Foundations of Data Science
Lecture 4, Module 2
Spring 2022

Rumi Chunara, PhD

So Far…

- What is Data Science?
- Data Handling
- Doing Data Science
- Types of Data
- Data cleaning, sampling, processing
- Entropy, Information
- SVD and PCA
- Feature and Model selection

Today

- Introduction to Machine Learning – what is it?
- Subsequently: Basic Algorithms (data practicalities, tradeoffs)

What is machine learning?

"A field of study that gives computers the ability to learn without being explicitly programmed." (1959)

Arthur Samuel, AI pioneer
Source: Stanford

What is machine learning?

- Supervised Learning
- Unsupervised Learning

Machine Learning

- **Supervised**
  - Given input samples *X* and output samples *y* such that *y = f(X)*, we would like to **learn f**, and evaluate it on new data.
  - **Focus today**

- **Unsupervised**
  - Given only input samples *X* of the data, we **compute a function f** such that *y = f(X).*

Machine Learning

- **Supervised:**
  - Is this image a cat, dog, car, house?
  - How would this user score that restaurant?
  - Is this email spam?
  - Is this blob a supernova?
- **Unsupervised**
  - Cluster some handwritten digit data into 10 classes.
  - What are the top 20 topics in Twitter right now?
  - Find and cluster distinct accents of people at NYU.

Supervised Machine Learning

Supervised learning (predictive modeling):

- Predict an outcome based on input data
  - **Example:** predict whether an email is spam or not

- Goal is generalization

Machine Learning Terminology

150 ***observations*** *(n = 150)*

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

target

4 features *(p = 4)*

Feature matrix has *n* rows and *p* columns

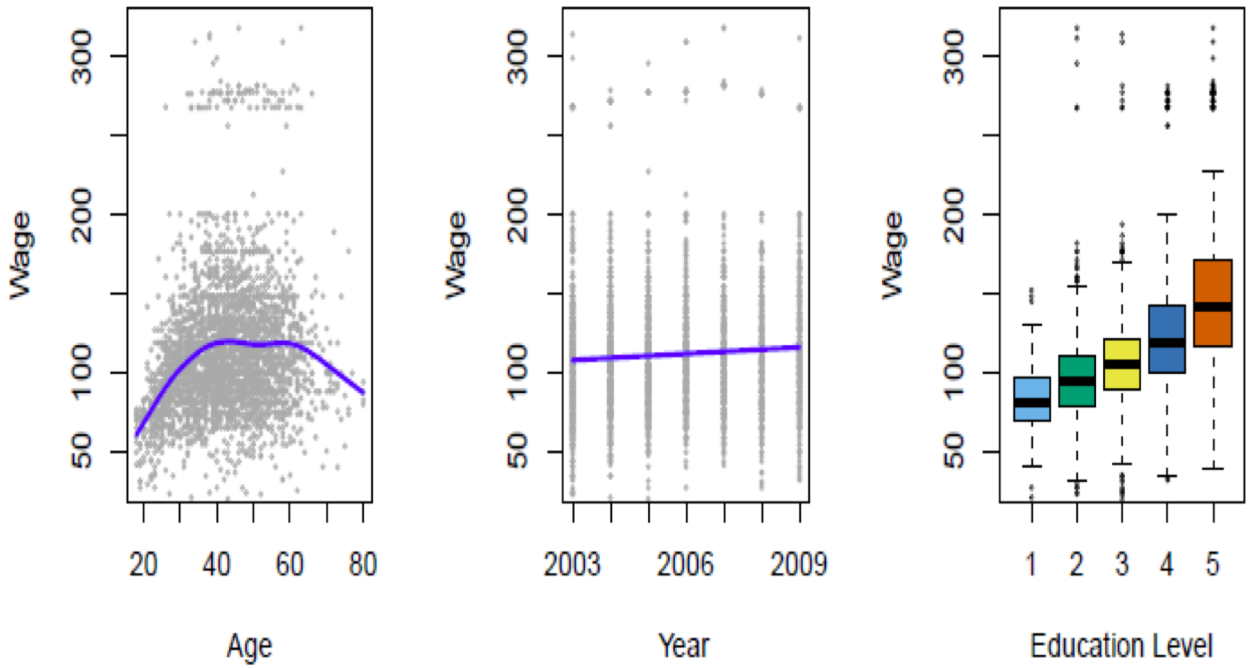
Target *y* is a vector with length *n*

Machine Learning Terminology

- **Observations** are also known as: samples, examples, instances, records
- **Features** are also known as: predictors, independent variables, inputs, regressors, covariates, attributes
- **Target** is also known as: outcome, label, response, dependent variable
- **Regression problems** have a continuous response
- **Classification problems** have a categorical response.

