Introduction to Machine Learning

Week 10 Day 01

DS 3000 – Foundations of Data Science

1

Reminders

HW 6

Thursday, November 7

FP3

Tuesday, November 12

2

Outline

Intro to Machine Learning (ML)

Types of ML Tasks

Training and Testing

Classification Case Study

3

Data Science

The interdisciplinary study and practice of computationally extracting meaningful insights from data

Three components:

- **Exploration** ➔ identifying patterns in data (messing around)
- **Prediction** ➔ making informed guesses
- **Inference** ➔ quantifying our degree of certainty

4

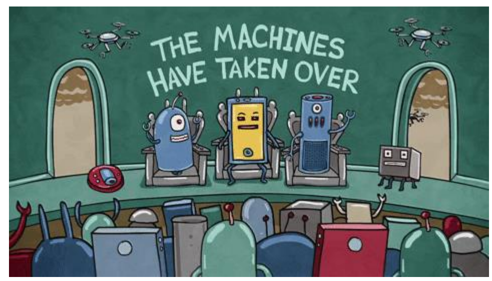
Machine Learning (ML)

The study of computer programs (algorithms) that can **learn by example**

Design predictive algorithms that learn from data

Replace humans in critical tasks

Subset of Artificial Intelligence (AI)



5

Machine Learning in Real Life

https://www.youtube.com/watch?v=z4K2F_OALPQ

6

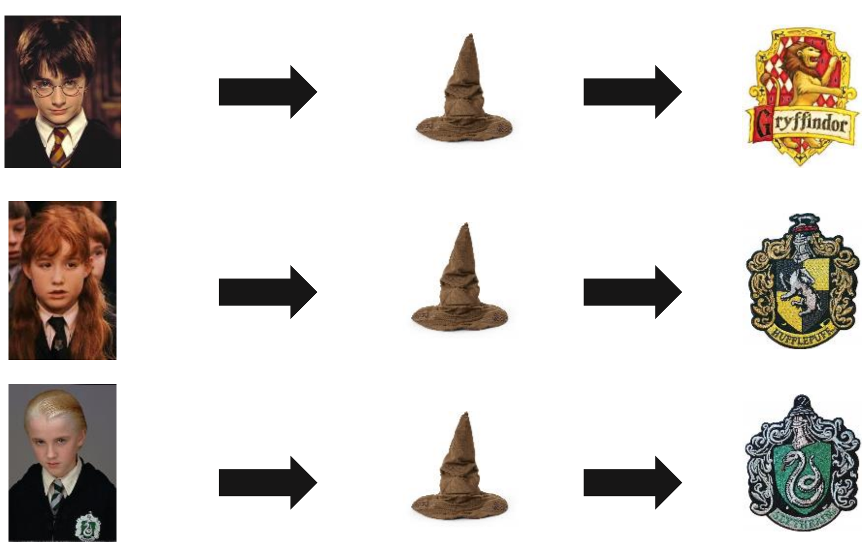
Sorting Hat as a Classifier

Sorting Hat learns from previous students
- Courage
- Hard Work
- Intelligence
- Ambition

