Math 10, Fall 2022
Introduction to Programming for Data Science

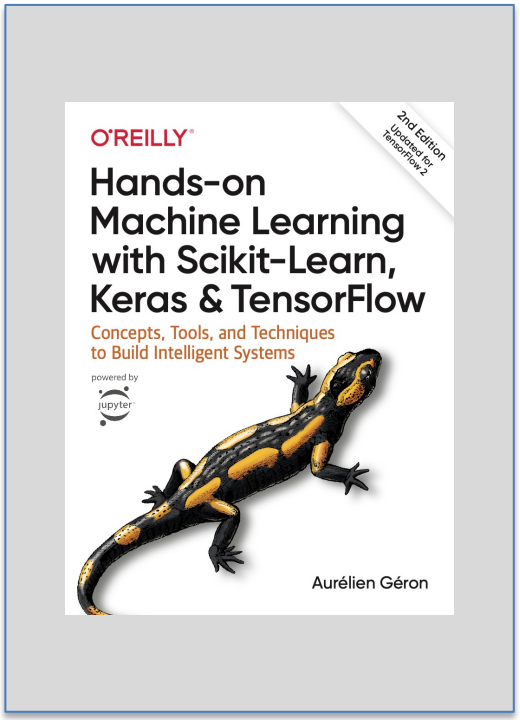
- First half of Math 10: Exploratory Data Analysis
- Second half: Introduction to Machine Learning

What is Machine Learning?

Informal definition, from *Hands on Machine Learning* by Aurélien Géron:

Machine Learning is the science (and art) of programming computers so they can learn from data.

(Pretty vague. Is computing a mean an example of Machine Learning?)



Two types of ML problems:

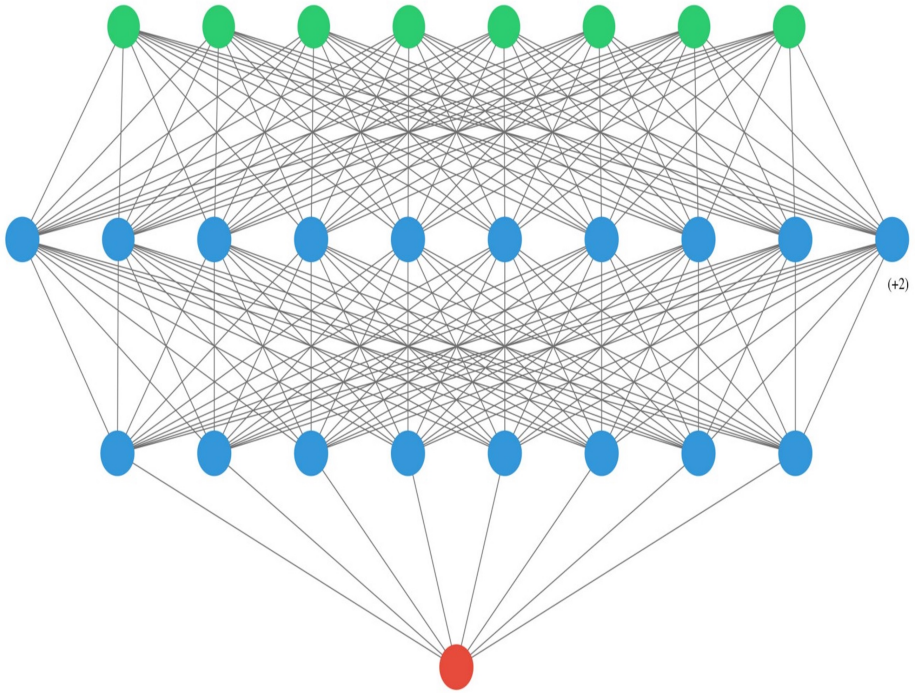
Supervised Learning: Have labeled data

Unsupervised Learning: No labels

Supervised Learning

Two types of supervised learning problems:

- Regression: predicting a quantitative value.
- Classification: predicting a category.

