10-301/601: Introduction to Machine Learning Lecture 1 – Problem Formulation & Notation

Henry Chai

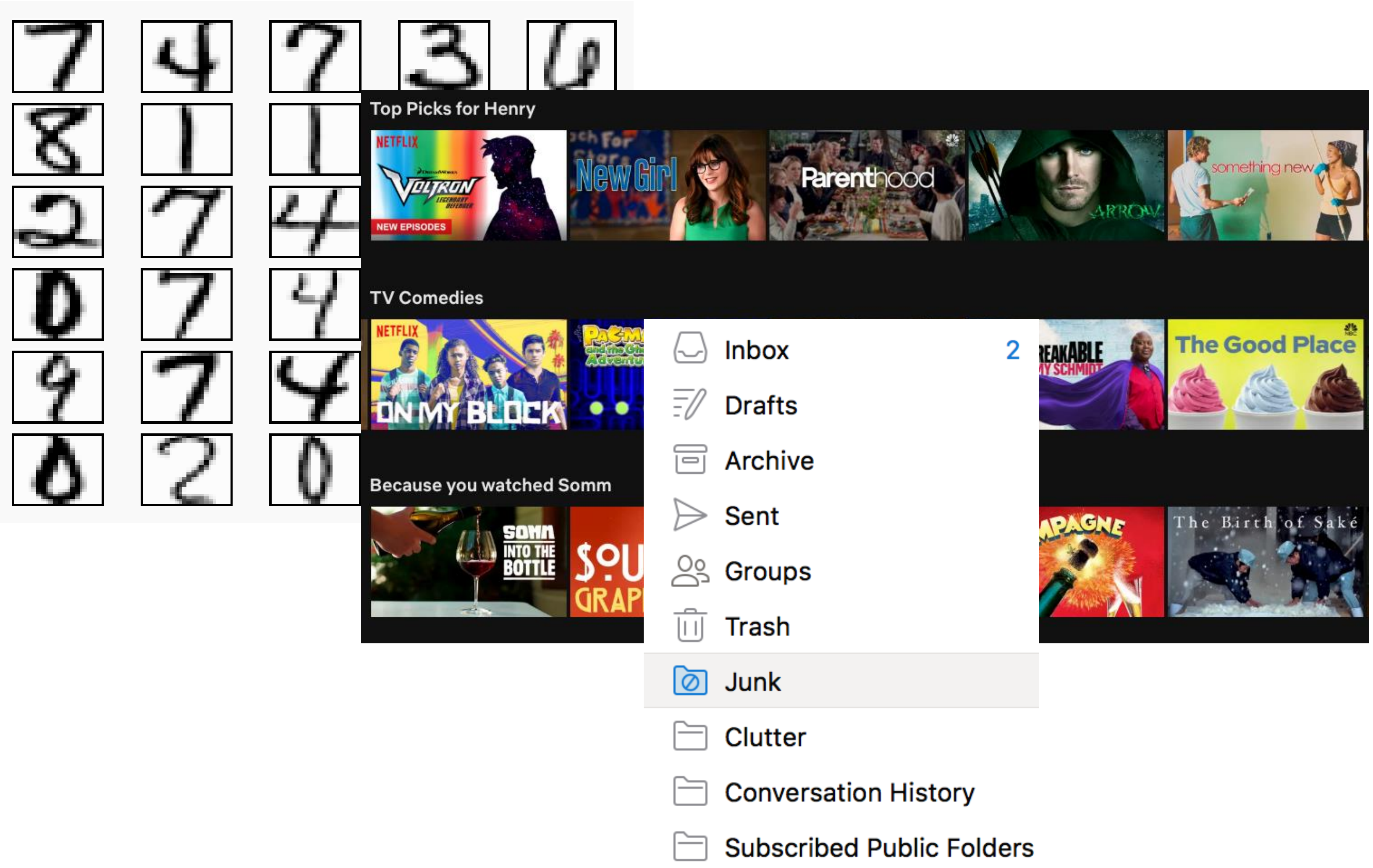
5/17/22

Front Matter

- Announcements:
  - HW1 released 5/17 (today!), due 5/24 at 1 PM
  - Recitation 1 on 5/19: review of prerequisite material
  - General advice for the summer:
    - Start HWs early!
    - Go to office hours! Starting tomorrow, 5/18
- Recommended Readings:
  - None

What is Machine Learning?