Supervised Learning Example

Predict salary using demographic data

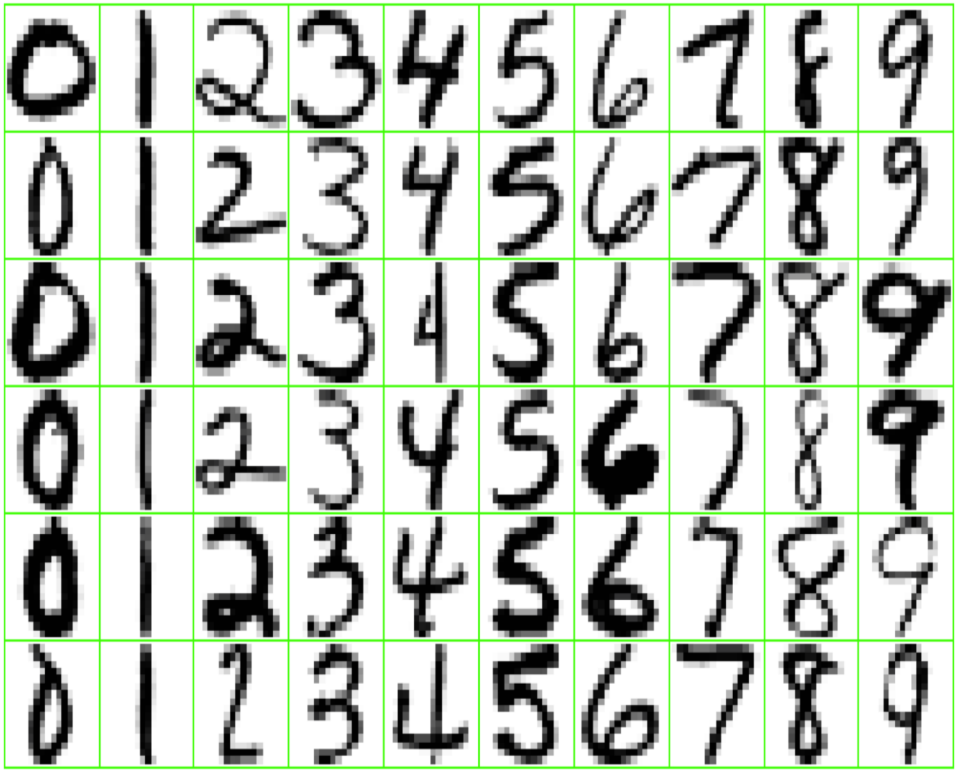


Income survey data for males from the central Atlantic region of the USA in 2009

Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

Supervised Learning Example

Identify the numbers in a handwritten zip code



Source: https://class.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/introduction.pdf

Categories of Supervised Learning

There are two categories of supervised learning:

Regression

- Outcome we are trying to predict is continuous
- Examples: price, blood pressure

Classification

- Outcome we are trying to predict is discrete (values in a finite, unordered set)
- Examples: spam/ham, COVID-19 test (positive or negative)

Regression or Classification?

Problem: Children born prematurely are at high risk of developing infections, many of which are not detected until after the baby is sick

Goal: Detect subtle patterns in the data that predict the time of infection before it occurs

Data: 16 vital signs such as heart rate, respiration rate, blood pressure, etc…

Impact: Model is able to predict the onset of infection 24 hours before the traditional symptoms of infection appear

Image: http://www.babycaretips4u.com/wp-content/uploads/2014/03/premature-baby.jpg
Case Study: http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695

Regression or Classification?

