Lecture 2

BU.330.775 Machine Learning

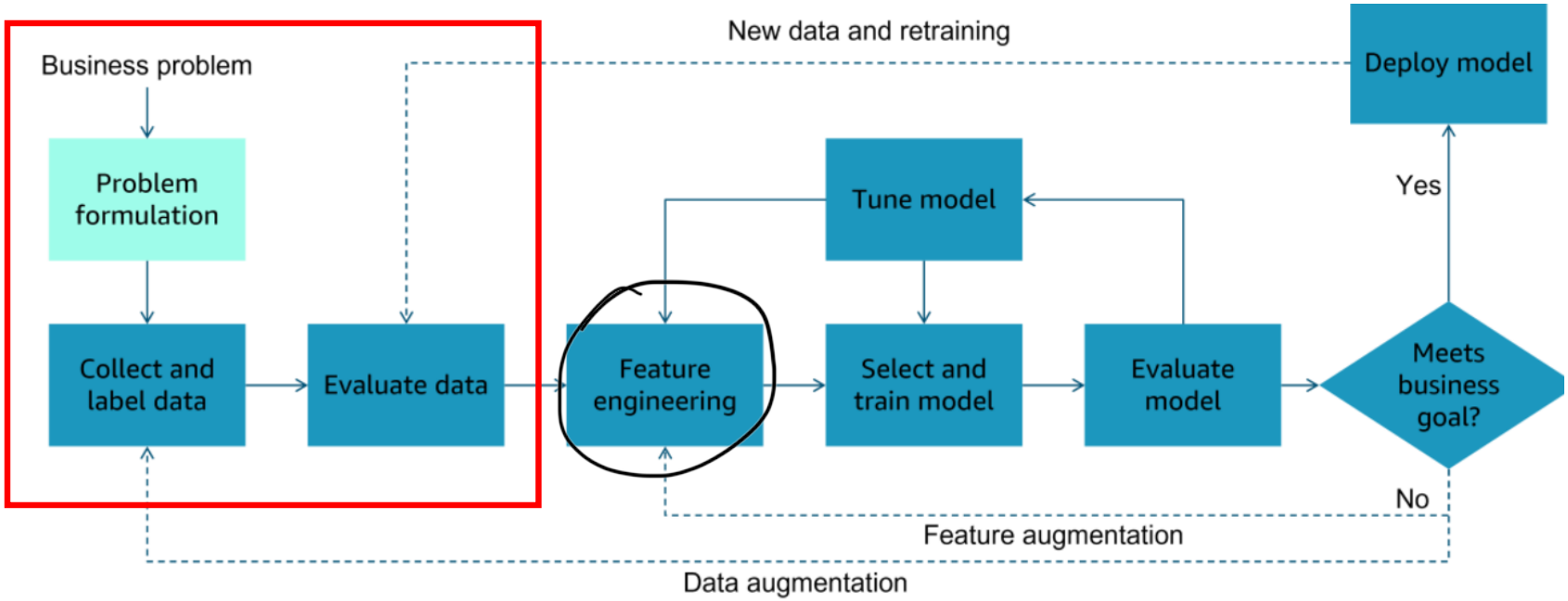
Minghong Xu, PhD.
Associate Professor

Today's Agenda

- Data preparation and preprocessing
- Exploratory data analysis
- Hands-on practice on California housing data

- Reminder: homework 1 due before class

Machine Learning Pipeline



Problem Formulation

- *Business objective: how to use? benefit?*
- *Is it machine learning problem?*
- *Supervised? Unsupervised?*
- *Target to predict?*
- *Data?*
- *Current solution?*

Example: Credit Card Fraud

- Business objective:
- Machine learning:
- Supervised/unsupervised:
- Target:
- Data:

Data Preparation

Data Terminologies

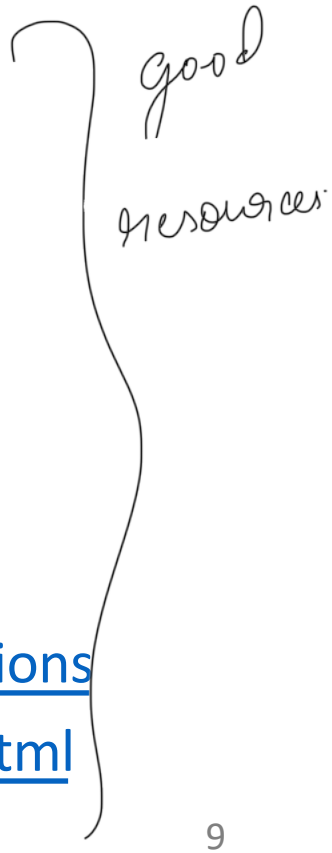
- Observation: instance/example/data point
  - A single record that represents one example of the problem being studied
- Feature: each piece of information included in the data representation
  - E.g. age, weight, height, diet, exercise or not…
  - Continuous feature
  - Categorical/discrete feature, not numerical
- Target: output variable

Questions to Consider

- *How much data do you have?*
- *Where is it?*
- *Do you have access?*
- *How to bring all data into one centralized repository?*
- *Is your data representative?*

Popular Open Data Repositories

- OpenML: https://openml.org
- Kaggle: https://www.kaggle.com
- PapersWithCode: https://paperswithcode.com/datasets
- UC Irvine: https://archive.ics.uci.edu
- AWS: https://registry.opendata.aws
- TensorFlow: https://www.tensorflow.org/datasets
- Nasdaq: https://data.nasdaq.com/institutional-investors
- World Health Organization: https://www.who.int/data/collections
- U.S. Census Bureau: https://www.census.gov/data/datasets.html



Some Data Portals

- http://dataportals.org
- https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

- Social media posts:
- https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public
- https://www.reddit.com/r/datasets/

Pipeline

- *How much data do you have?*
- *Where is it?*
- *Access?*
- ***How to bring all data into one centralized repository?***

- Data pipeline: a sequence of data processing components/steps
- A popular one: Extract, Transform, Load (ETL)

Data Evaluation

- *How much data do you have?*
- *Where is it?*
- *Access?*
- *How to bring all data into one centralized repository?*
- ***Is your data representative? Good quality?***

- Data evaluation!

Evaluation Tasks

- Data format
  - Common file formats for machine learning: csv, [redacted], json, etc.
- Data types
  - Structured, unstructured
- Descriptive statistics
  - Statistical: max, min, mean, median, std, var, etc.
  - Correlations
- Data visualization

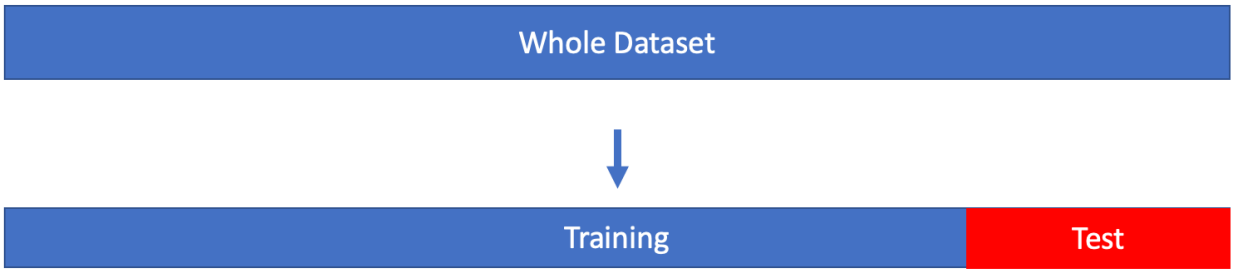
image, audio, video

JPEG PNG | MP3 WAV | MP4 AVI MKV

Numeric, Categorical, Text, Time series, Image audio, video.

