Foundations of Data Science
Lecture 3

Rumi Chunara, PhD

CS3943/9223

So Far…

- What is Data Science?
- Data Handling
- Doing Data Science
- Types of Data
- Data cleaning, sampling, processing
- Entropy, Information, Model Building (Feature and Model selection)
- Decision Trees

Today

- Intro to ML – what is it
- Two Basic Algorithms
  - Linear Regression
  - kNN

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 (Starting Today)
- Unsupervised Learning (Later)

Machine Learning

- **Supervised:** We are given input samples (X) and output samples (y) of a function y = f(X). We would like to "learn" f, and evaluate it on new data. Types:
  - **Classification:** y is discrete (class labels).
  - **Regression:** y is continuous, e.g. linear regression.

- **Unsupervised:** Given only samples X of the data, we compute a function f such that y = f(X) is "simpler".
  - **Clustering:** y is discrete
  - Y is continuous: **Matrix factorization, Kalman filtering, unsupervised neural networks**.

What is Machine Learning?

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational **learning** theory in artificial intelligence. In 1959, Arthur Samuel defined **machine learning** as a "Field of study that gives computers the ability to learn without being explicitly programmed".

Machine learning - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/**Machine_learning** Wikipedia

More about Machine learning

What is Machine Learning?

- One definition: “Machine learning is the semi-automatic extraction of knowledge from data.”

- **Automatic extraction:** A computer provides the insight
- **Semi-automatic:** Requires many smart decisions by a human

Supervised Machine Learning

Supervised learning
(aka “predictive modeling”):

- Predict an outcome based on input data
- Example: predict whether an email is spam
- Goal is “generalization”

ML Terminology

150 observations *(n = 150)*

Feature matrix “X” has n rows and p columns

Response “y” is a vector with length n

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*

response

4 features (p = 4)

ML Terminology

Observations are also known as: samples, examples, instances, records

Features are also known as: predictors, independent variables, inputs, regressors, covariates, attributes

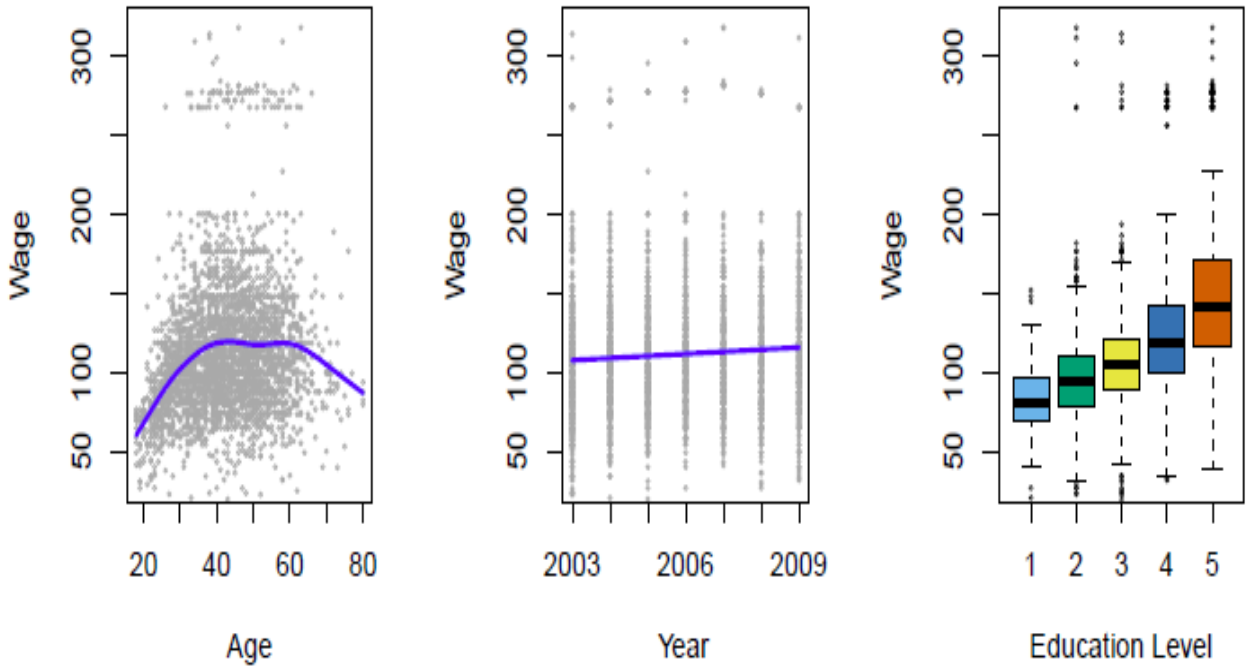
Response is also known as: outcome, label, target, dependent variable

Regression problems have a continuous response.
Classification problems have a categorical response.

The type of supervised learning problem has nothing to do with the features!

