Introduction to Data Science: SWE 227

Topics to be covered

- *1. Data.*
- *2. Data Science*
- *3. Data Handling*
- *4. Dataset*
- *5. Modeling*
- *6. Different Machine Learning Used Case*
- *7. Machine Learning:*
- *8. Deep Learning:*
- *9. Data Mining*
- *10. Misc.*

1. Data

Definitions

- ❑ *Units of information.*
- ❑ *Characteristics or information, usually numerical, that are collected through observation.*

Data + Context = Information

Stratification of Data:

Structured and Unstructured Data

Raw Data & Cleaned Data

What is Structured Data?

What is Unstructured Data?

What is Raw Data?

2. Data Science

Definitions

- ❑ *Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.*
- ❑ *“Fourth paradigm" of science (empirical, theoretical, computational and now data-driven)*
- ❑ ***Data science is an interdisciplinary field focused on extracting knowledge from data sets (The one that will be most suited for our cause)***

2. Data Science

The Data Science Process

The Data Science Process is similar to the scientific process - one of observation, model building, analysis and conclusion:

❑ *Ask questions*
❑ *Data Collection*
❑ *Data Exploration*
❑ *Data Modeling*
❑ *Data Analysis*
❑ *Visualization and Presentation of Results*

2. Data Science

What do data scientists do?

- ❑ Construction and curation of Datasets
- ❑ Exploratory Data Analysis
- ❑ Building and Improving Models
- ❑ Gaining Insights from Models
- ❑ Collaborating with Domain Experts to Create New Knowledge and Tech

Opportunities of Data Scientists

- ❑ High Paying Jobs
- ❑ Different Domains, countless opportunities. Finance, Biomedical, Marketing, Linguistics, Military, Astronomy and many more!
- ❑ **The sexy job in the next 10 years will be statisticians.**
  Hal Varian, Prof. Emeritus UC Berkeley Chief Economist, Google

3. Data Handling

Data Cleaning Techniques

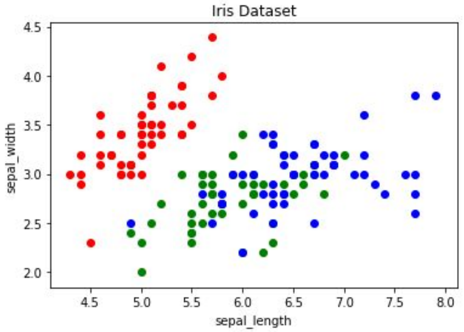
- ❑ Data Cleaning means removing “noisy” data samples.
- ❑ **The most boring and Important Step**
- ❑ Poor Data = Poor Models
- ❑ Primary Cleaning Constraints:
    - Range
    - Data-Type
    - Compulsory constraints
    - Cross-field examination
    - Unique Requirements
    - Regular Patterns
    - Accuracy

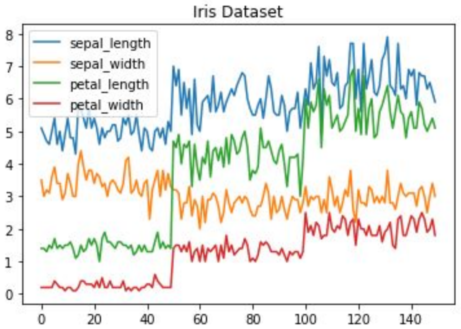
3. Data Handling

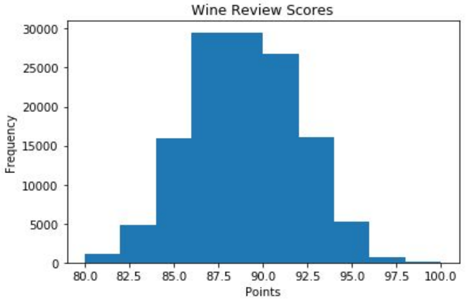
Data Visualization Techniques

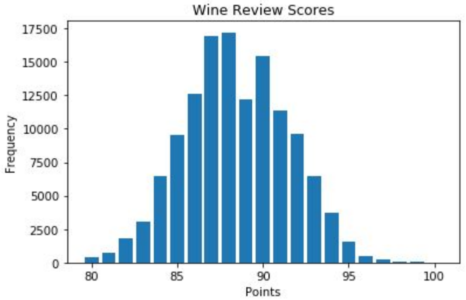
❑ Different types of Data Demand the use of Different Tools

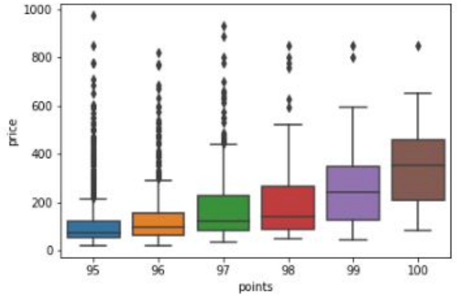
Some popular visualization techniques (CAN YOU NAME THEM?)

