of the weights in the linear model. To achieve this, we need to train and test from our updated dataset from Task 1 to perform the loss function followed by the least squares regression. For Support Vector Machine (Regression case) is another supervised learning technique is to separate data by a hyperplane (assuming data is separable). The goal is to optimize the margin, which is computed from the smallest distance of data points/residuals, which it can handle high dimensional data. For the purpose for this assessment, we apply using cross validation is to find accurate predictions within our dataset. To prevent issues when dealing with overfitting or underfitting our regression models.

METHOD PROCEDURE Task 2A to Task 2E

To achieve this, we follow under from the lab 2 Part 1 Material for the regression models, we first must compute our values for X and y from our updated dataset. Then we split the training and test datasets using 90% and 10% respectively with a random state of 123. Second, we normalise the on both datasets due to the impact of the model weights to make the data stabilised by checking the scale of the outputs and gradients that are affected by the inputs. In our regression models we must define, including linear regression and SVM with their default settings. One approach was to the train the model based on the entire training dataset and then evaluate the model based on the testing dataset. After computation, we have the following values of R-squared which is a statistical measure use to represent how good the model can fit. In this case, for the regression and SVM is 0.8325 and 0.86107 respectively. The second approach is to use (k=5) k-fold cross validation to train and evaluate the two models based on the average score. After computation, we have the following R-squared values for regression and SVM which are 0.8376 and 0.85248 respectively.

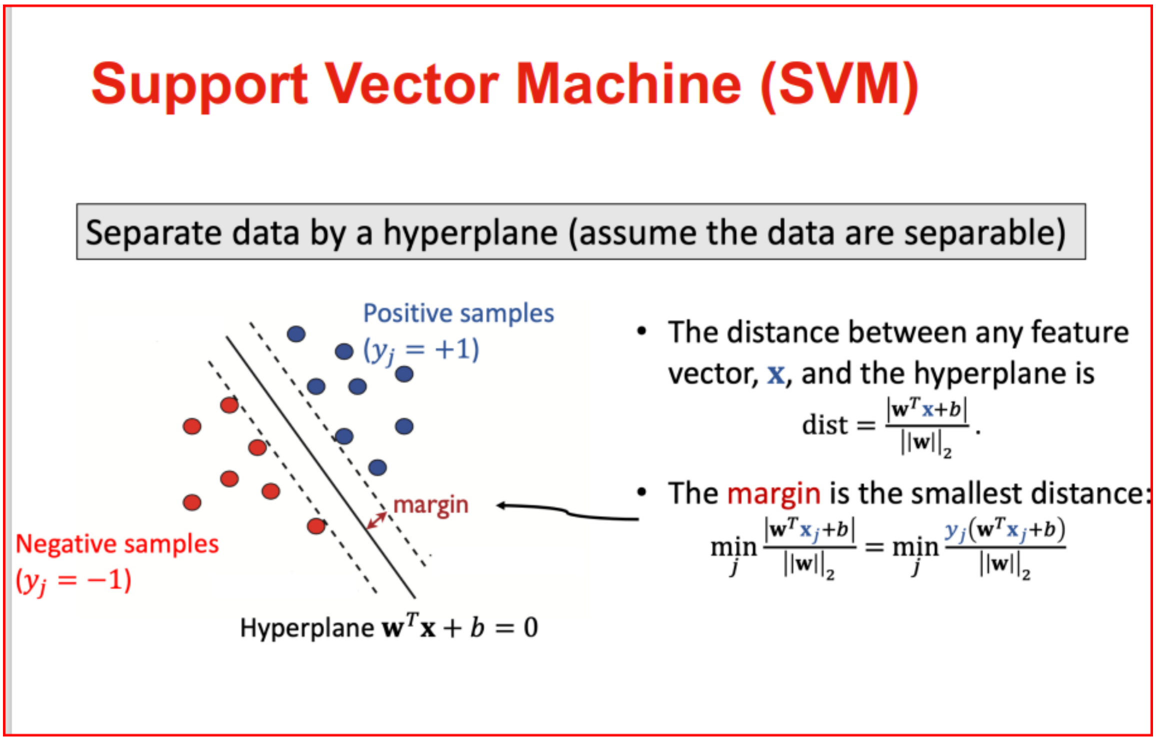
Next, we apply parameter finetuning for the regression and SVM models in separately to optimise the model performances and compare the cross-validation results before and after finetuning for each model. After computing the regression model, we have a result of 0.8376 and for the SVM model we have 0.9077 which is slightly higher by removing the coef0 component inside the grid parameters svr dictionary to make computation and compiling faster. Similar reasons were mentioned in the previous procedures applied during evaluation in default settings and cross-validation methods.

For the last part of this assessment, we evaluate the two optimised models (with the best parameter setting from Task 2c. for each model type) on the test set, then compare the results from Task(s) 2b. We first evaluate the trained Linear Regression model using the testing dataset given the r squared value of 0.83235 and for the SVM was 0.913388. Following by finding the predictions for the first five data points given in the training datasets for regression and SVM. For the final step, we save and load a trained model for the linear regression and svm models with the following R-squared values of 0.83235 and 0.913388 respectively.

TABLE/DISCUSSION:

	Regression Model	**SVM Model**
TASK 2A: Training models based on the entire training dataset. Including evaluation based on the testing dataset.	0.8323515207054641	0.8610735001837764
TASK 2B: Cross-Validation	0.8376013975750765	0.8524822144041153
TASK 2C: Parameter-finetuning	0.8376013975750768	0.9077343481000133
TASK 2D: Evaluate the trained models using testing dataset.	0.8323515207054651	0.9133880144997648

Overall, in all different cases the SVM has the highest value compared to the regression model due to the SVM capacity to capture data in higher dimensions which can easily separate data in a continuous space. **(See Figure 1. Lecture 2 Slide 31.)**



Support Vector Machine (SVM)

Separate data by a hyperplane (assume the data are separable)

- The distance between any feature vector, $\mathbf{x}$, and the hyperplane is
$$\text{dist} = \frac{|\mathbf{w}^T\mathbf{x}+b|}{||\mathbf{w}||_2}.$$
- The margin is the smallest distance:
$$\min_j \frac{|\mathbf{w}^T\mathbf{x}_j+b|}{||\mathbf{w}||_2} = \min_j \frac{y_j(\mathbf{w}^T\mathbf{x}_j+b)}{||\mathbf{w}||_2}$$

Figure 1. Lecture 2 Slide 31.