7

Sorting Hat as a Classifier

Student **Classifier** **House**



8

Predictions with ML

Improve weather forecasting to save lives, minimize injuries and property damage

Improve cancer diagnoses and treatment regimens to save lives

Improve business forecasts to maximize profits and secure people's jobs

Detect fraudulent credit-card purchases and insurance claims

9

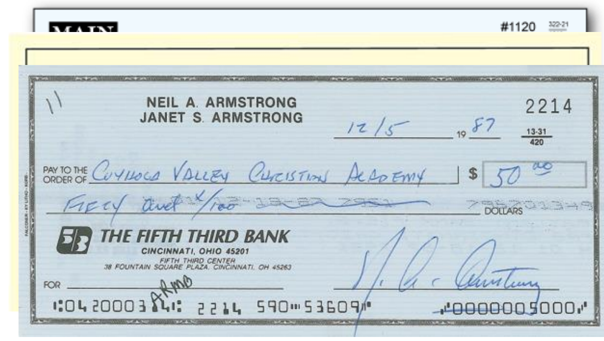
Popular Machine Learning Applications

Anomaly detection	Data mining social media (like Facebook, Twitter, LinkedIn)	Predict mortgage loan defaults
Chatbots	Detecting objects in scenes	Natural language translation (English to Spanish, French to Japanese, etc.)
Classifying emails as spam or not spam	Detecting patterns in data	Recommender systems ("people who bought this product also bought…")
Classifying news articles as sports, financial, politics, etc.	Diagnostic medicine	Self-Driving cars (more generally, autonomous vehicles)
Computer vision and image classification	Facial recognition	Sentiment analysis (like classifying movie reviews as positive, negative or neutral)
Credit-card fraud detection	Handwriting recognition	Spam filtering
Customer churn prediction	Insurance fraud detection	Time series predictions like stock-price forecasting and weather forecasting
Data compression	Intrusion detection in computer networks	Voice recognition
Data exploration	Marketing: Divide customers into clusters	

10

Machine Learning (ML)

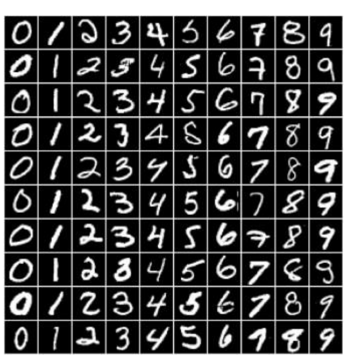
How do check scanners work? How do they know the dollar amount?



11

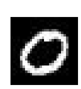
Machine Learning (ML)

ML algorithms can generalize from existing examples of a task



0 1 2 3 4 5 6 7 8 9

digit = 2

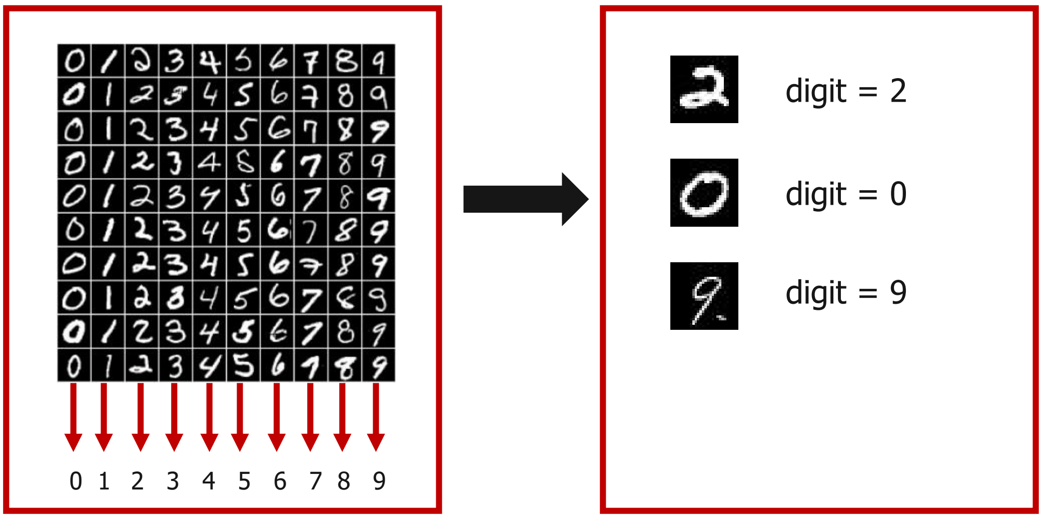


digit = 0

digit = 9

12

Machine Learning (ML)



Training **Testing**

13

Machine Learning (ML)

The process involves two main phases:

Training

Learning from the data and fitting a model

Testing

Estimating how well your model has been trained

How well can you generalize to new datasets?