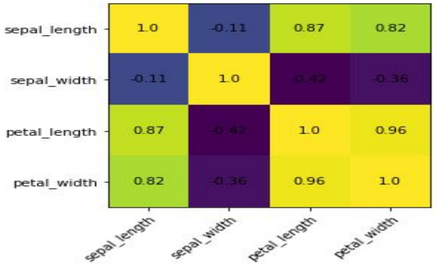


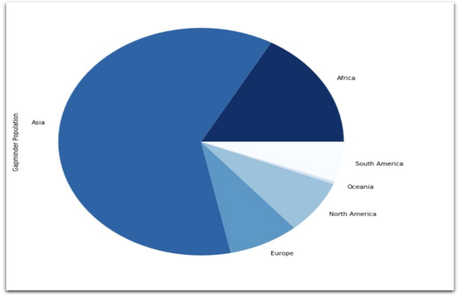


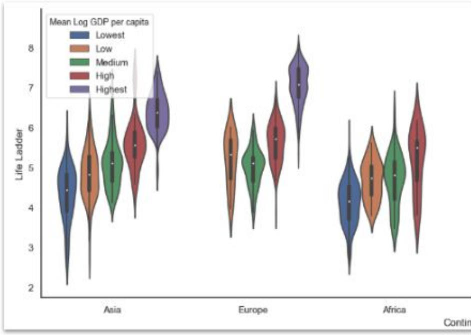












3. Data Handling

Data Visualization Tools in Python

- ❑ Matplotlib
- ❑ Pandas
- ❑ Seaborn
- ❑ Others

Other Data Visualization Tools

- ❑ Tableau
- ❑ Qlikview
- ❑ FusionCharts
- ❑ Highcharts
- ❑ Datawrapper
- ❑ Sisense
- ❑ Zoho Analytics
- ❑ Cluvio
- ❑ **Ms Excel**

4. Dataset

Some vocab

Dataset, for us: A dataset is a repository of well curated and cleaned data samples for analysis and modeling.
Benchmarking Dataset: A benchmark dataset is one that is used by researchers for testing their models for different tasks.
Annotator: The person who identifies/sorts the data samples from a dataset is called an annotator.
Metadata: Extra information present with the necessary data
Numerous different datasets have been developed.

For example, the list of leaf length of **all the trees** in the Sunderbans can be called a dataset. The persons who will be measuring and writing down the lengths are the annotators. If we add some extra information like ‘were birds found on the branches of this tree?’ against every tree, that will be regarded as a matadata.

4. Dataset

Data modalities

Image
Video
Text
Tabular Data
Others.

Example Dataset

❑ Imagenet is the most famous dataset for Image Classification.

❑ The dataset contains 14 million images.

❑ The images contain 21,000 different types of things or animals (leopard, electric ray, mountain tent, toilet tissue).

❑ The task is to develop a predictive model using this dataset that takes an input image and “classifies” it.

THIS IS CALLED AN IMAGE CLASSIFICATION PROBLEM. More on this later.

4. Dataset

Construction of a Dataset

- The construction of a dataset is a big task. It involves
  - Gathering domain knowledge
  - Correct annotation,
  - Cross checking the annotations,
  - Data cleaning,
  - Benchmark modeling
  - EDA (Exploratory Data Analysis)
  - Curation.
  - Publication

4. Dataset

Components of a Dataset

- ❑ **Train Data**: Train data or training data is the portion of the dataset that is used for training machine learning/deep learning models. For example, if there are 1,00,000 samples in the original dataset, we may use 70,000 or 80,000 (70% or 80%) for training the data.
- ❑ **Validation Data**: Validation data is the small portion of data that is used to evaluate the performance of the models at regular intervals. This sub-section might contain 10/20% of the total data.
- ❑ **Test Data**: Test data is the sacred sub-section from the whole data that the model is never exposed to. Once the training is over, the model is given these samples for a final evaluation of the model.

  The test data chunks are supposed to be coming from a slightly different distribution/ different data source compared to the training and validation set.

  For example, from a dataset for sleep staging might contain EEG data from 1000 patients. Say there exists multiple data samples from the same patients. In such a scenario, the test set should NEVER contain samples from patients whose samples are present in the training or validation set.

4. Dataset

Some Famous Datasets

- ❑ Image Datasets: Imagenet, Pascal VOC etc
- ❑ Video Datasets: Youtube Dataset, Aff-Wild etc
- ❑ Text Dataset: Twitter US Airline Sentiment Dataset, Large Movie Review Dataset etc.

