107062381 楊孝偉

Machine Learning Homework 2

In this homework I used support vector machine as my model to predict the 30-day mortality of patients.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space. Here, N is the number of input features that distinctly classifies the data points. To separate the two classes of data points (either the patient is alive or dead), there are many possible hyperplanes that could be chosen. The main goal is to find a plane that has the maximum margin. Maximizing the margin distance provides some reinforcement so that future data points can be classified easier.

There are several features available for me to use. I will be using all of the features, except ed_diagnosis, admission_datetime, sex and PATIENT ID. I extract the required features and split it intro training and testing data. The x_train is taking the values of `hm_hospitales_covid_structured_30d_train.csv`

The y_train is taking the values of hospital_outcome from `split_train_export_30d.csv`

These two are used to train the SVM model.

The x_test is taking the values of `fixed_test.csv`

Then I built the SVM model using the Scikit learn library and just call the related functions to implement the SVM model. The kernel is set to linear.

There are quite a lot of NaN cells in the original dataset used for training the model. This isn't good if not preprocessed first and directly fed to our model. Therefore, I used the pandas dataframe.fillna function to fill the NaN cells with the dataframe.mean value.