Machine Learning (Then)

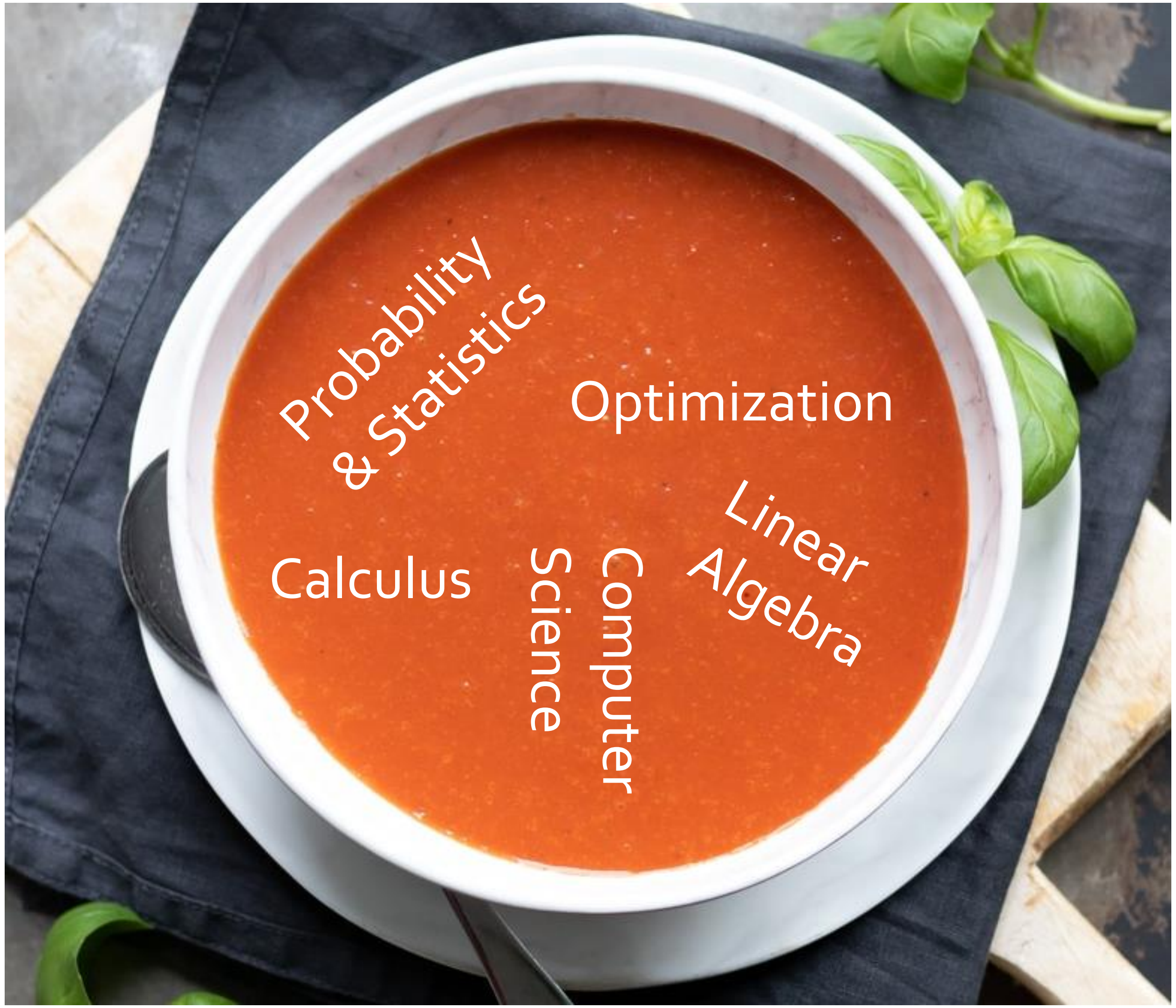


Machine Learning (Now)

Premise of Machine Learning

- There exists some pattern/behavior of interest
- The pattern/behavior is difficult to describe
- There is data
- Use data to “learn” the pattern

What is Machine Learning?