Some Bangladeshi Datasets

- ❑ Bengali.AI Grapheme Recognition Dataset
- ❑ Bangla-lekha Isolated
- ❑ Google Bangla ASR Dataset
- ❑ NO CORPUS

Countless datasets can be found in many other data science web-sites. Samples:
https://lionbridge.ai/datasets/14-best-text-classification-datasets-for-machine-learning/
https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

5. Modeling

What is modeling?

In Data Science, modeling can be referred to the process of constructing a system (based on data) that is about to find patterns in the data to train itself for giving predictive answers on unseen data.

What’s a Good Model?

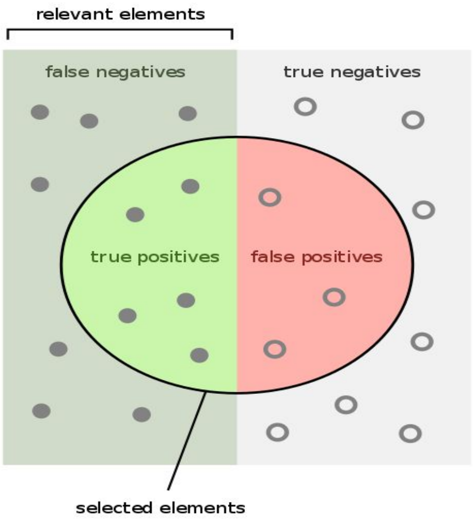
The model that “learns generalized patterns” from given data in an efficient way and is able to “perform” well on unseen data is called a good model.

Three Steps of Modeling

Training
Validation
Test
Can you recall what these mean?

5. Modeling

Performance Metrics



F1 Score: harmonic mean of the precision and recall.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

Precision: Ratio of correctly predicted positive observations to the total predicted positive observations. TP/TP+FP

Recall(Sensitivity, True Positive Rate): Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Recall = TP/TP+FN

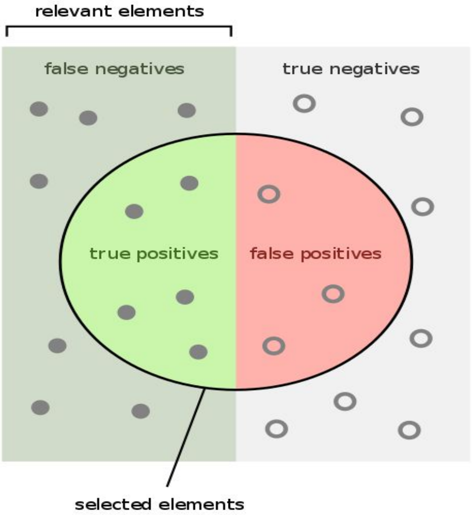
Specificity(True Negative Rate): TN/(TN+FP)

False Positive Rate: FPR = FP/(FP+TN)

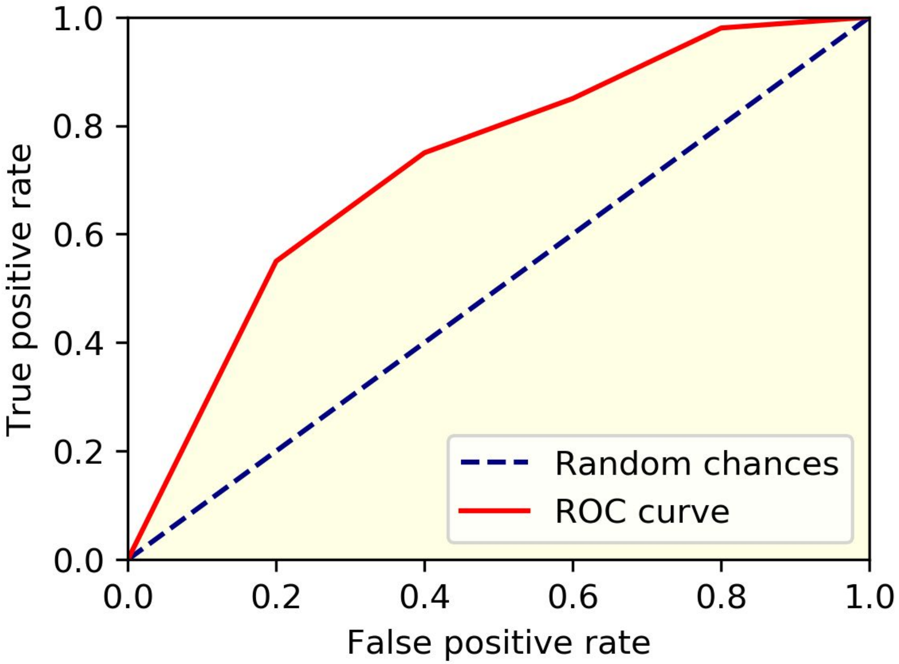
Accuracy: (TP+TN)/(T+N)

5. Modeling

Performance Metrics



AUC (Area Under Curve): AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). The ROC curve is obtained by plotting True positive rates against false positive rates for different instances of the same model.



