The Statistical Learning Framework

STAT 37710 / CMSC 35300

Rebecca Willett and Yuxin Chen

What is machine learning?

- spam filter

- automatic discovery of rules from history

What is machine learning?

- recommender system

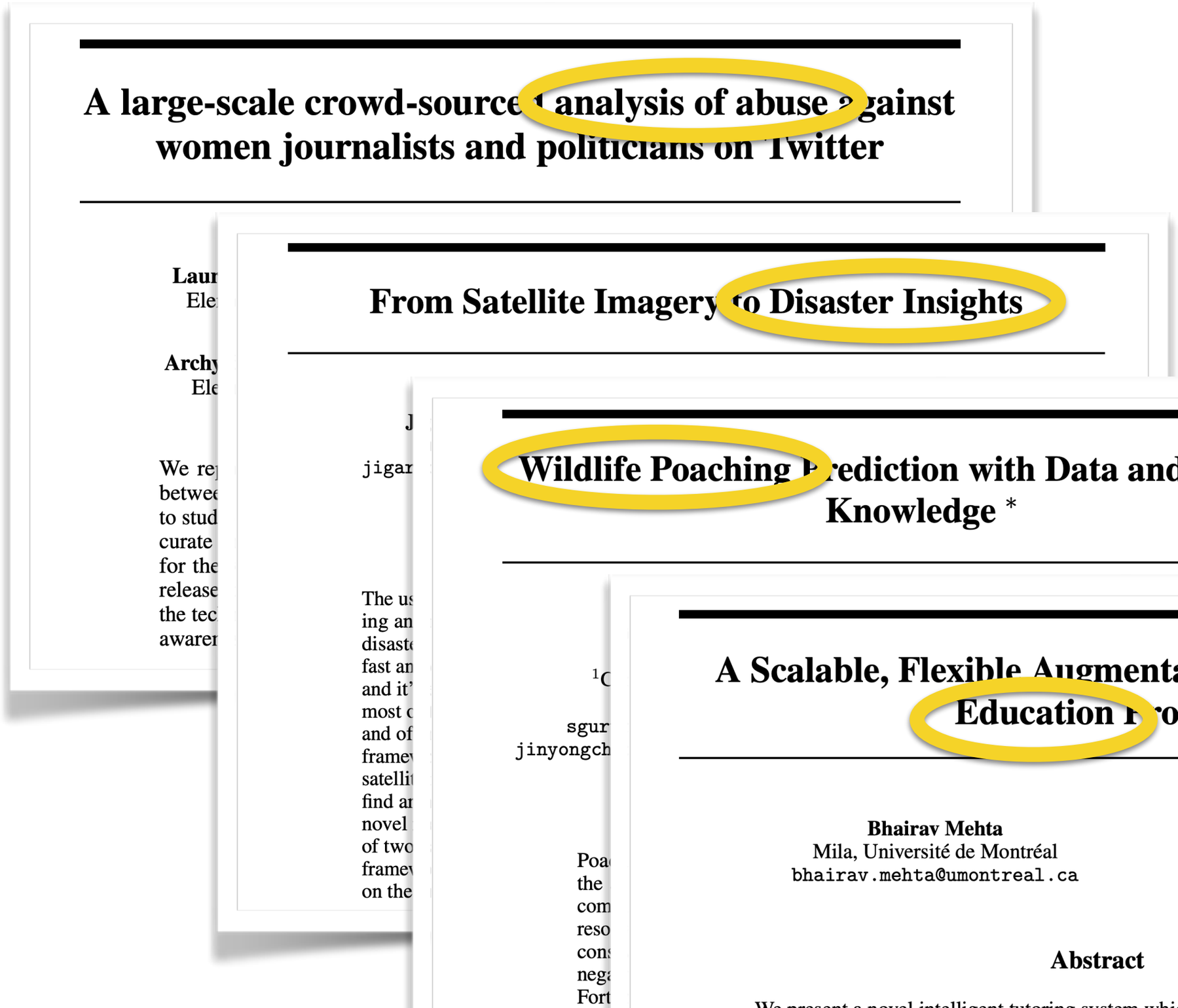
- extracting knowledge from previous experiences

