Unit 0
Introduction to Machine Learning

Prof. Phil Schniter

THE OHIO STATE UNIVERSITY

ECE 4300: Introduction to Machine Learning, Sp20

Learning objectives

- Be familiar with some examples of machine learning (ML)
- Understand ML problem formulation
  - Identify the goal
  - Identify the training data
- Be able to recognize ML tasks:
  - supervised vs. unsupervised
  - regression vs. classification
- For supervised ML tasks, identify predictors and target variables
- Understand the role of expert knowledge vs. data-driven learning

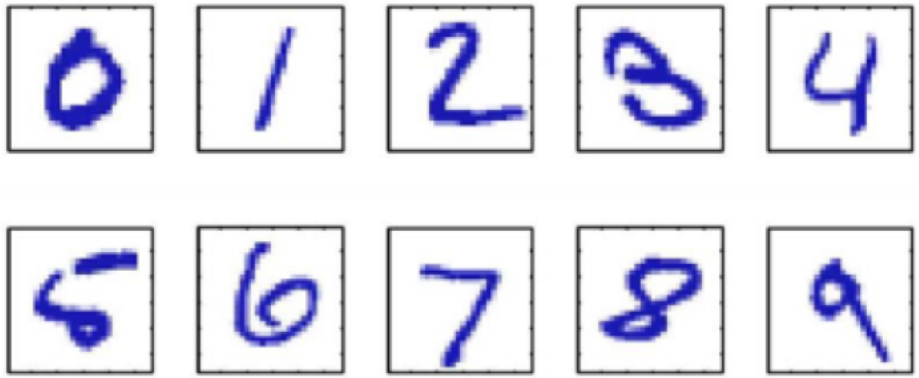
Outline

- What is machine learning?
- Types of machine learning tasks
- Why the hype today?

What is machine learning?

- Machine learning refers to information extraction methods that learn from data
- Why might we want this?
  - Sometimes experts do not know accurate mathematical models
    (e.g., images, speech signals, stock trajectories)
  - Sometime our (accurate) models depend on many time-varying parameters
    (e.g., noise cancellation)
  - Humans may be good at the task, but we want to automate it
    (e.g., speech/face recognition),

Example 1: Digit recognition

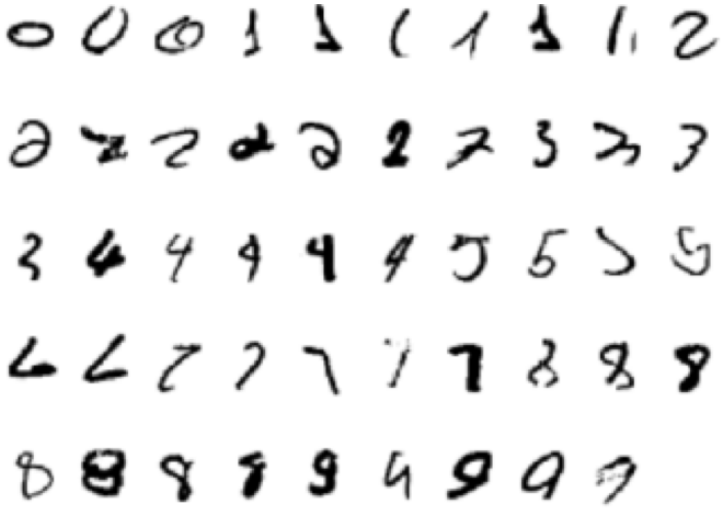


Images are 28 x 28 pixels

- Goal: Recognize a digit from a given image
- Problem statement: Design a function $f(\boldsymbol{x}) \to y$, where $\boldsymbol{x}$ is a 28×28 matrix and $y \in \{0, 1, \ldots, 9\}$.
- Challenge: We don't have good mathematical models for images of digits (e.g., how do we model a "5"?)
- Inspiration: Humans learn to recognize digits. Maybe machines can do the same!