6. Different Machine Learning Used Case

We can do EVERYTHING

- Common Variety:
  - Classification (Text Classification, Video Classification, Image Classification, Acoustic Scene Classification, Speaker Recognition)
  - Segmentation (Audio Segmentation, Video Segmentation, Image Segmentation (Semantic Segmentation, _ Segmentation)
  - Compression (Image Compression, Video Compression, Encoders)
- Special(?) Variety:
  - Named entity recognition
  - Image and video Captioning
  - Sentiment Analysis
  - Speech to Text
  - Text to Speech
  - Language Modeling.

7. Machine Learning

What is Machine Learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Long story short: It is different from Deterministic Programming

7. Machine Learning

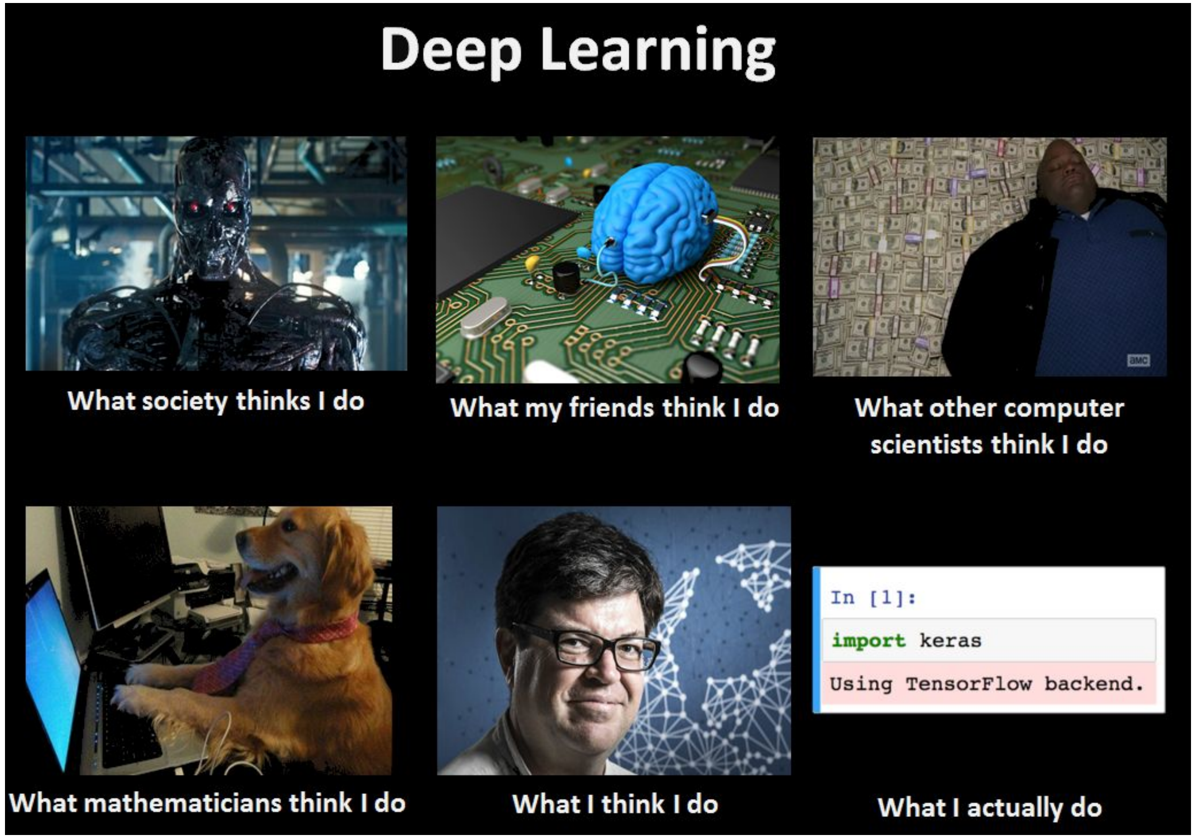
Different machine learning techniques

- ❏ Classifiers: (Bayesian, Maximum a posteriori, parameter estimation, decision tree, SVM, bag of words, N-gram models, association rules, nearest neighbor, locally weighted regression)
- ❏ Clustering (mixture models, k-means clustering, hierarchical clustering, distributional clustering).
- ❏ Ensemble techniques.

8. Deep Learning

What is Deep Learning?

❑ Definition of Deep Learning
❑ History of Deep Learning
❑ Possibilities and Variations
❑ Computational Constraints


Deep Learning
What society thinks I do
What my friends think I do
What other computer scientists think I do
What mathematicians think I do
What I think I do
What I actually do

```
In [1]:
import keras
Using TensorFlow backend.
```

- Will be discussed in great detail! Stay Tuned :3