What is machine learning

- Computed-aided discovery

- extracting “useful” patterns from examples

Machine learning is pervasive…



Helping Bees and Beekeepers with AI

Honey Authentication with Machine Learning Augmented Microscopy

Peter He (Department of Computing, Imperial College London) · **Alexis Gkantiragas** (Department of Molecular Biology, University College London) · **Gerard Glowacki** (H

Towards a Sustainable Food Supply Chain Powered By Artificial Intelligence

Volodymyr Kuleshov, Mariah Seymour, Danny Nemer, Nathan Fenner, Matthew Schwartz

Afresh Technologies and Stanford University

ntroduction

Improving Traffic Safety in Jakarta Through Video Analysis

João Caldeira, Alex Fout, Aniket Kesari, Raesetje Sefala, Katy Dupre, Joe Walsh

University of Chicago, Colorado Sta

icago

Intr

- **Problem:** ~2,000 people die
  accidents in Jakarta, Indones
- The city of Jakarta investe
  capture data on traffic beh

Next Hit Predictor - Self-exciting Risk Modeling for Predicting Next Locations of Serial Crimes

THE UNIVERSITY OF IOWA

Yunyi Li
The University of Iowa
yunyi-li@uiowa.edu

Tong Wang
The University of Iowa

picked

ct

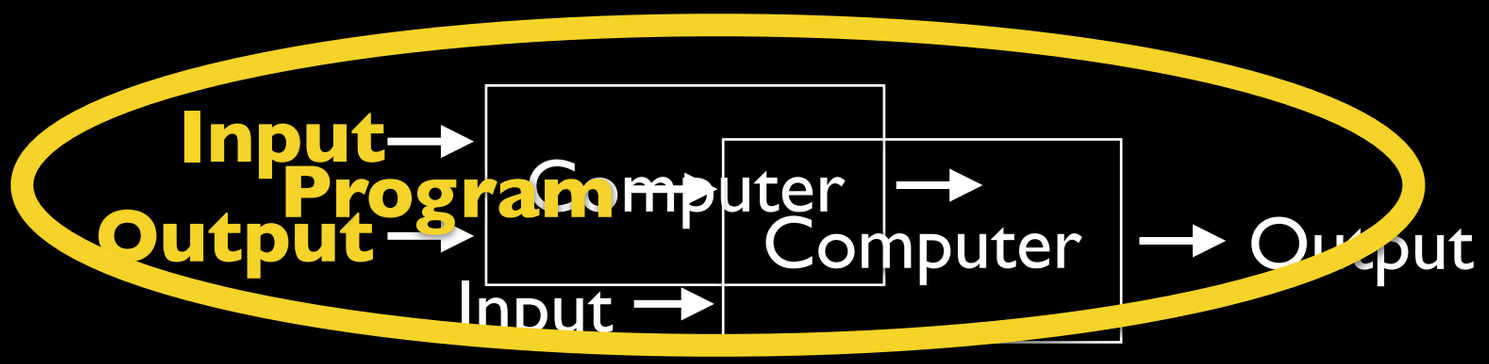
Definition of Macine learning

Arthur Samuel, 1952:

Machine learning is the ability to perform tasks without being explicitly programmed

Arthur Samuel, 1952:

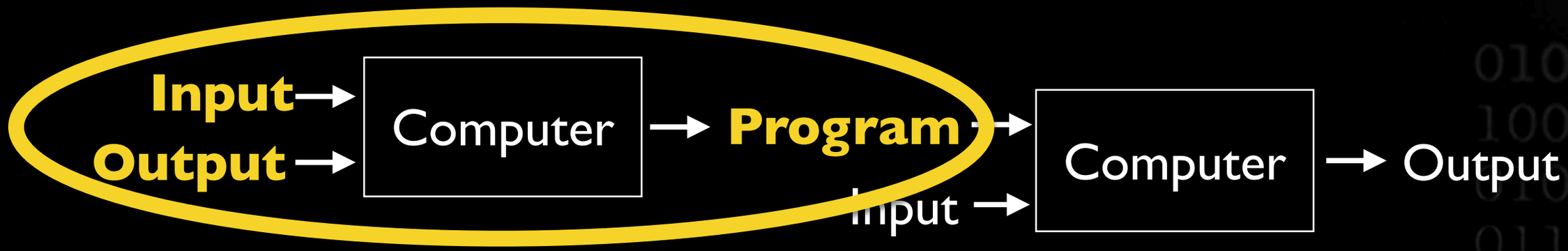
Machine learning is the ability to learn tasks without being explicitly programmed