Machine-learning approach to digit recognition

- Goal: Learn a function $f(\boldsymbol{x})$ that predicts the "label" y from the image matrix $\boldsymbol{x}$
- Need training data: Many $(\boldsymbol{x}, y)$ pairs
- Ex: MNIST database: 6000 examples of handwritten-digit images $\boldsymbol{x}$ labeled with the true value y
  - Collected by the National Institute of Standards (NIST) with the goal of reading zip-codes on letters
- Current systems achieve $<$0.21% error-rate on MNIST



Training examples from MNIST database.

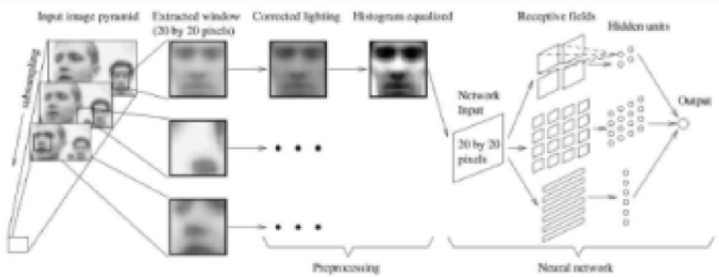
Each image is labeled with the truth.

Example 2: Face detection

- In each image region, determine if pixels show a face or non-face
- In this example, the label y is binary

Training data for face detection

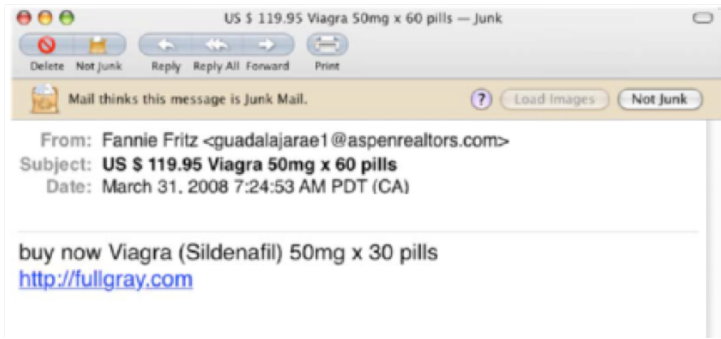
- A typical early face-recognition dataset:
  - 5000 face examples
    - All near frontal
    - All normalized (size, location, rotation, brightness)
    - Variations in age, gender, race, lighting across dataset
  - 10^8 non-face examples
- Many more datasets are available today
  - See, e.g., http://www.face-rec.org/databases/
- Many early face-detection algorithms were very complex



Rowley, Baluja, and Kanade, 1998

Example 3: Spam detection

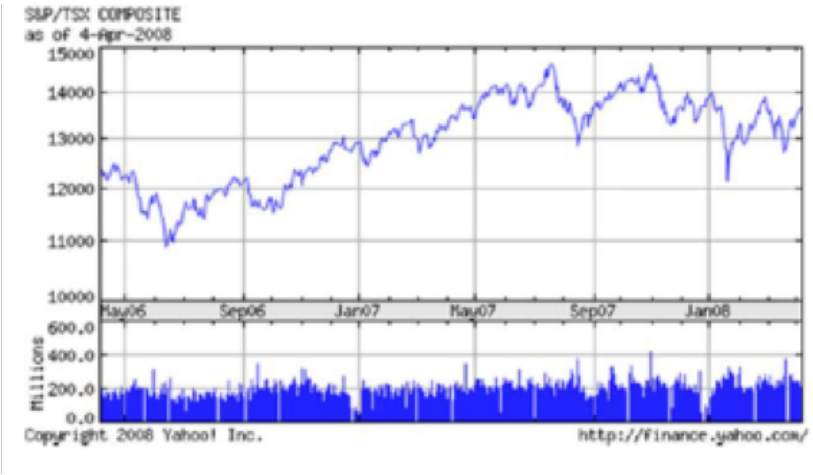
- Goal: Predict whether an email is junk or not
  - Labels y are binary ($1 =$ junk, $0 =$ not junk)
- Must represent each email numerically
  - Typical model is bag of words:
  - Enumerate all *common* English words, $j = 1, \ldots, p$
  - Represent each email as a vector $\boldsymbol{x} = [x_1, \ldots, x_p]^\mathsf{T}$, where $x_j = \#$ instances of word j.
- Challenges:
  - Very high-dimensional vector
  - System must adapt over time (keep up with spammers)