Source: https://veggiedesserts.com/easy-tomato-soup-recipe/

Things Machine Learning Isn’t

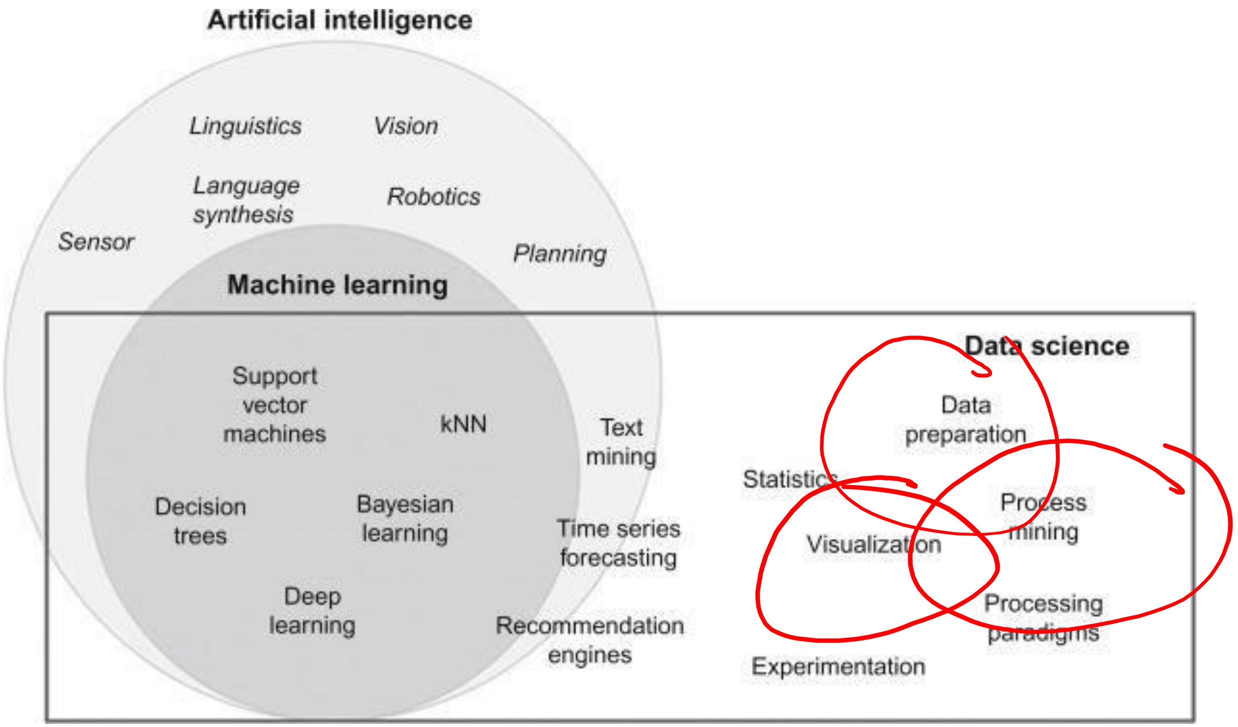
- Artificial intelligence
- Data science

Things Machine Learning Isn’t

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science

Things Machine Learning Isn’t

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data



Source: https://www.sciencedirect.com/topics/physics-and-astronomy/artificial-intelligence

What is ~~Machine Learning~~ 10-301/601?

Learning Paradigms:

What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Theoretical Foundations:

What principles guide learning?

□ probabilistic
□ information theoretic
□ evolutionary search
□ ML as optimization

Problem Formulation:

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real-valued	Regression
ordering	Ranking
sequence	Structured Prediction

Application Areas

Key challenges?

NLP, Speech, Computer Vision, Robotics, Medicine, Search

Facets of Building ML Systems:

How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

Big Ideas in ML:

Which are the ideas driving development of the field?

○ *inductive bias*
○ *generalization / overfitting*
○ *bias-variance decomposition*
○ *generative vs. discriminative*
○ *deep nets, graphical models*
○ *PAC learning*
○ *distant rewards*

What is ~~Machine Learning~~ 10-301/601?