Data Preprocessing

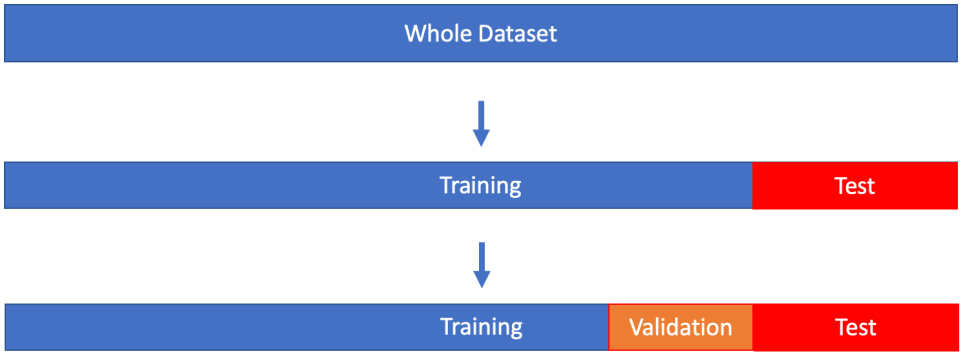
Training and Testing



- Most machine learning models have two stages: training and testing
- Training data is used to train the model, to learn patterns
- Testing data is used to evaluate the model's performance after training
- Training and testing sampling: a process that splits a data set into two parts, a training set and a testing set

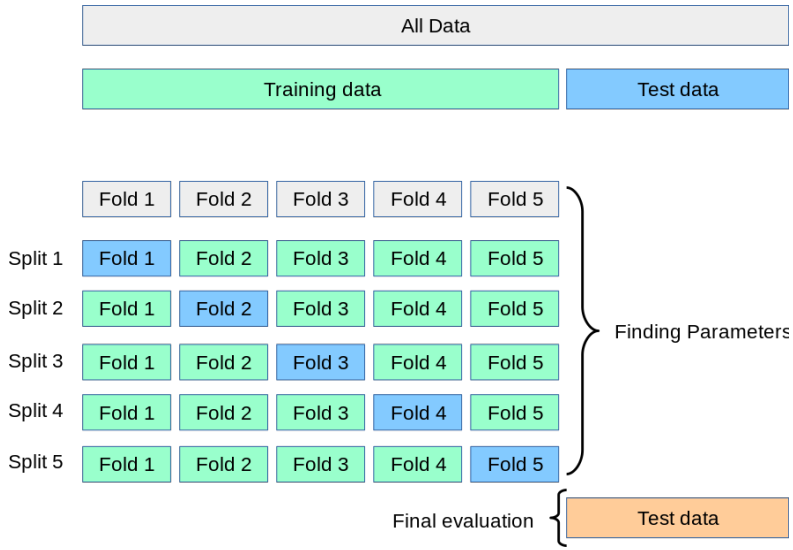
Validation

- Similar to preparing an exam
  - Do lots of exercises => Training
  - Mock exam => Validation
  - Final exam => Test



Cross-Validation

- Machine learning models need sufficient training, thus sufficient usage of data
- Cross-validation is useful for small datasets



https://scikit-learn.org/stable/modules/cross_validation.html

Sampling Strategy 1: Stratifying

- Divide subjects into subgroups called strata based on characteristics that they share
  - Race
  - Gender
  - Educational attainment
  - Etc.
- Once divided, each subgroup is randomly sampled
- Also called **Stratified Sampling**

Notes

to ensure each group (strata) has representation in sampling in equal proportion in test and train data set

Sampling Strategy 2: Shuffling

- Process of randomly rearranging the order of the rows/instances in a dataset
- Important especially when splitting data into training and testing sets or cross-validation
- Help to ensure that the model trained is not biased by the order of the data
- When not to shuffle: time-series

Handling Categorical Data

- Most machine learning algorithms prefer to work with numbers

- **Label encoding**: each category into a unique integer
  - When: if categorical variable has a natural order, e.g. low, medium, high
  - Also called ordinal encoding

  OrdinalEncoder

- **One-hot encoding**: convert into a set of binary (0/1) columns, one for each category
  - When: there is no inherent order

  OnehotEncoder

Handling Missing Value

- Most machine learning algorithms can not work with missing features
  - Need to take care before you can run the model