Email example

Example 4: Stock-price prediction

- Say you want to predict the price of a stock tomorrow
- What variables $(\boldsymbol{x}, y)$ would you use?
  - Is y discrete, as in previous examples?
- What is a non-machine learning approach?
- What is a machine learning approach?



Stock trajectory

Machine learning in many fields

- Transportation: self-driving cars
- Manufacturing: control, robotics, fault detection
- Telecommunications: spectrum monitoring, modulation detection
- Medicine: imaging, medical diagnosis
- Bioinformatics: motifs, alignment
- Web mining: search engines
- Finance: credit scoring, fraud detection
- Retail: market basket analysis, customer relationship management

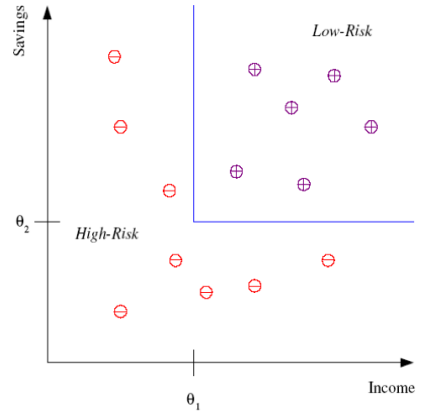
Many many more examples, growing by the day!

Outline

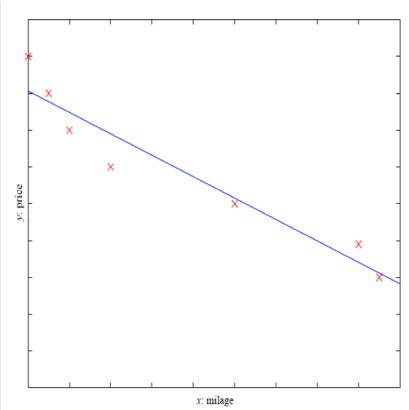
- What is machine learning?
- Types of machine learning tasks
- Why the hype today?

Supervised learning

- Goal: Predict "label" or "target" y from a vector of features $\boldsymbol{x} = [x_1, \ldots, x_p]^{\mathsf{T}}$
- Training: Many examples of $(\boldsymbol{x}, y)$ pairs
- Classification
  - The target y is discrete-valued
  - Example: Credit scoring: predict low/high risk $y \in \{0, 1\}$ based on income x_1 and savings x_2
- Regression
  - The target y is continuous-valued
  - Example: Predict car price $y \in \mathbb{R}$ based on mpg x_1 and brand x_2
  - The figure on the right shows only the use of x_1