Another definition

Tom Mitchell, 1997:

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



A brief history

- 1950: Alan Turing created the world-famous Turing Test — “The Imitation Game”
- 1952: Arthur Samuel coined the phrase “machine learning”, and create a computer program that improves checkers game
- 1956: The Dartmouth workshop started AI as a field (John McCarthy, Marvin Minsky, Nathaniel Rochester, Claude Shannon …)
- 1958: Frank Rosenblatt designed the first artificial neural network
- 1986: Geoffrey Hinton introduced backpropagation which enabled monumental leaps in ANNs
- 1995: Cortes and Vapnik “standardized” support vector machines
- 1997: Deep Blue beats a chess champion
- 1999: Computer-aided diagnosis catches more cancers (CAD Prototype at UChicago)
- 2009: Launch of ImageNet
- 2011: Watson competed on Jeopardy! and won against Ken Jennings and Brad Rutter
- 2012: AlexNet won the ImageNet competition, which led to the use of GPUs and Convolutional Neural Networks in machine learning
- 2016: AlphaGo defeated the human champion Lee Sedol in a best-of-five duel match
- …
- Now: AI dominates Silicon Valley and becomes pervasive

https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/

Related disciplines

information theory

statistics

philosophy
causality

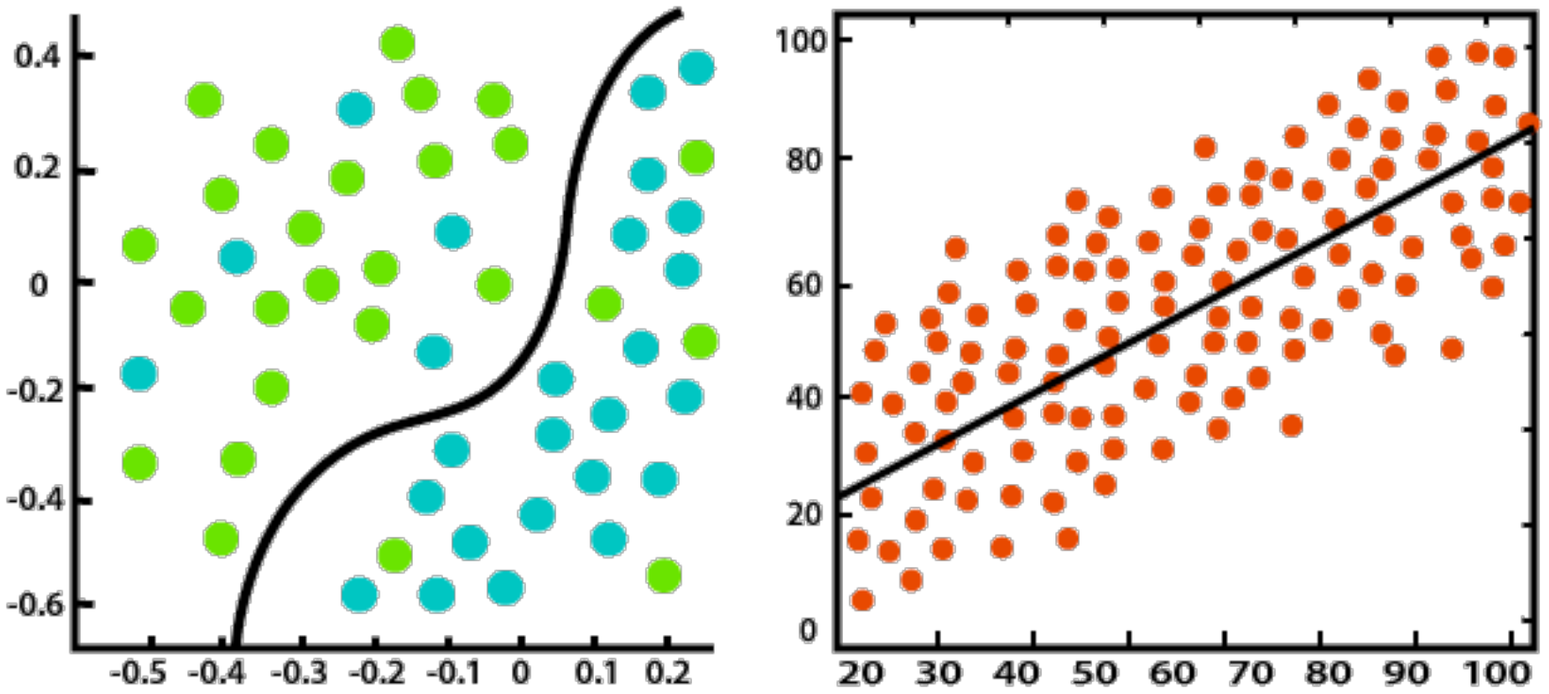
Machine Learning

algorithms
optimization

neural-informatics

Supervised machine learning

- classification vs regression



- learning rules from training data

The statistical learning framework

- Data
  - Domain set $\mathcal{X}$