- **Three options:**
- Get rid of the data points with missing values
- Get rid of the attributes with missing values
- Imputation: set the missing values to some value, e.g. zero, mean, median…

Examples

- Transaction or account activity analysis
  - Zero can be used to indicate no activity/transaction/balance
- Temperatures, blood pressure readings, other (usually) normal distributed cases
  - Mean can be used to represent average case
- House price prediction, outliers are significant or distribution is skewed
  - Median can be used to represent typical values
- *What are the potential issues for simply dropping?*

Note

Simple Impute (Strategy = median)

Handling Outliers

- Outliers may skew the models, longer training, less accurate
- Identifying outliers can be a difficult task
  - Statistical methods: Z-scores, quartiles
  - Visualizations: box plot, scatter plot, histogram
  - Machine learning: clustering, isolation forest
- *Once identified, drop or keep?*
- General "drop" cases:
  - Data errors
  - Sensitive models such as regressions

Notes

273 consider as outlier

IQR: $Q_3 - Q_1$

outlier $< Q_1 - 1.5 IQR$

outlier $> Q_3 + 1.5 IQR$

① KMean clustering

② DBSCAN

Feature Scaling

- Most machine learning algorithms do not perform well when attributes have very different scales
  - Results may be dominated by the large scale

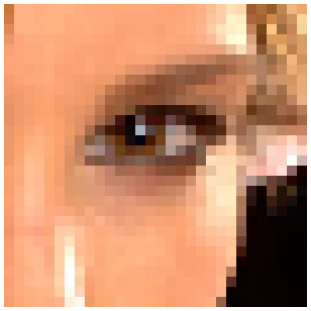
- **Min-max scaling/normalization**: transform into a given range from 0 to 1 — for data that doesn't follow Gaussian distribution
- **Standardization**: transform into standard normal distribution
  - Not a specific range, less affected by outliers — for data that follow Gaussian distribution.
- It is important to apply exactly the same transformation to the training set and the test set!

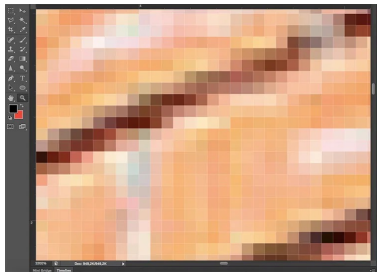
Unstructured Data: Images

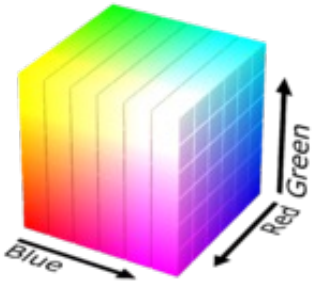
Image Data

- *What is an image?*
- A bunch of numbers between 0 to 255
- A 3-dimensional array

(0, 0, 0) black
(255, 255, 255) white



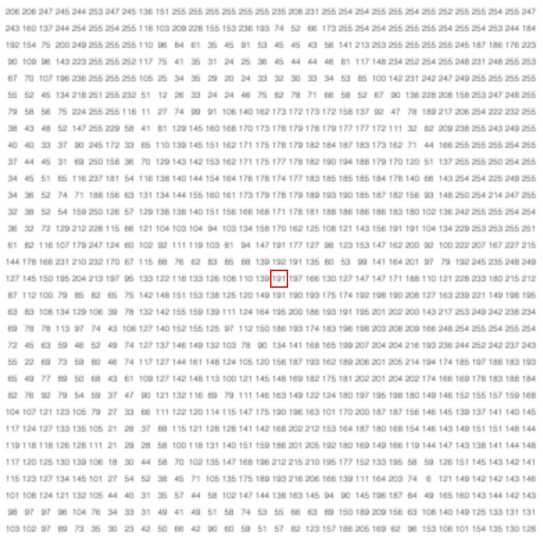




- https://convertacolor.com

Grayscale Image

- A grayscale (or graylevel) image: **the only colors are shades of gray**
  - One number for each pixel
  - Black/white image



