CS 480/680 1/6/26

Lecture 1

What is ML

Q0

Lecture Notes 0 – Intro to Machine Learning

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

January 6, 2026

Course Logistics

Intro to Machine Learning

Resources

- Course website
- LEARN – submit homework code
- Crowdmark – submit homework written part (single pdf)
- Piazza – discussions, asking questions, announcements
- ...

Format and grading

15 min

- Lectures – in person, not recorded
- In class participation – ask questions, answer my questions
- Homework – about weekly, about 1/2 of homeworks not graded
- Quizzes – announced, in class, about 3-4 total
- NO midterm exam
- Final exam (2h)
- TA Office hours: 2 hours / week (t.b. scheduled)
- Instructor Office Hour: Mondays 2-3pm, in person, location TBA

$$\boxed{\text{Grade} \approx 60\% \text{ (homework + quizzes)} + 35\% \text{ final exam} + 5\% \text{ participation}}$$

These percentages may change by $\pm 5\%$.

What will you learn in this course?

tasks

- Machine Learning (ML) problems (e.g. prediction, classification, clustering)
- ML models (learners, or **predictors**) (e.g. decision trees, neural networks (nn), nearest neighbors (NN))
- Training algorithms (=”learning algorithms”)
- ML concepts (e.g. bias and variance, decison regions, complexity (of a model))
- Statitics (e.g. MaxLikelihood, Bayesian inference)
- Optimization (e.g. local and global minimum, gradient descent)

Taxonomies

...all of them incomplete

- Machine Learning Problems
  - Unsupervised
  - Supervised
  - (Semi-supervised)
  - Reinforcement
- Machine learning models (statistical predictors)
  - Parametric
  - Non-parametric
- Statistical inference paradigms
  - Bayesian
  - Maximum Likelihood (ML)
  - Penalized Likelihood
  - Loss-based

  These lists are meant to show that in this course we will not adopt a particular paradigm, but we will touch on most of them.

Plan for 480/680

- Supervised Learning (Prediction)
  - Predictor examples
  - Basic concepts: decision region, loss function, generative vs discriminative, bias-variance tradeoff
  - Training predictors: gradient descent, [Newton method]
  - [Combining predictors: bagging, boosting, additive models]
  - Regularized predictors: model selection, support vector machines, L1 regularization,
  - Learning theory and model selection basics
- Unsupervised Learning
  - Clustering: parametric, non-parametric
  - [Graphical models intro]
  - [Non-linear dimension reduction and geometric learning]
  - [Semi-supervised learning]
  - [Modeling graph data]
- [Reinforcement Learning]

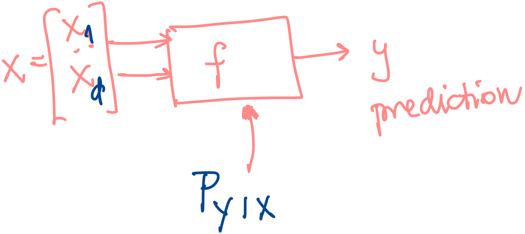
Supervised Learning (prediction)

Problem Given **input** $x \in \mathbb{R}^d$, **output** some property $y = f(x)$

- x is vector of attributes, features, inputs, covariates (if you are a statistician), ...
- y is **label**
- Data $\mathcal{D} = \{(x^1, y^1), \ldots (x^n, y^n)\}$ used for learning is **labeled data**
- First encounter with randomness $\mathcal{D}$ is sampled from a distribution P_{XY}. Goal is to learn $P_{Y|X}$.

label

$(x, y) \sim P_{XY}$



Example: Digit classification

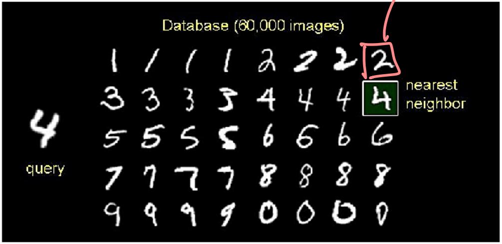
writer learner (predictn)