  - Label set $\mathcal{Y}$
  - A joint probability distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$ of training data, which consists of n points $(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
- Model
  - A prediction function $f : \mathcal{X} \to \mathcal{Y}$
- Measure of success
  - A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Measurement of success

- Given the prediction function f, and the loss function ℓ, the loss of an example is $\ell(f(x), y)$
- Training error/ empirical error of f

$$\hat{L}(f) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- Generalization error/ risk/ true error

$$L(f) \stackrel{\text{def}}{=} \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f(x), y)] = \int \ell(f(x), y) p(x, y) dx dy$$

Learning objective

- Given a collection of observations $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and a model class** $\mathscr{F}$, the goal of learning is to determine an estimator $f_S \in \mathscr{F}$, that best matches the observation (i.e., with the minimal error wrt a certain loss function)

Remarks**: The No Free Lunch Theorem states that every successful ML algorithm must make assumptions. This also means that there is no single ML algorithm that works for every settings.

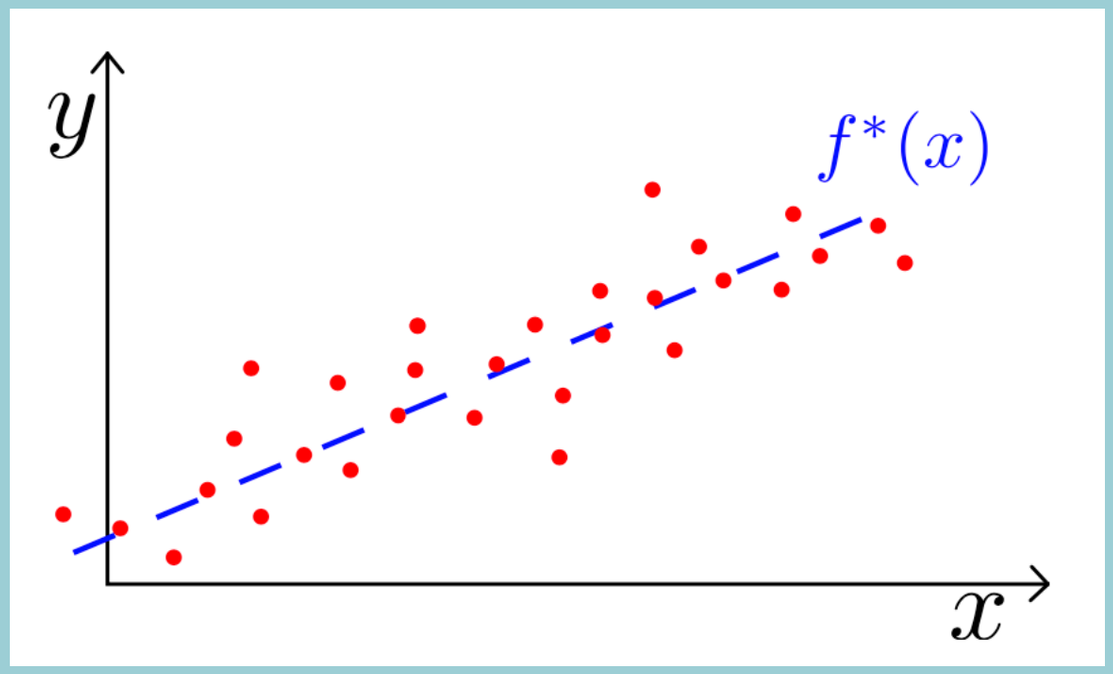
Example 1: linear regression

In regression we often consider $y = f^*(x) + \varepsilon$

Here $x \in \mathbb{R}^p, y \in \mathbb{R}, f^*$ is a fixed unknown function, and ε is random noise, e.g. standard Gaussian $N(0,\sigma), \sigma \in [0,\infty)$.