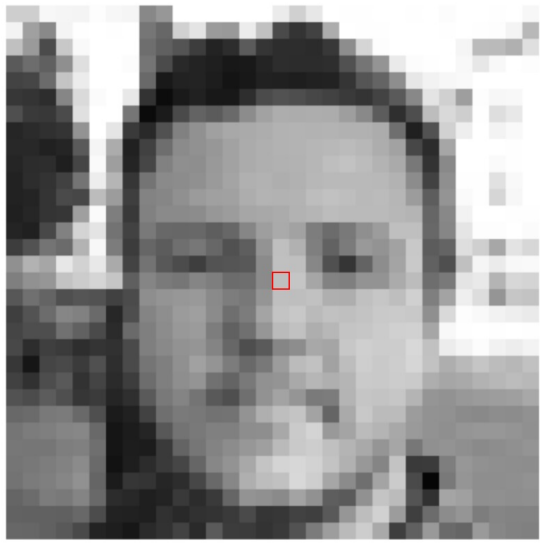


Image Representations

- https://www.boxentriq.com/code-breaking/pixel-values-extractor

- Size: width * height * # color channels
  - 3 color channels for colored images
  - 1 color channel (grayscale) for black/white images
- That's how many numbers we use to represent an image

- *How can we know the size of an image?*

ImageNet Dataset

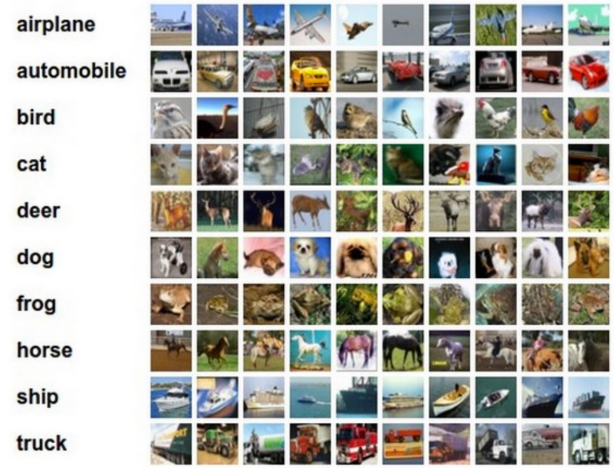
- https://www.image-net.org
- More than 14 million images, 21000 categories
- Hand-annotated
- Initiated by AI Researcher, Dr. Fei-Fei Li, in 2006
  - Expand and improve the data to train AI models
- Amazon Mechanical Turk to help with classification
  - Average: 50 images per minute

MNIST Dataset

- Modified National Institute of Standards and Technology Database
- A large database of handwritten digits
  - Commonly used for training various image processing systems
  - Taken from American high school students
  - Black/white images (grayscale levels), 28x28 pixel
  - 60,000 training images and 10,000 testing images

CIFAR-10 Dataset

- Canadian Institute For Advanced Research, 10 classes
  - Airplane, car, bird, cat, deer, dog, frog, horse, ship, truck
- Commonly used to train machine learning and computer vision algorithms
- 60000 images, 6000 images per class
- Images are 32x32x3 in size
  - width*height*3 color channels



Lab 2: California Housing data

- Pandas package
  - Derived from “Panel Data”
  - Handle and analyze structured data, such as tabular data, time-series, etc.
- Sklearn package
  - Originally scikit-learn, toolkit for scientific computing focus on machine learning
  - Library for machine learning
  - Built on top of NumPy, SciPy and matplotlib
- **Note that** we follow lecture notes sequence in this exercise, the actual preprocessing sequence in your own project might be different

Python Indentation

- To nest the code blocks
  - Function definition
  - Conditional statements
  - Loops
- Use “tab” not multiple “space”
- Auto controlled in Google Colab
- Sometimes you need to manually remove/control it

Next Week

- Supervised machine learning models – part I
- Model training, regularization, and evaluation
- Homework 2 due before class

References

- AWS Academy Machine Learning Foundations