CS 480/680

1/6/26

Lecture 1

Q0

Intro to ML

Prediction pbs by type of output

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

- 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

Grade ≈ 60% (homework + quizzes) + 35% final exam + 5% participation

These percentages may change by ±5%.

480 vs 680

more [difficult] questions on HW & Exam

some sections of material not required

What will you learn in this course?

- 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] 486/686

Supervised Learning (prediction)

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

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

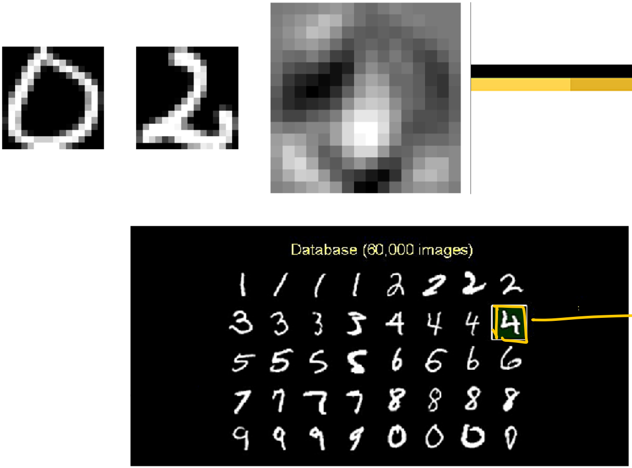
$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \Rightarrow \boxed{\quad f \quad} \longrightarrow \hat{y}$$

"guess" for y

$$(x^i, y^i) \sim \text{iid } P_{XY} \qquad i = 1:n$$

$$P_{Y|X}$$

Example: Digit classification



$y = 4 \in 0{:}9$

28×28

$x \in \{0:255\}$

CS480/680 Winter 2023 - Lecture 1 - Pascal Poupart PAGE 8

UNIVERSITY OF WATERLOO

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

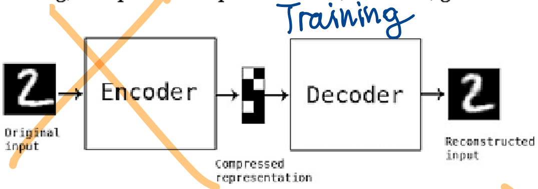
2 2

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"

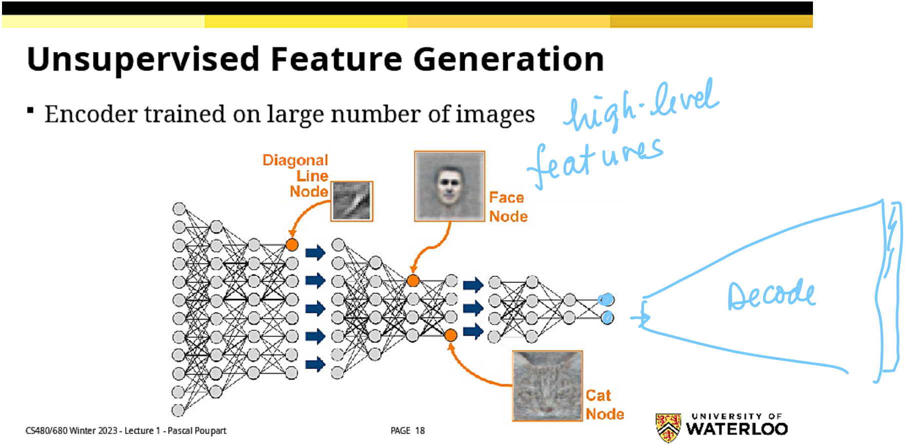
Unsupervised Learning

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



generate digit images

generative model



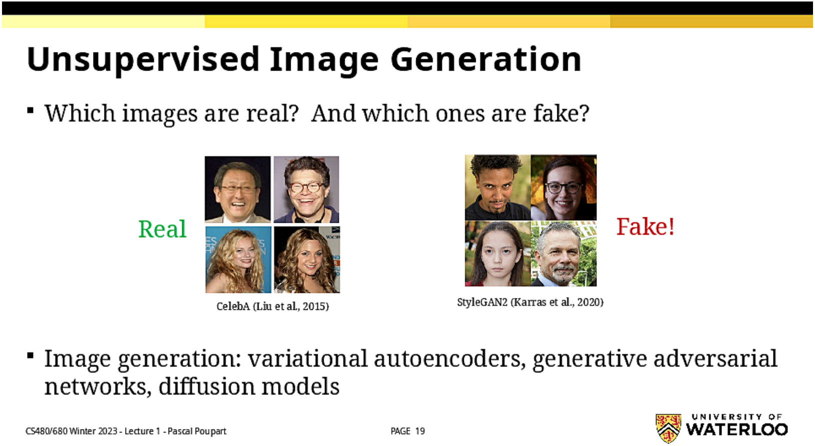
Unsupervised Feature Generation

- Encoder trained on large number of images

CS480/680 Winter 2023 - Lecture 1 - Pascal Poupart

PAGE 18

UNIVERSITY OF WATERLOO



Unsupervised Image Generation

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

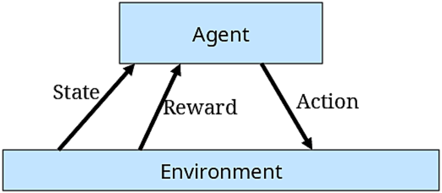
CelebA (Liu et al., 2015)