For example for linear regression, $\mathcal{F} = \{f : \mathbb{R}^p \to \mathbb{R} \mid f(x) = x^T w, w \in \mathbb{R}^p\}$.

We aim to find $f^*(x) = x^\top w^*$ for some $w^* \in \mathbb{R}^p$ that best matches observations $\{(x_i, y_i)\}_{i=\{1,\ldots,n\}}$



Example 2: binary classification

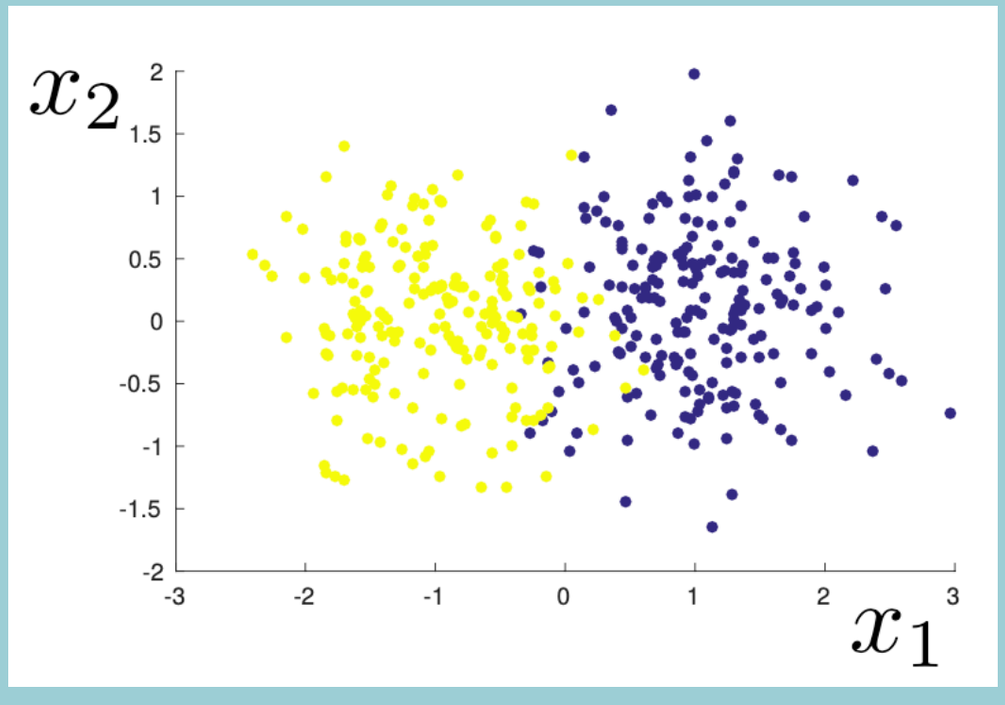
In binary classification, $x \in \mathbb{R}^p, y \in \{-1,1\}$ (or $y \in \{0,1\}$)

A basic example of data model is a mixture of two Gaussians:

$$x \sim \begin{cases} \frac{1}{Z} N(\mu_+, R_+) & \text{if } y = 1 \\ \frac{1}{Z} N(\mu_-, R_-) & \text{if } y = -1 \end{cases}$$

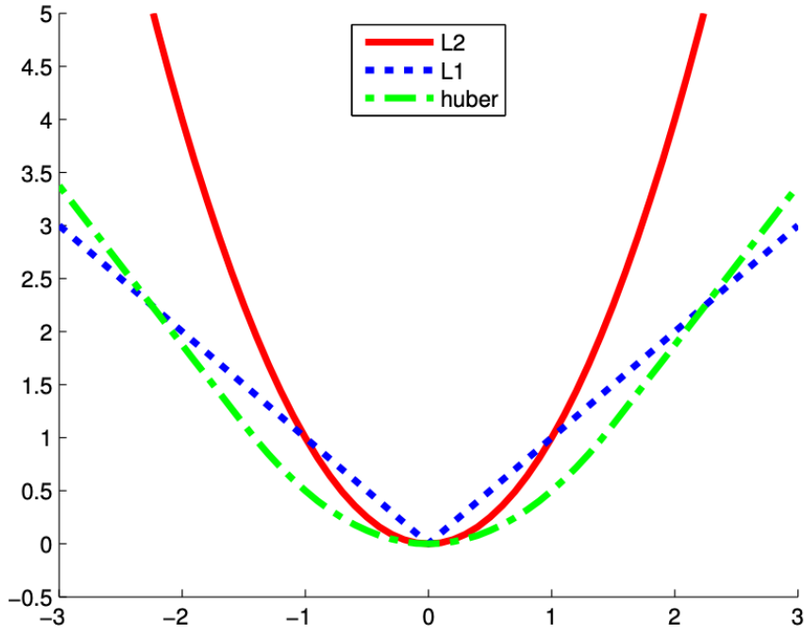
where $\mu_+, \mu_- \in \mathbb{R}^p$, $R_+, R_- \in \mathbb{R}^{p \times p}_{\geq 0}$ and Z is a normalization factor.

