SIT 720 - Machine Learning

Lecturer: Chandan Karmakar | karmakar@deakin.edu.au

School of Information Technology,
Deakin University, VIC 3125, Australia.

Assessment Task 3 (40 marks)

Submission Instruction

1. Student should insert Python code or text responses into the cell followed by the question.
2. For answers regarding discussion or explanation, **maximum five sentences are suggested**.
3. Rename this notebook file appending your student ID. For example, for student ID 1234, the submitted file name should be A3_1234.ipynb.
4. Insert your student ID and name in the following cell.

In [ ]:

```
# Student ID:

# Student name:
```

Background

Environment and its changes are the most complex system. It is unarguably accepted that the temperature changes are greately affected by various environmental factors. Many of them are positively related to the change, whereas, some have negative correlation. In this assesment task, you will analyse relationship among various environmental factors, which affect temperature.

The dataset

Dataset file name: weather_dataset.csv

Dataset description: The dataset contains total 10 features. Each row contains an hourly record of weather status and the data was recorded for the time period between 2006 and 2016.

Features and labels:

1. recording_date_time (date_time): Date and time the data was recorded
2. precip_type (string): Precipitation status, blank (no value) indicates unknown status
3. temperature (float): Temperature in degree Celsius
4. apparent_temperature (float): Feel like temperature in degree Celsius
5. humidity (float): Percentage amount of water vapour in the air
6. wind_speed (float): Speed of the wind in km per hour
7. wind_bearing (int): The direction of wind in degree in geo-polar co-ordinate. Value 0 means perfect east, 90 means perfect north, 180 and 270 means west and south respectively.
8. visibility (float): Distance in km that is visible in naked eyes.
9. cloud_cover (float): The fraction of the sky obscured by clouds. The value is 1 if the observed area is fully cloudy, 0 if no clouds and other fractional value indicates the portion of the area covered by clouds.
10. pressure (float): Air pressure or atmospheric in milibars

Part 1: Linear Regression: (25 marks)

1. Load the dataset and split the data for training and testing - consider the data of last 2 years (2015 and 2016) for testing. Now exclude recording_date_time column from both training and test sets. Display the shape of training and test sets. **(3 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

1. Consider the 'temperature' as the target. List the insignificant features for predicting temperature, if any. Explain your findings. **(5 marks)**
   [Hint for students: See the "7.3 Relevance and Covariance among features or variables" for more information.]

In [ ]:

```
# INSERT your code (or comment) here
```

1. Now create a linear model considering the 'temperature' as the target variable and other columns as features (you can optionally remove non-contributing features). Show the test performance (as Mean Absolute Error, MAE) of the model. **(5 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

1. Find the feature which shows maximum correlation with "pressure". Create a linear regression model to predict temperature using these two features ('pressure' and the one which shows maximum correlation). Compare the performance of this simplified model with the model developed in the previous question (Q-3). Explain the performance variation, if any. **(6 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

1. Apportion the complete dataset into training and test sets, with an 40-60 split. **(6 marks)**

   (a) Train a linear regression model without considering overfitting scenario and report the test performance.

   (b) Create an optimal regularised linear regression model and report the test performance.

   (c) Explain the reason behind the performance variation, if any.

In [ ]:

```
# INSERT your answer in maximum five sentences.
```

Part 2: Logistic Regression: (9 marks)

1. Can the same target (temperature, mentioned in Part-1) be used for logistic regression? Why? **(2 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

1. Split the dataset as 70-30% for training and testing. Create a logistic regression model to predict the 'precip_type'. Report the prediction accuracy of your model whether the "precip_type" is "rain" or not (use decision threshold of 0.45). **(5 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

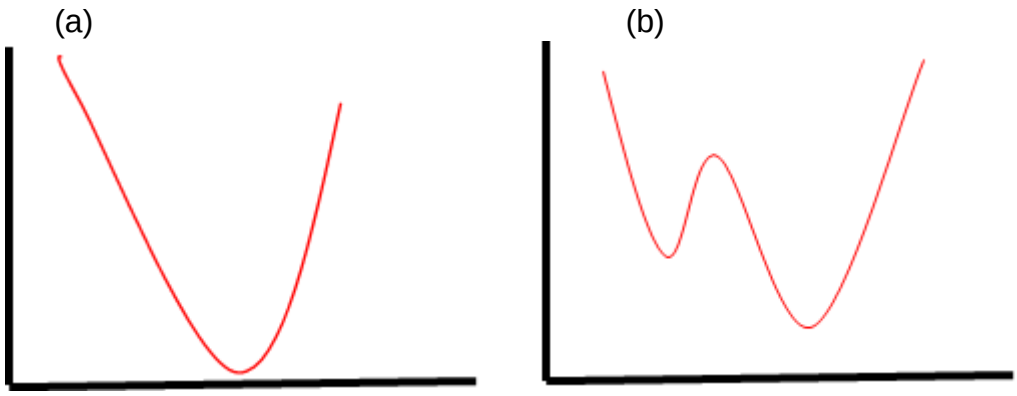
1. Discuss the test performance using precision, recall and confusion matrix. **(2 marks)**

In [ ]:

```
# INSERT your code (or comment) here
```

Part 3: Objective function optimisation: (6 marks)

Let's consider the line graphs shown below and answer the following questions [Hint: See weekly content 7.4-7.10],



a. Which of the above figures represents the convex objective function and why? (**1 marks**)

b. Which hyper-parameter can help to reach the convergence point and the impact of value selection? (**2 marks**)

c. How can we find the global minima for the objective function shown in Figure-b? *[N.B. Conceptual description will be accepted.]* (**3 marks**)