StyleGAN2 (Karras et al., 2020)

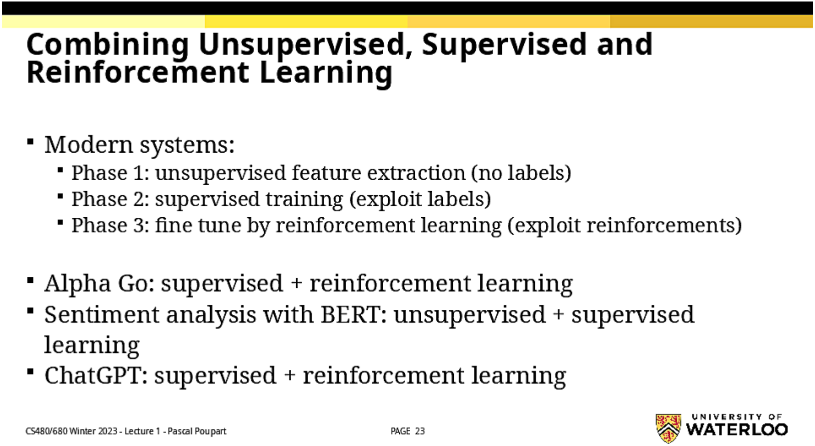
- Image generation: variational autoencoders, generative adversarial networks, diffusion models

CS480/680 Winter 2023 - Lecture 1 - Pascal Poupart PAGE 19

Reinforcement Learning Problem



Goal: Learn to choose actions that maximize rewards



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

CS480/680 Winter 2023 - Lecture 1 - Pascal Poupart PAGE 23

UNIVERSITY OF WATERLOO

This course

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

Lecture Notes I – Examples of Predictors

Marina Meilă
mmp@uwaterloo.ca

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

January 6, 2026

Prediction problems by the type of output

The "learning" paradigm and vocabulary

The Nearest-Neighbor and kernel predictors

Some concepts in Classification

Linear predictors

- Least squares regression
- Linear Discriminant Analysis (LDA)
- QDA (Quadratic Discriminant Analysis)
- Logistic Regression
- The PERCEPTRON algorithm

Classification and regression tree(s) (CART)

The Naive Bayes classifier

Reading HTF Ch.: 2.3.1 Linear regression, 2.3.2 Nearest neighbor, 4.1–4 Linear classification, 6.1–3. Kernel regression, 6.6.2 kernel classifiers, 6.6.3 Naive Bayes, 9.2 CART, 11.3 Neural networks, Murphy Ch.: 1.4.2 nearest neighbors, 1.4.4 linear regression, 1.4.5 logistic regression, 3.5 and 10.2.1 Naive Bayes,4.2.1–3 linear and quadratic discriminant, 14.7.3– kernel regression, locally weighted regression, 16.2.1–4 CART, (16.5 neural nets), Bach Ch.:

Prediction problems by the type of output

In supervised learning, the problem is *predicting* the value of an **output** (or **response** – typically in regression, or **label** – typically in classification) variable Y from the values of some observed variables called **inputs** (or **predictors, features, attributes**) $(X_1, X_2, \dots X_d) = X$. Typically we will consider that the input $X \in \mathbb{R}^d$.

Prediction problems by the type of output

In supervised learning, the problem is *predicting* the value of an **output** (or **response** – typically in regression, or **label** – typically in classification) variable Y from the values of some observed variables called **inputs** (or **predictors, features, attributes**) $(X_1, X_2, \ldots X_d) = X$. Typically we will consider that the input $X \in \mathbb{R}^d$. Prediction problems are classified by the type of response $Y \in \mathcal{Y}$:

- *regression*: $\boxed{Y \in \mathbb{R}}$
- *binary classification*: $Y \in \boxed{\{-1, +1\}}$
- *multiway classification*: $Y \underline{\in \{y_1, \ldots y_m\}}$ a finite set
- *ranking*: $Y \in \mathbb{S}_p$ the set of permutations of p objects
- *multilabel classification* $Y \subseteq \{y_1, \ldots y_m\}$ a finite set (i.e. each X can have several labels)
- *structured prediction* $Y \in \Omega_V$ the state space of a graphical model over a set of [discrete] variables V

Example (**Regression.)**

- Y is the proportion of high-school students who go to college from a given school in given year. X are school attributes like class size, amount of funding, curriculum (note that they aren't all naturally real valued), median income per family, and other inputs like the state of the economy, etc. Note also that $Y \in [0, 1]$ here.
- $Y \geq 0$ is the income of a person, and X_j are attributes like education, age, years out of school, skills, past income, type of employment.
  Economic forecasts are another example of regression. Note that in this problem as well as in the previous, the Y in the previous period, if observed, could be used as a predictor variable for the next Y. This is typical of structured prediction problems.
- Weather prediction is typically a regression problem, as winds, rainfall, temperatures are continuous-valued variables.
- Predicting the box office totals of a movie. What should the inputs be?
- Predicting perovskite degradation. Perovskites are a type of crystal considered promising for the fabrication of solar cells. In standard use, such a material must have a life time Y of 30 years. How can one predict which material will last that long without waiting for 30 years?
  Y is time to degradation, X_j are material composition, experimental conditions, and measurements of initial values of physical parameters.