ECE 720-x50: Introduction to AI

Mini Project 1: Machine Learning

-- TASK 1 --

You are asked to construct three regression models.

There are two data sets that should be used to build the models:

- **Edmonton Weather Data**
  https://www.kaggle.com/amit321/edmonton-weather-data/download
- **Mosquito Trap Data**
  https://data.edmonton.ca/Environmental-Services/Mosquito-Trap-Data/dg7f-ubac

The models should represent a function between a number of mosquitos and weather features. In other words, you should construct **one** linear regression for:

$$\text{no\_of\_Mosquito} = \text{function}(\text{weather\_features}^{A})$$

and **one** linear and **one** polynomial models for:

$$\text{no\_of\_Mosquito}_{\text{Female}} = \text{function}(\text{weather\_features}^{B})$$

or:

$$\text{no\_of\_Mosquito}_{\text{Male}} = \text{function}(\text{weather\_features}^{C})$$

Notes:

- You need to combine both data sets to prepare a data set suitable for model construction
- Both data sets have different 'time grid' so you need to resolve it
- You have to aggregate data points over time, and different mosquito species

Use different cost functions, perform analysis of constructed models, apply normalization and standardization, feature selection.

-- TASK 2 --

You are asked to construct a predictor using three different models/approaches: Logistic Regression, Support Vector Machine, and Random Forest.

In the case of datasets, you have two choices:

Choice A: Two data sets from Task 1:

Edmonton Weather Data

https://www.kaggle.com/amit321/edmonton-weather-data/download

Mosquito Trap Data

https://data.edmonton.ca/Environmental-Services/Mosquito-Trap-Data/dg7f-ubac

The classification should be:

Mosquito_Female OR Mosquite_Male = function(weather_features^{D})

Choice B:

Ionosphere Dataset

https://archive.ics.uci.edu/ml/datasets/Ionosphere

The Ionosphere Dataset requires the prediction of structure in the atmosphere given radar returns targeting free electrons in the ionosphere. It is a binary (2-class) classification problem. There are 351 observations with 34 input variables and 1 output variable.

The classification should be:

g for good and b for bad = function(input_variables^{E})

Please, perform 10-fold cross validation of each model and t-test in order to identify which model is the best.

Deliverables:

- Jupyter Notebook with all your activities and texts describing them.

Deadline:

- Tuesday March 16th, 11:55 PM

NOTE: subscripts A, B, C, D and E means that in each of the models you can/should utilize a subset of input variables.