14

ML Workflow

Representation → **Evaluation** → **Optimization**

- ✓ Preprocess data
- ✓ Choose features
- ✓ Choose algorithm(s)

- ✓ Decide on criteria to distinguish good vs. bad algorithms

- ✓ Choose the settings/parameters to maximize the model fit

15

Feature Representation/Extraction

The process of representing raw data in a meaningful way for ML tasks

Need to quantify the properties of the data

These are the variables based on which you will make predictions

Known as **features**, predictors, or attributes (sometimes IVs too)

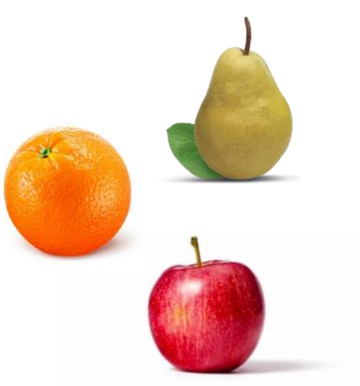
16

Feature Representation/Extraction

	Raw data	Feature	Value	Feature representation
Product Review (Text)	It is really comfy, even more comfortable than the Vive or the Rift!	# of positive words # of negative words Exclamation	2 0 yes	List of positive and negative words
Picture				A matrix of color values (pixels)
Object		**Feature** width height weight color score	**Value** 2.95″ 2.71″ 14oz 0.53	A set of attribute values

17

ICA5: Fruits Dataset



Features

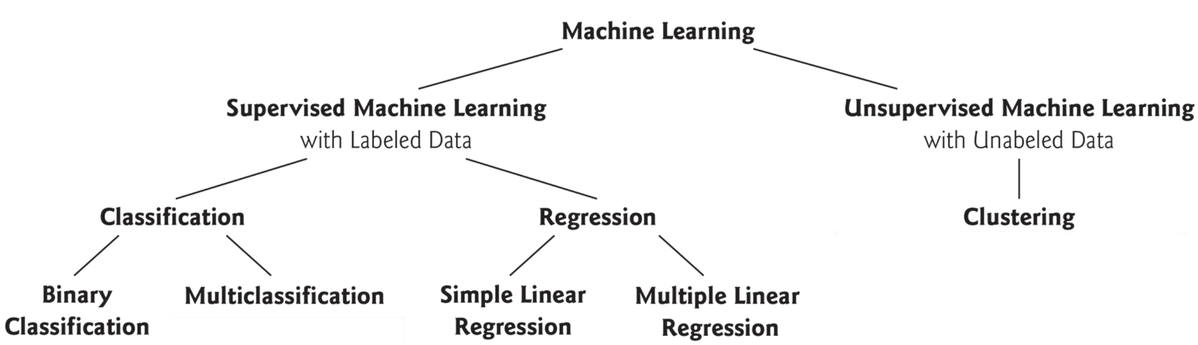
- Width
- Height
- Weight
- Dominant color

18

Types of ML Tasks

19

Two Types of ML Tasks



20

Two Types of ML Tasks

Supervised Learning

Learn to predict target values from labeled data

- Classification (target values are categorical/discrete classes)
- Regression (target values are numeric/continuous values)

