**Solution Sheet**

1. Which model have you used for probability prediction? Explain your model.

The model used for the probability prediction is an Artificial Neural Network(ANN).

I have encoded all the values in the dataset and trained the ANN on it. The ANN has a basic Sequential architecture with 1 input layer,3 middle layers and 1 output layer. The input layer has input dimensions 28 and the output layer has the dimension 1.

How an ANN works is that each node in an ANN has connections to the previous layers node and next layer node, but never to a node of the same layer. Each connection in an ANN has some weights attached to it. The main job of an ANN while training is to get the best value of those weights. The value of all the previous layer nodes connected to the current layer node is multiplied with each weight, summed up and then an activation function is applied on the sum. The activation function applied in the ANN is ReLu(this prevents the values from going into negative). When the output node’s value is calculated, it differs a bit from the expected value of the output node.

To fix this error in value, we backtrack through the model from output layer to input layer, modifying the weight as we see fit. This whole process is done for every value in the Train Dataset(doing it once fully for the training dataset is called an epoch). Our ANN runs for 500 epochs. The ANN is validated on a specific set of validation values. This calculates how good our model is performing given some input values and how close its predicted values are to the actual values

Running the ANN for so many iterations gives us the best values for the weights. Each epoch’s checkpoint (weight values) is saved if the validation loss in that epoch is the smallest till now.

The best checkpoint is loaded and the testdataset is taken as an input. The ANN then predicts the infected probability for each row in the test dataset.

1. Which model have you used for Diuresis Time series prediction? Explain your model.

The model used for the Diuresis Time Series prediction is an LSTM (Long Short Term Memory) model. LSTM networks manage to keep contextual information of inputs by integrating a loop that allows information to flow from one step to the next. Basically, the LSTM network decided what values to remember and what values to forget. The core concept of LSTM’s are the cell state, and it’s various gates. The gates decide which values to pass forward to the next step, or which value to disregard completely(forget).

Our LSTM has a sequential architecture with 2 recurrent LSTM layers and 1 Dense layer.

We define a window size for the LSTM. We take it as 3. A window is basically considering specific columns(in the window) to predict values of another column.

In our LSTM, we take columns 1,2,3 to predict columns 4

We take columns 2,3,4 to predict column 5

We take columns 3,4,5 to predict column 6

We take columns 4,5,6 to predict column 7

In our LSTM, we take the loss function as MSE (Mean squared error). This error is calculated between actual values and predicted values. This error is then used with backtracking to modify the weights of each connection.

We train the LSTM for each row in the Diuresis dataset. Each completion of all the rows is called an epoch. We do the same for 400 epochs.

This is basically the training procedure of the LSTM. Knowing our final outputs, we can modify the weights of the connections in the LSTM accordingly.

We can then use the final weights to take column 5,6,7 as input and predict column 8 27th March) as output.

We use these 27th March Diuresis values and replace them in the Training dataset, and train the previously defined ANN model on the new values. Then, apply the model on the test dataset and get the infected probability values.