Minor Project 1 – Cognitive Application

(Ayush Goyal, Manipal Institute of Technology, CSE,aiqqia.ag@gmail.com)

Theory Questions

1. **What is overfitting and how to avoid it?**

Overfitting in Machine Learning is a deficiency in Machine Learning that hinders the accuracy as well as the performance of the model. It happens when a model captures noise (randomness) instead of signal (the real effect). As a result, the model performs impressively in a training set, but performs poorly in a test set.
To avoid overfitting, we could use the following practices:
One good way to train models is to split the days into training and test sets - evaluate and update the model till it performs well on both training and testing data. We can use cross-validation of an appropriate number of folds to do train and test the model on different portions of data at a time to get a better idea of the accuracy.
Feature engineering - feature selection & feature engineering
Regularization (Lasso & Ridge)
In neural networks - we can make sure the network isn't very deep, we can have a dropout model in case there might be many redundant nodes in the neural network.

2. **What is RMSE and MSE? How can you calculate them?**

Root Mean Squared Error (RMSE) is the square root of the mean square error. That is probably the most easily interpreted statistic, since it has the same units as the quantity plotted on the vertical axis. The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point. The two should be similar for a reasonable fit.
The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error) and square the value. Then you add up all those values for all data points, and, in the case of a fit with two parameters such as a linear fit, divide by the number of points minus two. The squaring is done so negative values do not cancel positive values. The smaller the Mean Squared Error, the closer the fit is to the data. The MSE has the units squared of whatever is plotted on the vertical axis.

3. **What is Line of best fit?**

The Linear Regression model attempts to find the relationship between variables by finding the best fit line.
The line of best fit is a line that best represents the relationship between a scatter plot of data points. Just by looking at the data points, we can't predict anything. Whereas, if we use these data points to create the best fit line, we can easily predict the future values.

4. **Explain multivariant linear regression using a real-life example.**

   Simple linear regression is finding a line that best fits/correlates two features to give a continuous output. When we use more features, regression with more than two features is multivariate linear regression.
   A real-life example can be: Housing price prediction by using features - size of house, no of bedrooms, location, furnished/unfurnished, etc.

5. **How can we improve the accuracy of a linear regression model?**

   Accuracy of a linear regression model can be improved in the following ways:
   a. Model tuning
   b. Handling missing values and outliers
   c. Encoding categorical variables
   d. Feature selection and scaling
   e. Hyper parameter tuning

Simple Linear Regression with Python

Dataset: https://s3.us-west-2.amazonaws.com/public.gamelab.fun/dataset/salary_data.csv

Reference: https://towardsdatascience.com/machine-learning-simple-linear-regression-with-python-f04ecfdadc13

I have implemented the **Simple Linear Regression** algorithm over the given dataset, and I have run it in Jupyter notebook. The ".ipynb" file and the dataset along with the graph shown is implemented and the same is uploaded on my GitHub repository, the link to which is given below.

GitHub Repository: https://github.com/aiqqia/Machine-Learning