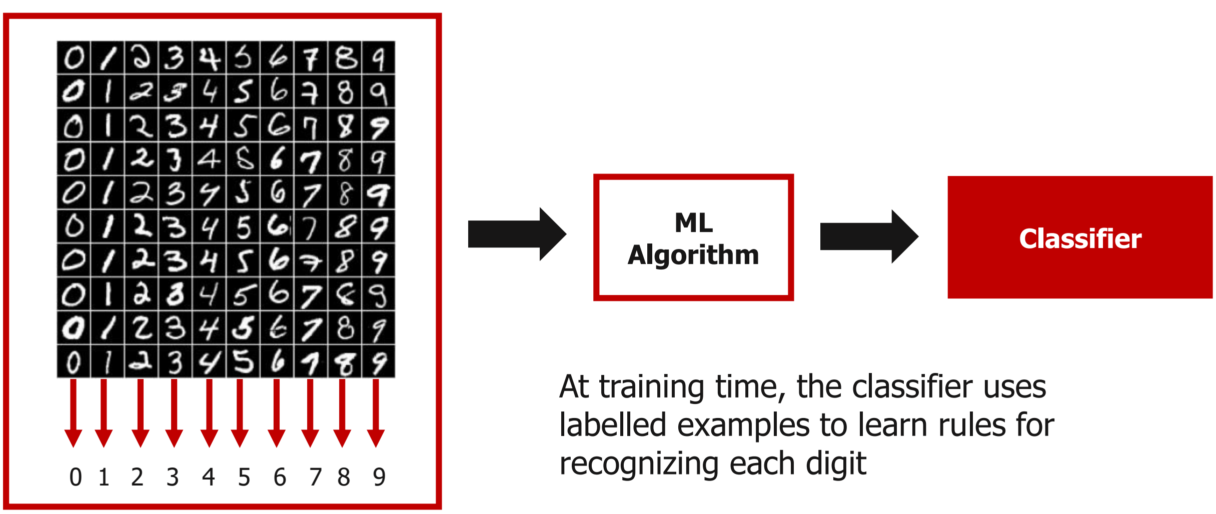
Unsupervised Learning

Find structure in unlabeled data

- Clustering (find groups of similar instances in the data)

21

Supervised Learning: Training



Training data

At training time, the classifier uses labelled examples to learn rules for recognizing each digit

22

Supervised Learning: Testing

Classifier

digit = 2

Test data input

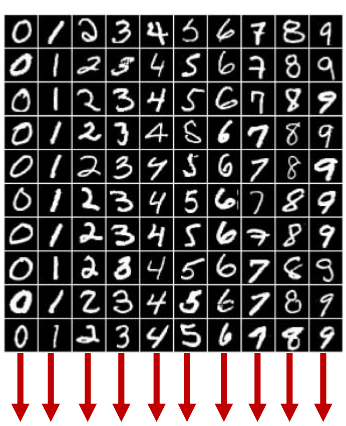
Model based on training data

Predicted label

After training, at prediction time, the trained model is used to predict the digit label for new instances using the learned rules.