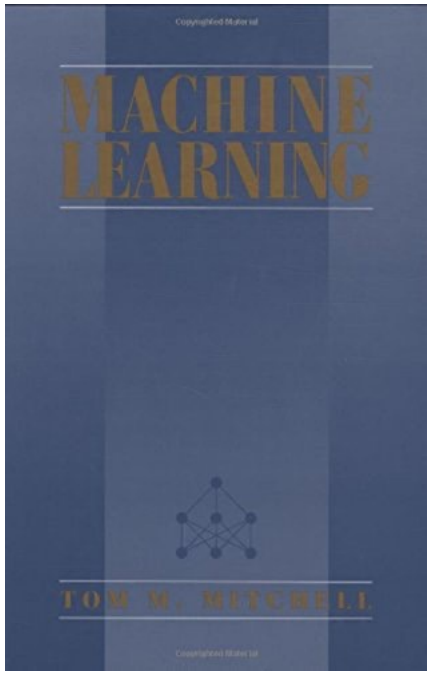


A more formal Definition

(Adapted from Mitchell, *Machine Learning*, 1997.)

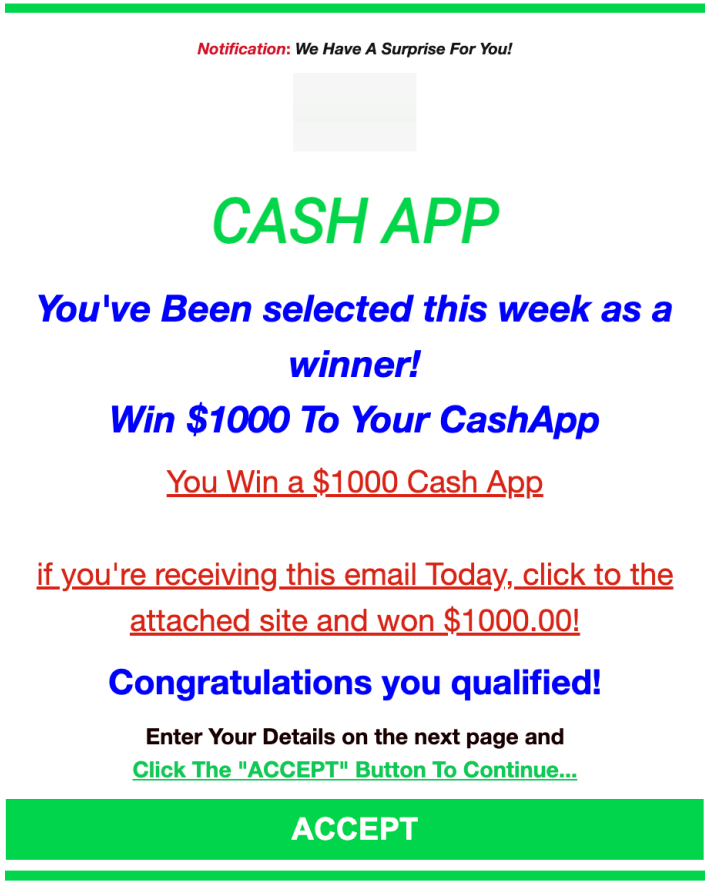
A definition of learning from data:

Consider a collection of tasks T, a performance measure P, a baseline strategy B, and an algorithm A which depends on a set of training data D. The algorithm A is said to learn from the data D, if its performance at tasks in T, as measured by P, is better than the baseline strategy B.



Example: Classification

- Task: Determine if an email is spam.
- Baseline strategy: Predict spam if the email contains 3 or more exclamation points (!).
- Training data: Emails that have been identified as spam/not spam.
- Performance measure: Percentage correctly identified.



Example: Classification

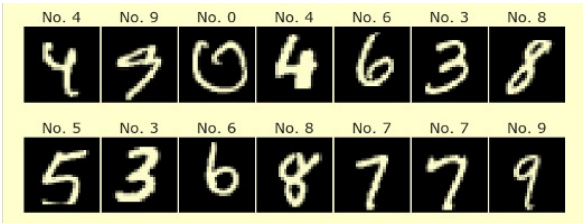
- Task: Estimate the probability that a Titanic passenger survived.
- Baseline strategy: Use the average survival rate as prediction.
- Training data: Survival outcomes and passenger characteristics.
- Performance measure: Log loss (severe penalty for being both confident and wrong)

Example: Regression

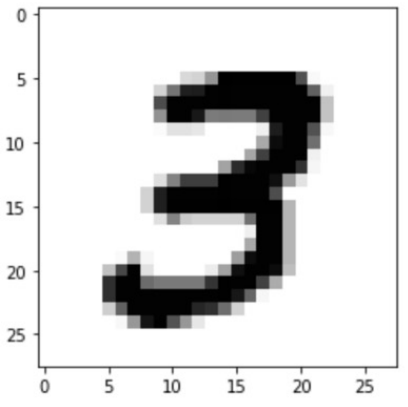
- Task: Predict prices of houses in King County, Washington, based on characteristics.
- Baseline strategy: Always predict the median house value.
- Training data: Prices of certain houses in the county.
- Performance measure: Mean Absolute Error (less concern with outliers than Mean Squared Error)