Supervised Learning Example

Predict salary using demographic data

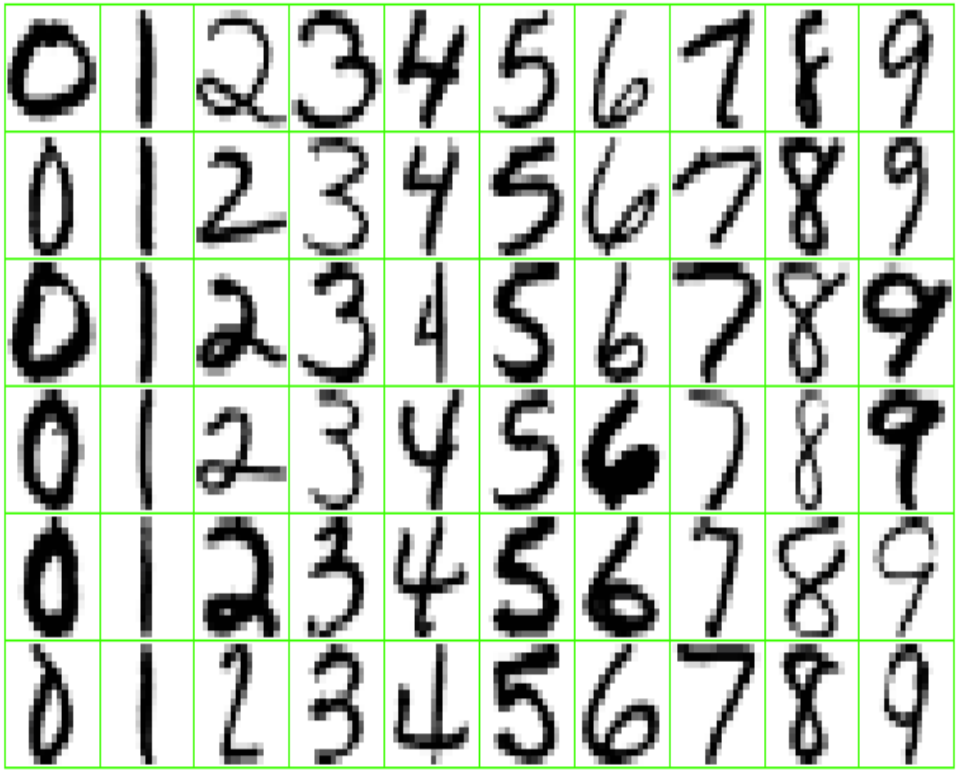


Income survey data for males from the central Atlantic region of the USA in 2009

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 categorical (values in a finite, unordered set)
- Examples: spam/ham, cancer class of tissue sample

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

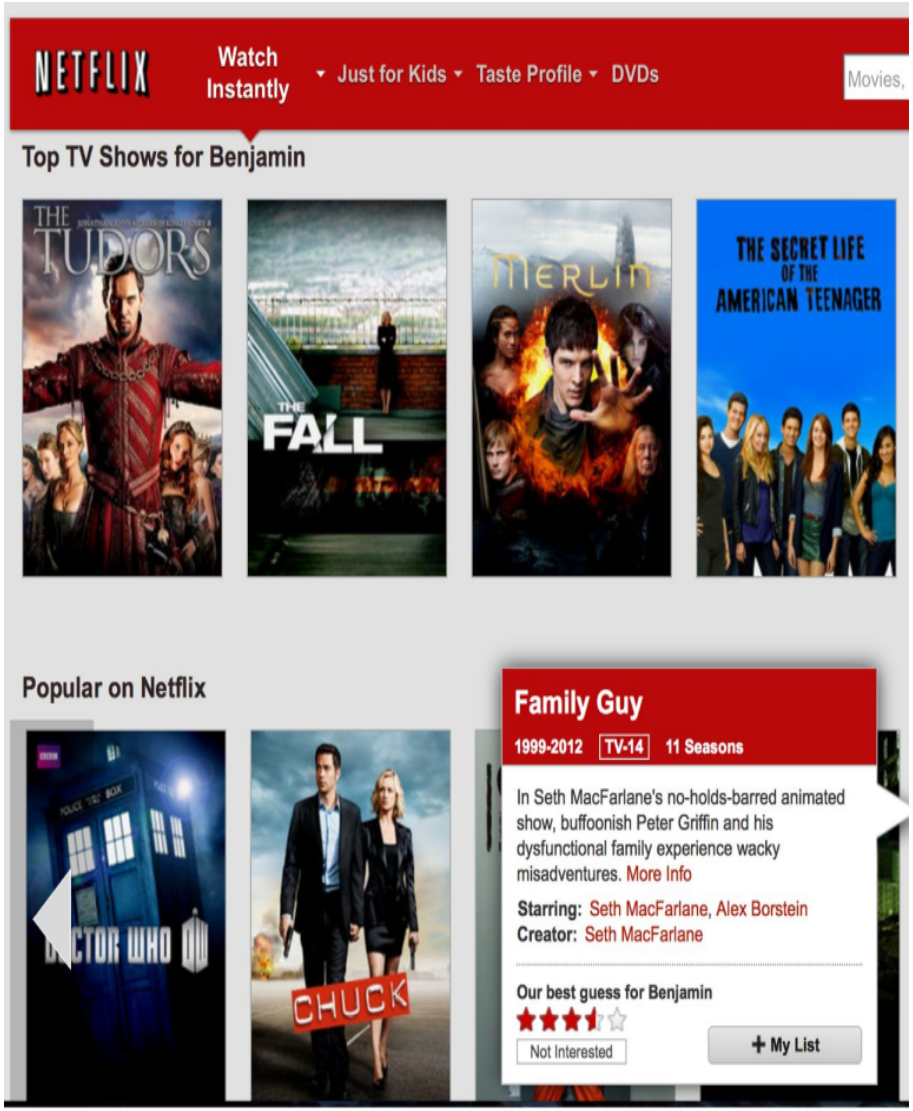
Goal: Detect subtle patterns in the data that predicts infection before it occurs