- Supervised Models
  - Decision Trees
  - KNN
  - Naïve Bayes
  - Perceptron
  - Logistic Regression
  - SVMs
  - Linear Regression
  - Neural Networks
- Unsupervised Models
  - K-means
  - GMMs
  - PCA
- Graphical Models
  - Bayesian Networks
  - HMMs
- Learning Theory
- Reinforcement Learning
- Important Concepts
  - Feature Engineering and Kernels
  - Regularization and Overfitting
  - Experimental Design
  - Ensemble Methods

Defining a Machine Learning Task (Mitchell, 97)

- A computer program **learns** if its *performance*, *P*, at some *task*, *T*, improves with *experience*, *E*.
- Three components
  - Task, T
  - Performance metric, P
  - Experience, E

Defining a Machine Learning Task: Example

- Learning to approve loans/lines of credit

- Three components
  - Task, T

    Decide whether to extend a loan
  - Performance metric, P

    # of people who "default" on their loan
  - Experience, E

    Interviews w/ loan officers

Defining a Machine Learning Task: Example

- Learning to approve loans/lines of credit

- Three components
  - Task, T

    Predict the probability that someone defaults on their loan
  - Performance metric, P

    Accuracy over 10 years
  - Experience, E

    Historical records on loan defaults

Things Machine Learning Isn’t

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral?

⚠ When survey is active, respond at **pollev.com/301601polls**

Lecture 1 Polls

0 done

0 underway

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

When poll is active, respond at **pollev.com/301601polls**

Do you agree or disagree with the following sentence: "Because machine learning uses algorithms, math and data, it is inherently neutral or impartial."

Agree

Unsure

Disagree

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Things Machine Learning Isn’t

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

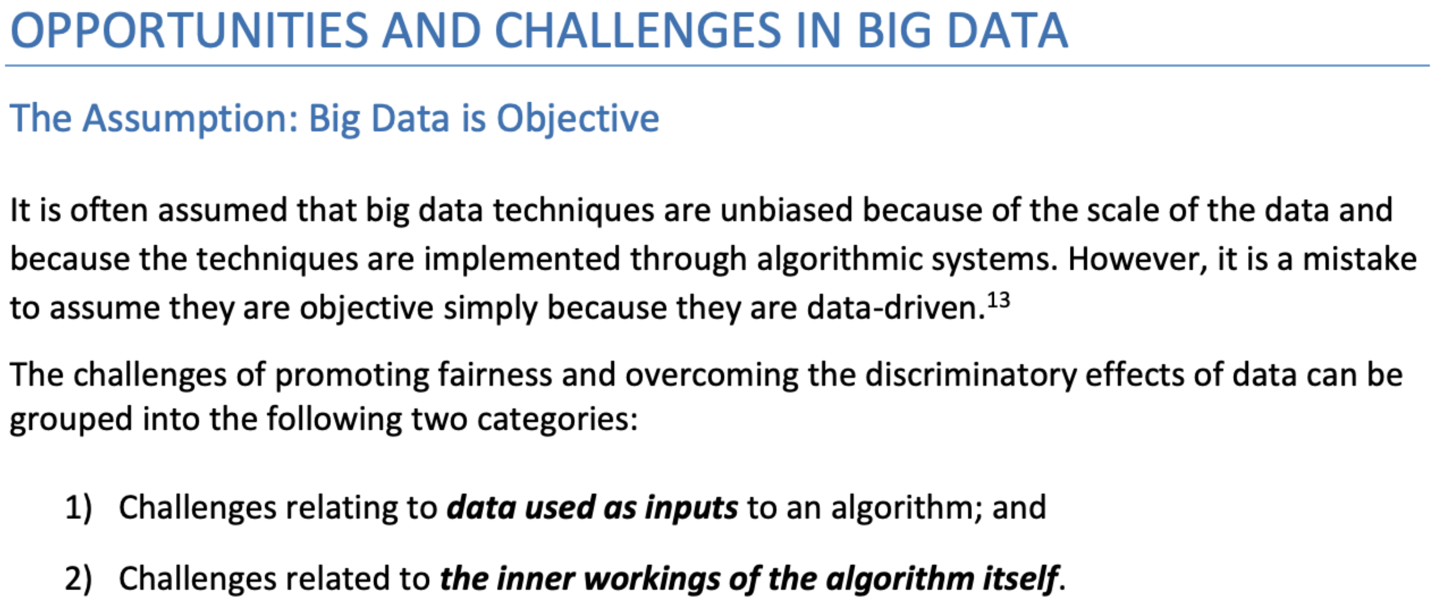
May 2016

THE WHITE HOUSE
WASHINGTON

Source: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Things Machine Learning Isn’t

- Artificial intelligence: Creating machines that can mimic human behavior/cognition
- Data science: Extracting knowledge/insights from noisy, unstructured data
- Neutral



OPPORTUNITIES AND CHALLENGES IN BIG DATA

The Assumption: Big Data is Objective

It is often assumed that big data techniques are unbiased because of the scale of the data and because the techniques are implemented through algorithmic systems. However, it is a mistake to assume they are objective simply because they are data-driven.[13]

The challenges of promoting fairness and overcoming the discriminatory effects of data can be grouped into the following two categories:

1) Challenges relating to ***data used as inputs*** to an algorithm; and
2) Challenges related to ***the inner workings of the algorithm itself***.