$y \rightarrow x \rightarrow \hat{y}$

Multiclass classification

Supervised Learning

- Nearest neighbour:



$x^i \in \{0:63\}^{16 \times 16 \text{ OR } 28 \times 28} = \mathcal{X}$

$y^i = 2 \in \mathcal{Y} = \{0, 1, \ldots 9\}$

Unsupervised learning

- Learn **the structure** of a distribution P_X from **unlabeled** data $\{x^1, x^2, \ldots x^n\}$ $\sim P_X$
  - Clustering – find groups in data (if they exist)
  - Dimension reduction – PCA, non-linear dimension reduction
  - Sparse dependencies – graphical models, sparse regression
  - Causal structure

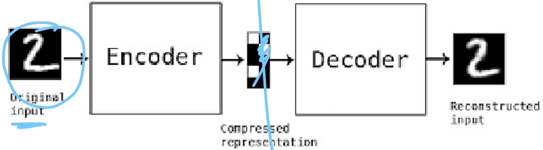
7 7 2 2

Unsupervised learning

- Learn **the structure** of a distribution P_X from **unlabeled** data $\{x^1, x^2, \ldots x^n\}$
  - Clustering – find groups in data (if they exist)
  - Dimension reduction – PCA, non-linear dimension reduction
  - Sparse dependencies – graphical models, sparse regression
  - Causal structure
- Learn a **distribution** P_X from **unlabeled** data $\{x^1, x^2, \ldots x^n\}$ – Density estimation, Autoencoders, GANs, "generative models"

Unsupervised Learning

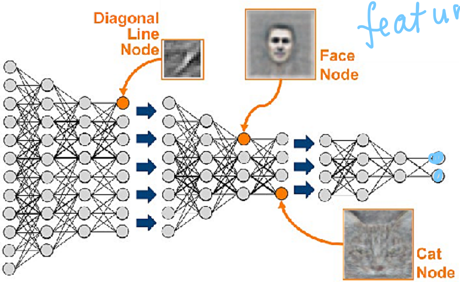
- Output is not given as part of training set
- Find model that explains the data
  - E.g. clustering, compressed representation, features, generative models



Unsupervised Feature Generation

- Encoder trained on large number of images

high-level features



Decode

Unsupervised Image Generation

- Which images are real? And which ones are fake?

Real

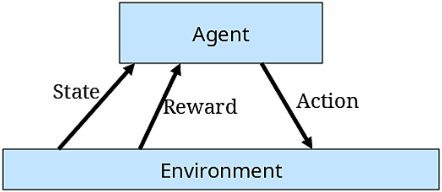
CelebA (Liu et al., 2015)

StyleGAN2 (Karras et al., 2020)

Fake!

- Image generation: variational autoencoders, generative adversarial networks, diffusion models

Reinforcement Learning Problem

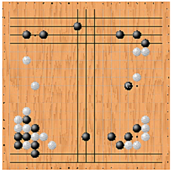


Goal: Learn to choose actions that maximize rewards

Game Playing

- Example: Go (one of the oldest and hardest board games)

- **Agent:** player
- **Environment:** opponent
- **State:** board configuration
- **Action:** next stone location
- **Reward:** +1 win / -1 loose



2016: AlphaGo defeats top player Lee Sedol (4-1)
Game 2 move 37: AlphaGo plays unexpected move (odds 1/10,000)

Combining Unsupervised, Supervised and Reinforcement Learning

- Modern systems:
  - Phase 1: unsupervised feature extraction (no labels)
  - Phase 2: supervised training (exploit labels)
  - Phase 3: fine tune by reinforcement learning (exploit reinforcements)

- Alpha Go: supervised + reinforcement learning
- Sentiment analysis with BERT: unsupervised + supervised learning
- ChatGPT: supervised + reinforcement learning

This course

- Supervised learning
- Unsupervised learning (some)
- But not Reinforcement Learning – see CS 486/686/885
  `https://cs.uwaterloo.ca/ ppoupart`