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?



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"?

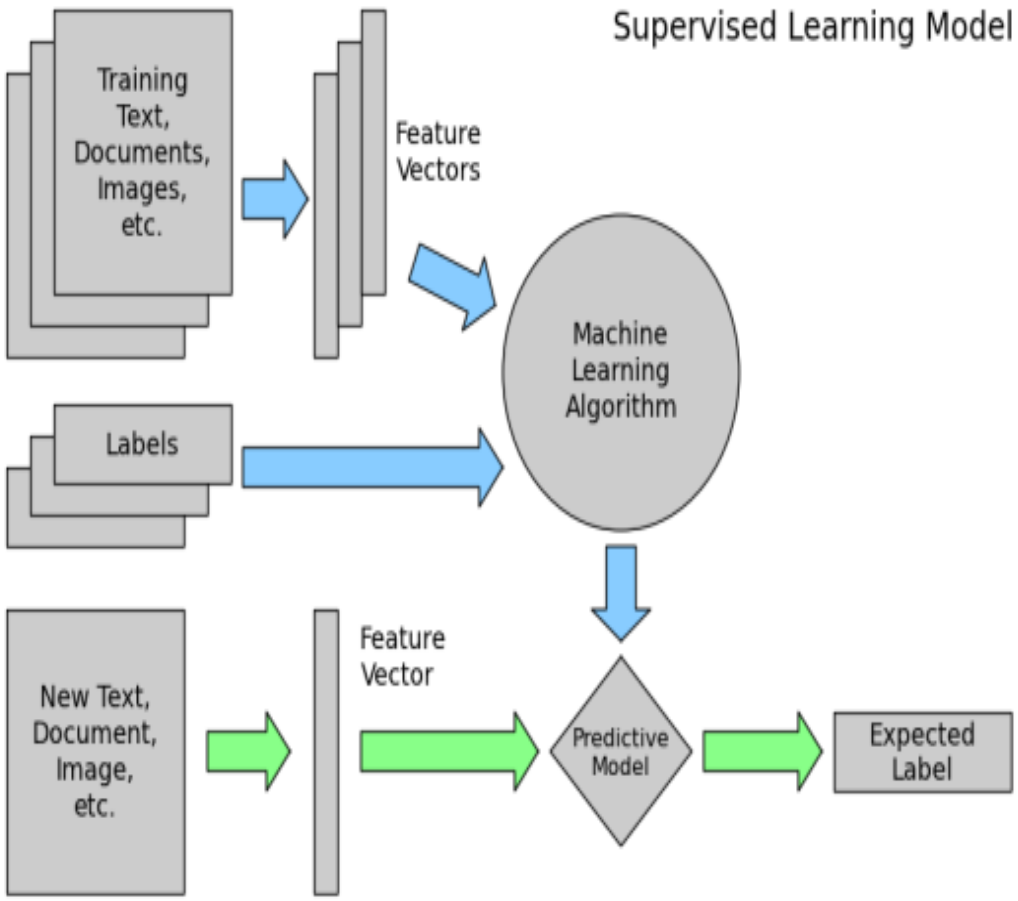
1. Train a **machine learning model** using **labeled data**
   - "Labeled data" is data with a response variable
   - "Machine learning model" learns the relationship between the features and the response

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

The primary goal of supervised learning is to build a model that "generalizes": It accurately predicts the **future** rather than the **past**!

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"

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

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 hand-written digit data into 10 classes.
  - What are the top 20 topics in Twitter right now?
  - Find and cluster distinct accents of people at NYU.

Semi-supervised Learning

- Sometimes the question of whether an analysis should be considered supervised or unsupervised is less clear-cut.
- Suppose that we have a set of n observations.
- For m of the observations, where $m < n$, we have both predictor measurements and a response measurement.
- For the remaining $n - m$ observations, we have predictor measurements but no response measurement.
- Such a scenario can arise if the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect.

Techniques

- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Linear Regression
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Topic Models