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Supervised Learning

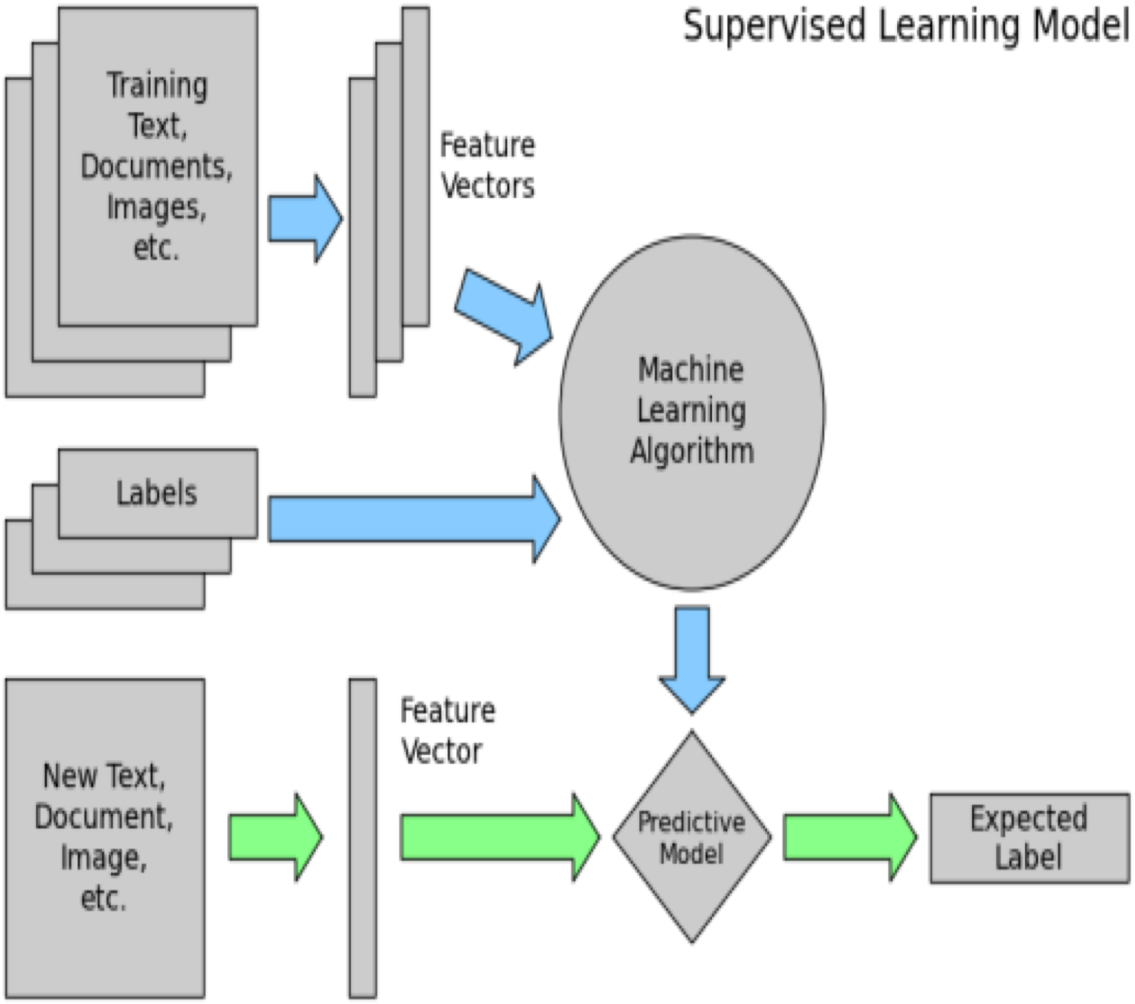
How does supervised learning "work"?

- **Train** a **machine learning model** using **labeled data**
  - "Labeled data" is data with a response variable
  - "Machine learning model" learns the relationship between the features and the target (**a function *f***)

- Make predictions on **new data** for which the response is unknown

Supervised Learning

How does supervised learning "work"?



Supervised Learning Example

Supervised learning example: Dog detector

- Input data: Images from Google
- Features: Numerical representations of the images
- Response: Dog (yes or no), hand-labeled

1. Train a **machine learning model** using **labeled data**
   - Model learns the relationship between the image data and the "dog status"

1. Make predictions on **new data** for which the response is unknown
   - Give it a new image, predicts the "dog status" automatically

Techniques

- **Supervised Learning:**
  - Decision Trees
  - kNN (k Nearest Neighbors)
  - Linear Regression
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests

Questions we aim to answer

- Why pick a certain algorithm?
- How does the data need to be pre-processed to use the selection?
- Managing data challenges (missing data, large data sets)
- How do we evaluate the implementation?
- What are the broader considerations (for the end user, etc.)