2

Handling Data & Datasets

First steps into Data Science

2.1
What are **Datasets?**

Data and Datasets

- Data is information in any form
- Features are attributes of data that hold information
- Data grouped together in a logical manner is called a dataset
- Datasets are a collection of examples used to train ML algorithms

Data in datasets can be of many forms

- Text data, also called a *corpus*
- Image data
- Audio data
- Video data
- Numeric data

The 2 main purposes of datasets

1. To train a model
2. To evaluate a model

2.2
Where do I find **Datasets?**

- Datasets can be gathered from a variety of sources, some common sources are:
  - Kaggle https://www.kaggle.com/datasets
  - Google Dataset Search https://datasetsearch.research.google.com/
  - Amazon Web Services Datasets https://registry.opendata.aws/

- You can even **create your own dataset**

2.3
Types of Data.

Quantitative Data

Data is called **quantitative** if it is measured by **numbers** and has a **magnitude**. These can be further broken down to:

- Discrete Data - Refers to numeric data that are **whole numbers** *(Discrete numbers == Whole Numbers)*
  - Example: Number of students in a class is 34 or 40
- Continuous Data - Refers to numeric data that can be expressed in a **decimal format**
  - Example: Weight, Time, Distance. Such as 75.3 kg

Qualitative or Categorical Data

Data is called **qualitative** if it is measured used to express categorical features (not numerical):

- Nominal Data - This is categorical data used to express names or labels which **do not have an order** or can be measured
  - Example: Male or Female, Country of Origin, etc.
- Ordinal Data - This is like nominal data but **has ordering** associated with it
  - Example: Income brackets, Size: Small < Medium < Large

2.4
Dataset **Splitting**.

Datasets are just a collection of vast amount of data. For us to use them **to train ML models we need to break them into subsets**.

Datasets can be split into 3 general categories

- Training set
- Validation set
- Test set

Training Set

- This is the largest subset
- It contains the data that will be used to **fit the model**
  - *Fit the model, refers to teaching the model*
- The labels help a model to come up with the weights/rules to learn the data type

Validation Set

- Validation set is used to **evaluate all the parameters** of the model
- This data helps identifying shortcomings such as
  - Overfitting
  - Underfitting

Test Set

- Test set should always contain data **unknown to the model**
- This set is used to test the accuracy of the model
- Other metrics other than accuracy can be used as well

Dataset split ratios

Depending on the size of your dataset and the task you are trying to achieve, the splits can be of various ratios.

For example, you can have train/val/test splits of

- 80/10/10
- 60/20/20
- 50/30/20

2.5
Data **Preprocessing**.

The problem with real-world data

- Majority of real-world data is highly susceptible to have missing and inconsistent data
- Data could also be noisy
  - Noisy is a term used to refer to meaningless features
  - It could be an extra feature, corrupted data, or distorted data

Noisy Data causes a lot of problems in ML

- Model might associate the noise to an outcome, which is false
- Noise may distort the underlying pattern the model could have learned
- Duplicate values may give an incorrect view to the overall data
- Outlier and inconsistent data can hampler model’s learning

Solution: Preprocess the Data

Data preprocessing improves the overall data quality. This means the model can be fitted better. We can broadly outline 3 steps of data-preprocessing:

1. Data Cleaning
2. Data Integration
3. Data Reduction / Dimension Reduction

Data Cleaning

Data cleaning is the first step where you remove the faulty data. We will go over three strategies for data cleaning:

1. Missing Values
2. Noisy Data
3. Removing Outliers

Data Cleaning: Handling Missing Values

- Ignore those tuples (datapoints)
  - Can be considered when the dataset is huge and the tuple in question has multiple missing values
  - Although this strategy depends on the context
- Fill in missing values
  - Use an average value to fill in missing ones
  - Assign randomized values

Data Cleaning: Handling Noisy Data

- Binning
  - Divide data into equal-sized bins so they can be dealt with individually
  - Data in a bin can be replaced by mean, median or boundary
- Regression
  - Smoothen noisy by fitting all the data points in a regression function

Data Cleaning: Removing Outliers

- Create clusters of data, where similar data points with be clustered together
- Values that do not lie in any cluster can be treated as outliers and removed

Data Integration

- Data Integration is referred to the process where data present in multiple sources are merged together into a single larger data store
- Data Integration is especially needed in real-world scenarios
  - Detecting presence of nodules from CT scan images
  - Only option is to integrate the images from multiple medical nodes to form a large dataset

Strategies of Data Integration

We can follow a general guideline when integrating data:

1. Schema Integration and Object Matching
    a. Different stores can be of different dimensions or formats
    b. Translate everything to one format
2. Decide which overlapping features should remain and what should be discarded
3. Remove redundant attributes from all data sources
4. Detect and resolve data value conflicts

Data Reduction / Dimension Reduction

- Often, the datasets might be too large for models
- It may also be the case we don’t need every feature to determine an outcome
- Therefore we prefer to reduce the **dimensionality** aka. Features of the data

Strategies of Dimensionality Reduction

We can follow a general guideline when integrating data:

1. Reduce the number of redundant features
    a. Ex: Do we need NID number to detect heart-attack?

…

Strategies of Dimensionality Reduction cont.

2. Data Compression

   a. We can use encoding technologies to decrease size
      i. Especially beneficial for audio, video, image data
      ii. Ensure the compression uses lossless reduction
   b. Discretization
      i. Turn continuous values into categories of ranges
         Ex: Put 12, 15, 18, 19 into ranges 10-15 and 16-20
   c. Attribute subset selection
      i. Select a subset of attributes/features that provide the most information

END.