Source: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Defining a Machine Learning Task: Example

- Learning to "years remaining"

- Three components
  - Task, T

    Predict how much longer someone will live
  - Performance metric, P

    (squared) relative ratio of predicted : actual lifespan
  - Experience, E

    collection of demographic information

Defining a Machine Learning Task: Example

- Learning to
- Three components
  - Task, T

    character recognition for signatures
  - Performance metric, P

    time/ accuracy in predictions
  - Experience, E

    collect users past signature(s)

Defining a Machine Learning Task: Example

- Learning to
- Three components
  - Task, T

    playing chess
  - Performance metric, P

    win rate

    # of games the program wins (augmented w/ move counts)
  - Experience, E

    - games played against the computer
    - data of games played

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised) binary classification task**

features

labels

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

data points

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **<u>binary classification</u> task**

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **<u>classification</u> task**

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Risk
Yes	Low	Normal	Low Risk
No	Medium	Normal	Low Risk
No	Low	Abnormal	Medium Risk
Yes	Medium	Normal	High Risk
Yes	High	Abnormal	High Risk

Our first Machine Learning Task

- Learning to diagnose heart disease

  as a **(supervised)** **<u>regression</u> task**

features

targets

data points

Family History	Resting Blood Pressure	Cholesterol	Medical Costs
Yes	Low	Normal	$0
No	Medium	Normal	$20
No	Low	Abnormal	$30
Yes	Medium	Normal	$100
Yes	High	Abnormal	$5000

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Is this a "good" Classifier?

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the dataset

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)

training dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **error rate** is the proportion of data points where the prediction is wrong

Training vs. Testing

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset (Yes)
- A **test** dataset is used to evaluate a classifier's **predictions**

test dataset

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	**Predictions**
No	Low	Normal	No	Yes
No	High	Abnormal	Yes	Yes
Yes	Medium	Abnormal	Yes	Yes

- The **test error rate** is the proportion of data points in the test dataset where the prediction is wrong $(1/3)$

A Typical (Supervised) Machine Learning Routine

- Step 1 – training
  - Input: a labelled training dataset
  - Output: a classifier
- Step 2 – testing
  - Inputs: a classifier, a test dataset
  - Output: predictions for each test data point
- Step 3 – evaluation
  - Inputs: predictions from step 2, test dataset labels
  - Output: some measure of how good the predictions are; usually (but not always) error rate

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset

labels

data points

Heart Disease?
No
No
Yes
Yes
Yes

- This classifier completely ignores the features…

Our first Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Majority vote classifier: always predict the most common label in the **training** dataset

labels

data points

Heart Disease?	Predictions
No	Yes
No	Yes
Yes	Yes
Yes	Yes
Yes	Yes

- The training error rate is $2/5$

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?	Predictions
Yes	Low	Normal	No	No
No	Medium	Normal	No	No
No	Low	Abnormal	Yes	Yes
Yes	Medium	Normal	Yes	Yes
Yes	High	Abnormal	Yes	Yes

- The training error rate is 0!

When poll is active, respond at **pollev.com/301601polls**

Is the memorizer learning?

Yes

No

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Our second Machine Learning Classifier

- A **classifier** is a function that takes feature values as input and outputs a label
- Memorizer: if a set of features exists in the **training** dataset, predict its corresponding label; otherwise, predict the majority vote
- The memorizer (typically) does not **generalize** well, i.e., it does not perform well on unseen data points
- In some sense, good generalization, i.e., the ability to make accurate predictions given a small training dataset, is the whole point of machine learning!

Key Takeaways

- Components of a machine learning problem
- Machine learning vs. artificial intelligence vs. data science
- Algorithmic bias
- Components of a labelled dataset for supervised learning
- Training vs. test datasets
- Majority vote & memorizer classifiers