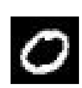
23

Supervised Learning



Training

digit = 2



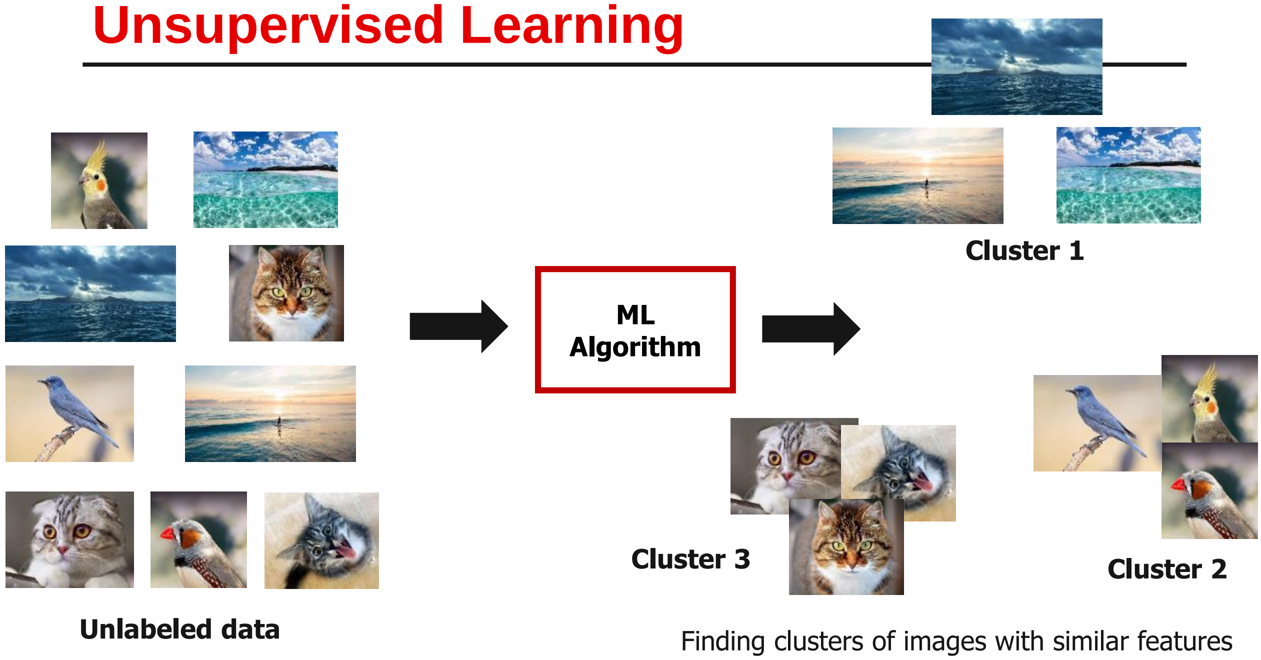
digit = 0

digit = 9

Testing

24

Unsupervised Learning



Finding clusters of images with similar features

25

Training and Testing

Ideally, use two different data sets (one for training and one for testing)

Training and test sets are assumed to have been sampled independently from an infinite population

Never test your algorithm on your training data

What if you have just one dataset?

- Percentage-split
- Cross-validation

26

Training and Testing

Percentage-split method:

Involves randomly dividing the dataset into training and test sets

A certain percentage is used for each

- 75% training & 25% is typical (default in Sci-kit Learn)
- 70/30 or 80/20 are common too

27

Steps in a ML Case Study

1. Load the dataset
2. Explore the data with pandas and visualizations
3. Transform your data (variable coding, normalization, etc.)
4. Split the data for training and testing
5. Create the model
6. Train and test the model
7. Tune the model and evaluate its accuracy
8. Make predictions on live data that the model hasn’t seen before

28

Supervised Learning: Classification

29

Supervised Learning

You train machine-learning models on datasets that consist of rows and columns.

Each row represents a **data *sample***.

Each column represents a **feature** of that sample.

In supervised machine learning, each sample has an **associated label called a *target*** (like “dog” or “cat”).

This is the **value you’re trying to predict for new data** that you present to your models.

30

Fruits Dataset

weight	width	height	color_R	color_G	color_B	fruit
4.3	6.2	7.2	0.5	0.16	0.03	apple
6.9	6.3	7.7	0.88	0.76	0.36	apple
3.1	7.2	7.1	0.71	0.97	0.52	orange
5.8	6.5	7.1	0.37	0.34	0.58	pear
4.4	8	6.4	0.49	0.09	0.03	orange
6.9	6	8.2	0.75	0.94	0.84	orange
3.8	8.1	7.3	0.21	0.37	0.96	pear
6.9	7.6	7.8	0.01	0.43	0.32	apple
6.2	6.5	6.1	0.08	0.84	0.11	orange
5.5	6.3	6.1	0.84	0.5	0.93	pear
4	7.7	7.1	0.53	0.02	0.66	apple

31

Classification

Classification algorithms predict the discrete classes (categories) to which samples belong

Binary classification uses two classes

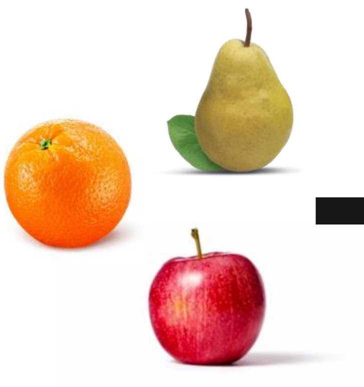
e.g., “spam” or “not spam” in an email classification application

Multi-classification uses more than two classes

e.g., the 10 classes, 0 through 9, in the Digits dataset.