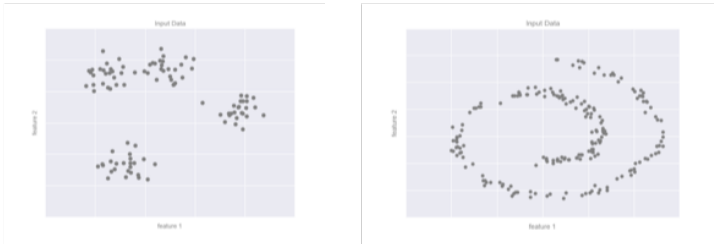
Classification for credit scoring

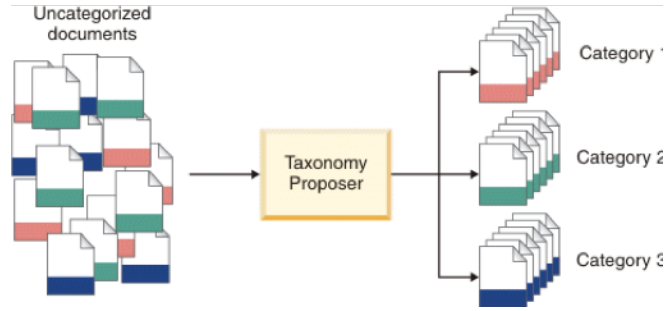


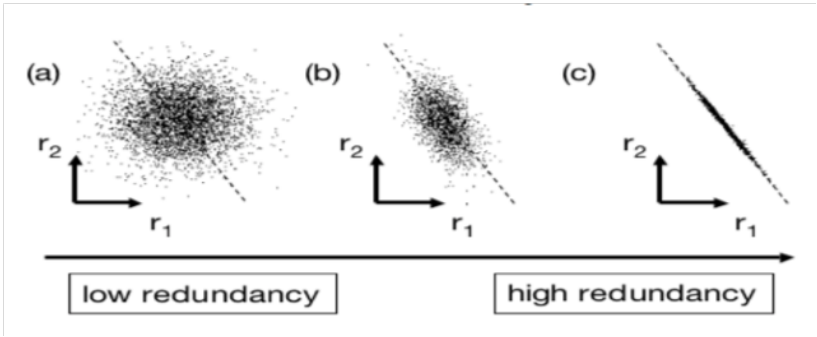
Regression for car pricing

Unsupervised learning

- Goal: learn structure of data $\boldsymbol{x}$
- Training: Many examples of $\boldsymbol{x}$ (but no y)
- Clustering
  - Grouping similar instances of $\boldsymbol{x}$
  - Examples: document clustering, customer segementation, motif discovery in bioinformatics
- Principal component analysis (PCA)
  - Finding correlations in data
  - Used for dimensionality reduction







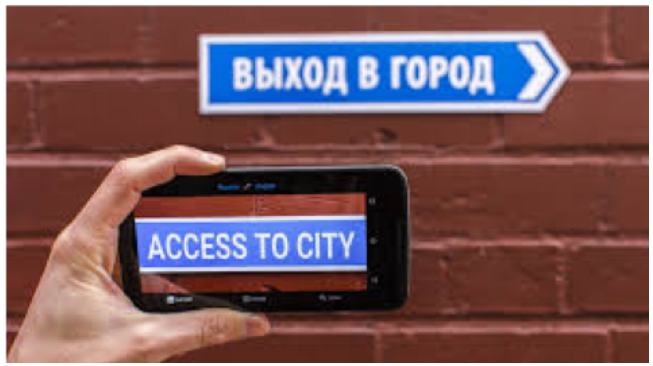
- Other methods include non-negative matrix factorization (NMF) and Gaussian-mixture modeling (GMM). Will discuss later...

Outline

- What is machine learning?
- Types of machine learning tasks
- Why the hype today?

What is ML doing today?

- Machine translation
- Autonomous driving
- Jeopardy
- Difficult games like Alpha Go
- Many, many other things!



Why now?

- Machine learning is not new
  - Many of the core ideas were developed during 1950s–1990s
- So why is it so popular now?
  1. Big data
     - Massive connectivity (e.g., internet, smartphones)
     - Massive storage available
     - Massive datasets have been collected
  2. Computational advances
     - Cumulative effects of Moore's law
     - Distributed computation (e.g., server farms)
     - Special purpose hardware: GPUs

server farm

graphical processing unit

In-class exercise

- Break into small groups
- Think of an application that interests you
- Identify a specific task that might be tackled using machine learning
  - What is the goal?
  - What type of ML task is it? (e.g., classification, regression,...)
  - What is the training data?
  - Why use ML and not some expert-driven method?