The prediction function/model $f_\theta \in \mathcal{F}$ is parametrized by $\theta = (\mu_+, \mu_-, R_+, R_-) \in \Theta$, and we aim to find θ^* (which corresponds to f_{θ^*}) that best batches the observation

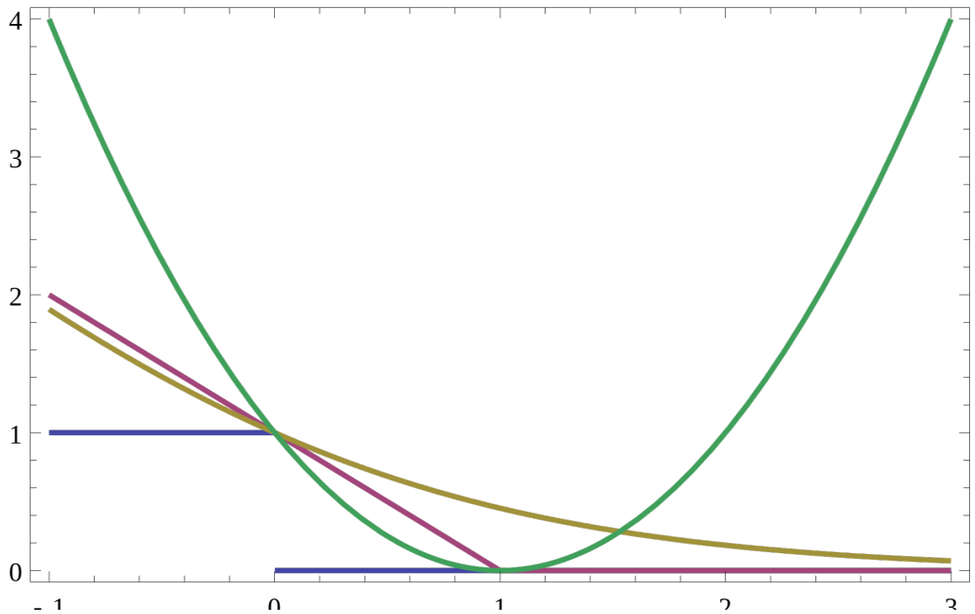


Examples of loss function ℓ

- Regression loss:
  - square loss: $\ell(f(x), y) = (y - f(x))^2$
  - absolute loss $\ell(f(x), y) = |y - f(x)|$
  - huber loss: Quadratic for $|y - f(x)| < \delta$ and linear for $|y - f(x)| > \delta$
    - robust and differentiable
  - …



- Classification loss:
  - 0-1 loss:
    - (misclassification error)
  - hinge loss : $\ell(f(x), y) = \max(0, 1 - yf(x))$
    - (penalize correct predictions when not confident)
  - log loss: $\ell(f(x), y) = \frac{1}{\ln 2} \log(1 + \exp(-yf(x)))$
    - always wants more "margin" $yf(x)$
  - square loss $\ell(f(x), y) = (y - f(x))^2$
    - magnify penalties if $|y - f(x)|$ is large
  - …



Where we are

- A brief tour of machine learning
    - From the Turing test (conceptual idea) to contemporary machine learning models
- The statistical learning framework
    - Function estimation given data and model class
    - Measurement of success defined via loss function

What's next

- Given a (parametric) model class $\mathscr{F}$, and loss function ℓ, how to estimate model parameter from observations
- Mean square error (square loss) $\ell(\hat{y}, y) = (\hat{y} - y)^2$ and bias-variance decomposition