32

Distinguishing between Different Types of Fruit



Classifier

Apple

Orange

Pear

33

Fruits Dataset

Each row represents a sample (a fruit)

weight	width	height	color_R	color_G	color_B	fruit
4.3	6.2	7.2	0.5	0.16	0.03	apple
6.9	6.3	7.7	0.88	0.76	0.36	apple
3.1	7.2	7.1	0.71	0.97	0.52	orange
5.8	6.5	7.1	0.37	0.34	0.58	pear
4.4	8	6.4	0.49	0.09	0.03	orange
6.9	6	8.2	0.75	0.94	0.84	orange
3.8	8.1	7.3	0.21	0.37	0.96	pear
6.9	7.6	7.8	0.01	0.43	0.32	apple
6.2	6.5	6.1	0.08	0.84	0.11	orange
5.5	6.3	6.1	0.84	0.5	0.93	pear
4	7.7	7.1	0.53	0.02	0.66	apple

34

Fruits Dataset

Columns represent **features** of each sample

Conventionally the last column is the **target** (label)

Features | Target

weight	width	height	color_R	color_G	color_B	fruit
6.2	8	8.1	0.5	0.62	0.67	apple
6.1	7.5	6.6	0.37	0.97	0.06	apple
4.7	7.5	6.2	0.23	0.4	0.54	orange
5.8	7	7.3	0.1	0.51	0.34	pear
3.1	6.8	6.1	0.47	0.55	0.27	orange
4.5	7.8	6.6	0.34	0.78	0.19	orange
3.4	7	6.1	0.15	0.47	0.37	pear
6.6	6	6.6	0.15	0.32	0.7	apple
6	6	7.8	0.31	0.29	0.48	orange
5.8	7.6	7.3	0.45	0.98	0.15	pear
6.5	6.5	6.4	0.32	0.46	0.29	apple

35

36

k-Nearest Neighbors

37

Birds of a feather flock together

38

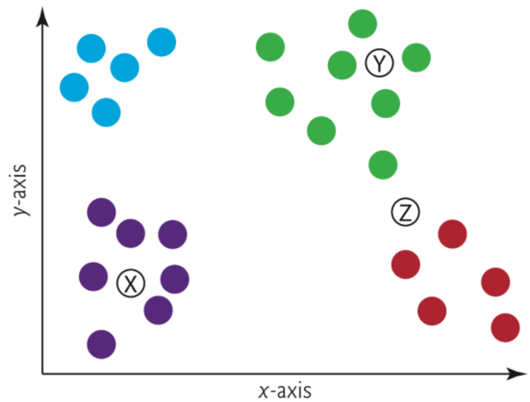
k-Nearest Neighbors

Predict a sample's class by looking at the ***k* training samples nearest in "distance"** to the **sample**

Filled dots represent four distinct classes A (blue), B (green), C (red) and D (purple)

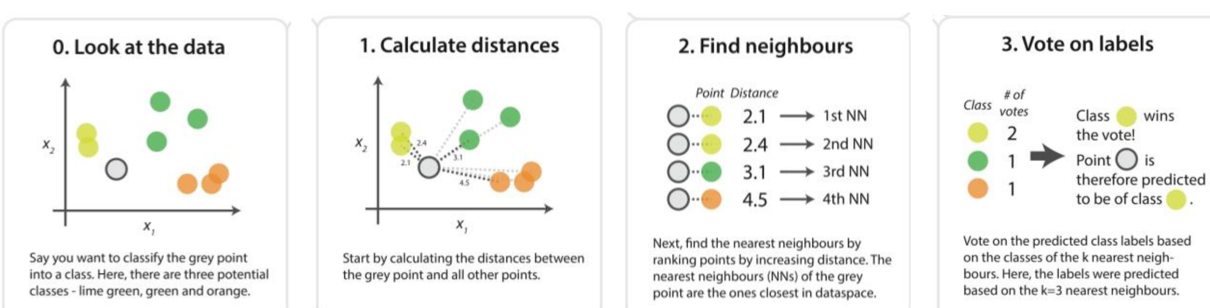
Class with the most "votes" wins

- **Odd *k* value avoids ties** — there's never an equal number of votes



39

k-Nearest Neighbors Algorithm



40

k-Nearest Neighbors Algorithm

Given a training set with features and labels, and given a new instance to be classified:

1. Find the **k-**most similar instances to the new sample in the training set
    - based on distance between the new sample and instances

2. Get the labels of the **k-**most similar instances in the training set

3. Predict the label for the new sample by combining the labels of the **k-**most similar instances
    e.g. simple majority vote

41