Source: Wikimedia Commons, Hannah Lewis House, Jon Roanhaus

Example: Classification

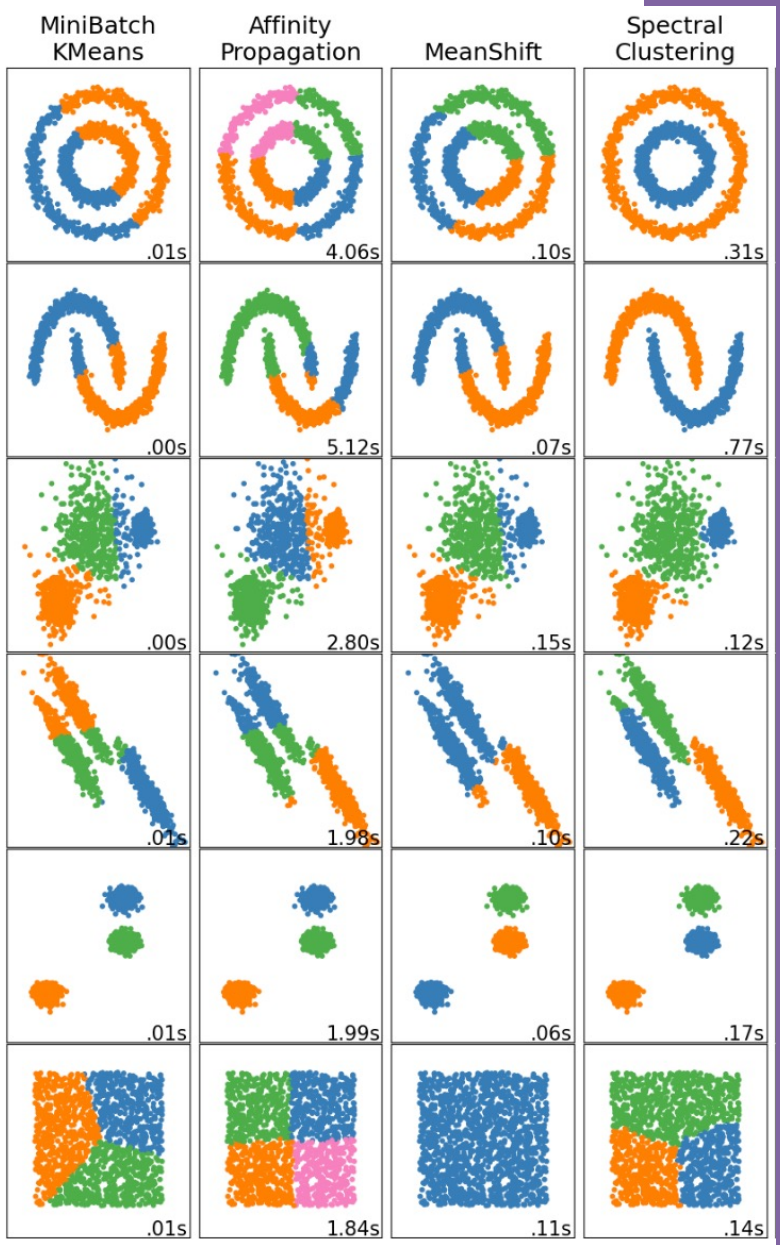


- Task: Identify the values of handwritten digits (given pixel values).
- Baseline strategy: Always predict 0.
- Training data: Handwritten digits together with their correct values.
- Performance measure: Percentage of digits correctly identified.



Example: Unsupervised Learning

- Task: Divide data into K distinct clusters.
- Baseline strategy: Assign randomly.
- Training data: 100 sample points.
- Performance measure: Average distance of a sample point to the nearest centroid.



Example: Unsupervised Learning

- Task: Generate art.
- Baseline strategy: Random pixel values.
- Training data: Collection of artworks together with evaluations by an expert.
- Performance measure: Was an expert tricked into thinking the artwork was made by a human?

Example: Unsupervised Learning

- Task: Reduce the dimensionality of NumPy arrays representing images of faces.
- Baseline strategy: Keep only the center-most 36 pixel values
- Training data: A collection of images of faces.
- Performance measure: Similarity of the reduced face to the original image.