

Introduction to Machine Learning

Reading

- Chapters 1
 - Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.
 - Available online: <https://hastie.su.domains/Papers/ESLII.pdf>
- Chapters 1
 - James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.
 - Available online: <https://www.statlearning.com/>
- ML intro from a statistical point of view

What is machine learning?

- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”

Mitchell, Tom M. (1997). Machine Learning. McGraw-Hill, New York

The task, T

- Ideally, solve tasks that are hard/tedious/repetitive/error-prone for humans to do
- (Part of) the artificial intelligence problem
- Collect some data and learn to perform well on new data (from the same distribution)
- Many of these tasks existed before the term “machine learning” was invented
 - E.g., in statistics, signal processing, etc.
 - Ultimately, learning is a statistical task
 - most modern learning techniques were developed by the statistics community

Example 1: Email Spam

- A classical ML problem from the 1990s
- Given an email (including sender, subject, body), decide whether it is spam or ham (legit email)
- We are given a training dataset of ~4K emails and need to learn to make a decision

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.*

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

Example 2: Handwritten digit recognition

- Another classical problem from the 1990s
- Given a grayscale image (i.e., a matrix of numbers in $[0,1]$), identify the digit in the image
- Used to recognize handwritten amounts in checks
 - One of the first real applications of neural nets in the 1990s
- Once again, we are given a number of training images in order to learn patterns
 - Dataset has 60K images

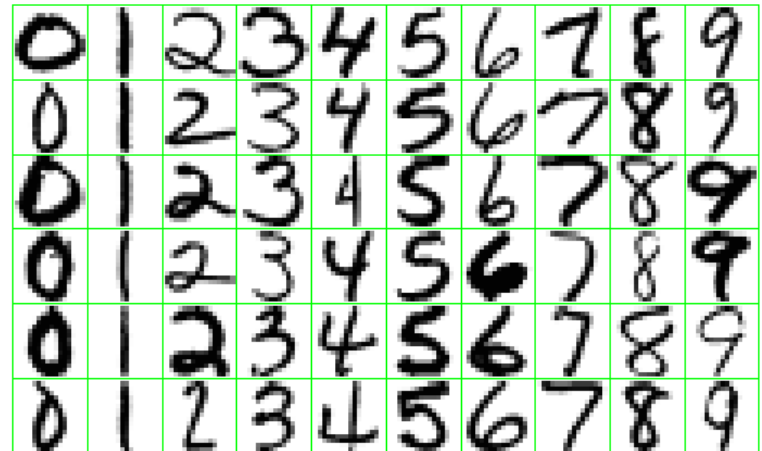


FIGURE 1.2. Examples of handwritten digits from U.S. postal envelopes.

Example 3: Playing Go

- Go is an ancient board game
- During each turn, a player can place a stone on any non-occupied cell
- The goal is to surround the opponent's stones
- Board is 19x19, and each cell can contain white/black/no stone
 - Total state space is $\sim 3^{361}$ (can't brute force search)
- The learning task is to learn a strategy for playing Go
 - There is no training data per se – the algorithm has to try different actions and observe outcomes
 - Researchers recently developed AI that can beat the best human players, using reinforcement learning



Classification

- Classical machine/deep learning task
- Given an input example, determine which category the example belongs to
- Examples:
 - image classification (is this an image of a cat, dog, etc.)
 - object recognition (identify all objects in an image)
 - spam detection
 - determine whether a bank customer should be given a loan or not

Regression

- Another classical task
- Given an input example, determine some continuous-valued property of the example
- Examples:
 - predict expected claim amount that an insured person will make
 - establish the effect of education on a person's salary
 - determine distance to car in front given an image

Anomaly Detection

- Essentially a binary classification problem
- Examples:
 - detect fraud from banking data
 - determine whether a patient has cancer from an MRI image
- Anomaly detection is a very old and well studied problem
 - Also known as hypothesis testing
 - Originally studied in radar object tracking, where the goal was to detect enemy planes using radar data (during the Cold War)

Natural Language Processing

- Transcription: given an audio waveform (or an image with written text), output a sequence of characters describing the words
- Translation: given a statement in one language, translate it to another
- Sentiment analysis: given a paragraph or an article, determine its overall sentiment (positive, negative, neutral, etc.)
- Summary: given a paragraph or an article, summarize the main points
- Generative NLP: generate conversation, describe an image, generate code, etc.
 - All the rage these days, with large language models

Synthesis and Generative Models

- Given a set of training examples, generate new examples that look similar to the training data
- Examples: given images of cats, generate new images of cats, e.g., in a different environment, colors, etc.
- Not fully formalized, but generated images need to have sufficient variability and be different from training data
- Generative Adversarial Networks (GANs) received a lot of attention in the 2010s
- Transformers are the latest rage
 - GPT = Generative Pre-trained Transformer
 - Generate text, images, control actions given an input of arbitrary length

Reinforcement Learning

- Given a dynamical system (described by differential equations, or a Markov decision process), learn a controller that maximizes a given reward function
- Examples: learn to drive a car from images, play games
- There is no training data per se – during the training process, the controller generates (s, a, s', r) tuples
 - s is the system's current state (e.g., position)
 - a is the applied action
 - s' is the next state
 - r is the observed reward after applying a and entering s'

The Performance Measure, P

- For classification tasks, the natural measure is accuracy, i.e., the proportion of examples that are classified correctly
- For regression tasks, one can use a variety of distance measures, e.g., difference between predicted and true value
- In RL, the measure may be the expected reward over some horizon
- Measure is always evaluated on a test set, since we are interested in the learned model's performance on unseen data
 - Will talk later about the distinction between training and test data

The Experience, E

- Supervised learning – we are given both training examples **and** corresponding labels (e.g., labeled images)
- Unsupervised learning – given unlabeled data, we would like to learn the entire distribution that generated the data
 - The above are related – we can think of supervised learning as unsupervised learning where we learn the joint distribution of examples and labels
- Reinforcement learning – the algorithm explores the state space as part of learning