UMSL

CMPSCI- 4300 - Introduction to Artificial Intelligence
Department of Computer Science

Semester: 2022 Spring

Instructor - Navneet Kaur

Phase 1: Data analysis & preparation

The most important task in this phase is to find a dataset for your project.

Tips for Selecting/Picking the Dataset:

- You can pick a dataset from UCI ML database or Kaggle.
- If you are learning machine learning for the first time, a binary classification problem will probably be easier (not a regression problem).
    - A problem is 'binary classification' if your output column is 0 or 1. For a binary classification problem, the single layer model is a logistic regression model.
    - If your output column has continuous values, even if they are between 0 and 1, it is a regression problem, not classification. For a regression problem, the single layer model is a linear regression model.
- Before selecting the dataset make sure that the data is not imbalanced
    - In the case of Binary Classification check by calculating: what percentage of the output labels are 0 and what percentage are 1 .
    - In the case of Regression, check if the values are uniformly distributed or not: by plotting (using visualization technique refer below) the distribution of the output variable.
- If your dataset is heavily imbalanced (for example, too many examples of one class and very few examples of the other class) it may be a better idea to choose a different dataset.

Restrictions while choosing/selecting a dataset.

- You are not allowed to pick a time-series dataset. An example time series data is a stock price dataset.
- You are also not allowed to select a dataset consisting of image inputs or text inputs (natural language processing datasets).
- Do not choose the "Iris flower dataset", "Pima diabetes dataset", or the "Wine quality dataset".
- Your tabular data should have at least around a 1000 rows and at least 3 features (columns) in addition to the output column.

UMSL

CMPSCI- 4300 - Introduction to Artificial Intelligence
Department of Computer Science

Semester: 2022 Spring

Instructor - Navneet Kaur

Pre- Requisites:

- Before working on this phase, please practice “Activities 1 and 2” .
- You may also find this short lecture on ‘how to clean a tabular dataset for machine learning’ useful to learn how to clean your data.
- Refer the Example Report

Tasks

- Once you have found a dataset of your choice, the next step is to upload the data in a Python Notebook. Refer Activity 1 Basics of Python for uploading your dataset to the notebook.
- Describe your data to check the features of your dataset
- Discuss the range of the values (min, max, mean, median, etc.) by using `data.describe()`.
- Visualize your data.For example, plot histograms showing distribution of each input feature. For visualization Refer Visualization Techniques. You can also refer these videos: Video 1 , Video 2
- Check the distribution of the output label and visualize/plot that.
- Normalize your data. Check Activity 2 or Refer the video Normalization Technique

In your report:

- Include Abstract and Introduction
- Discuss why you chose to work on this project.
- Describe the dataset and its source.
- Show your plots/graphs/charts.
- Discuss the distribution of the output labels..
- Discuss how you normalize your data

UMSL

CMPSCI- 4300 - Introduction to Artificial Intelligence
Department of Computer Science

Semester: 2022 Spring

Instructor - Navneet Kaur