STAT 535

9/29/22

Lecture 1

Course overview
ML - what's in a name
Prediction problems

L0, L1 posted
Quiz 0

Lecture Notes 0 – Intro to Machine Learning

Marina Meilă
mmp@stat.washington.edu

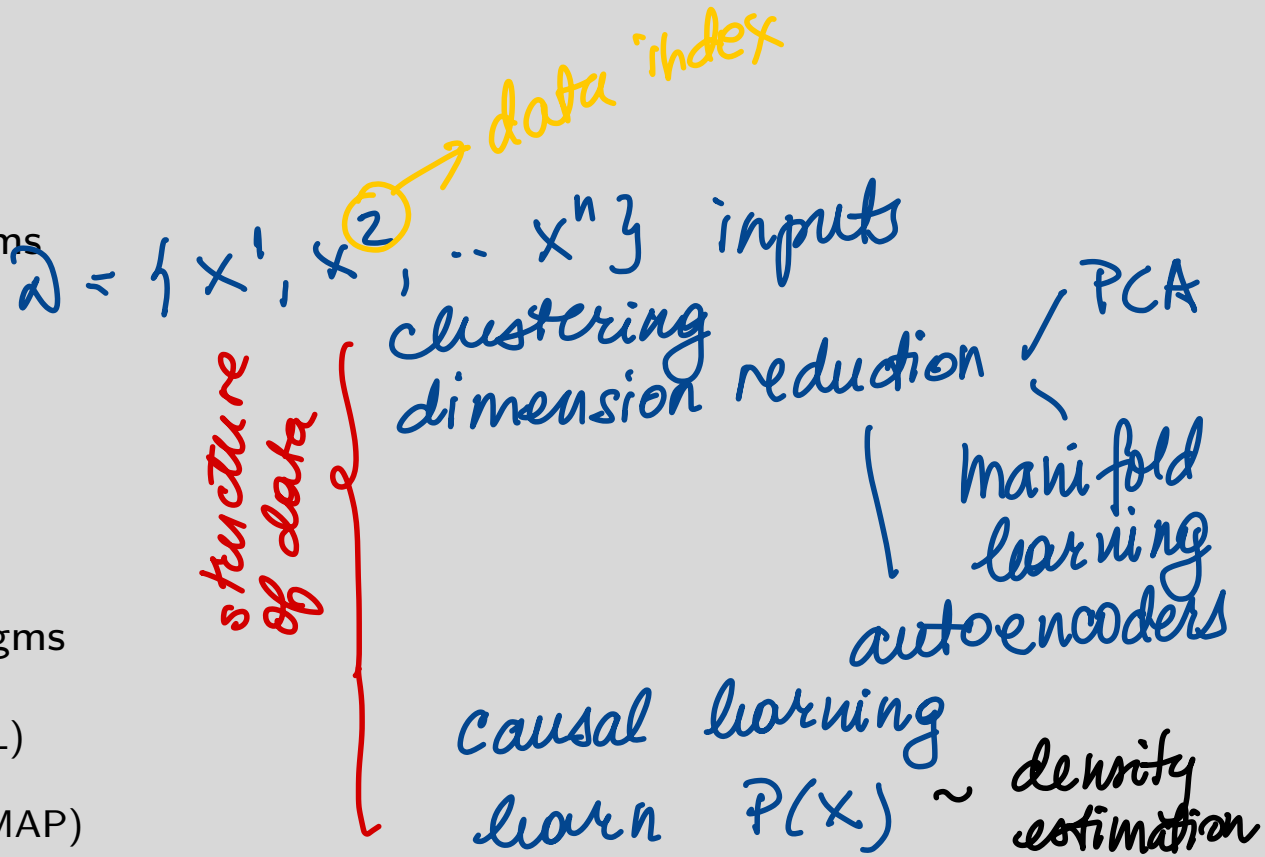
Department of Statistics
University of Washington

September 29, 2022

What's in a name? Or where does "Machine Learning" /"Statistical Learning" come from?

- **What's in this sequence?**
  - Data analysis problems (e.g. clustering, classification)
  - Statistical models (e.g. exponential family models, graphical models)
  - Statistical methods (e.g. Support Vector Machines)
  - Algorithms (e.g. message passing, K-means). There is a continuum between algorithms, methods, and some of the other items on this list.
  - Mathematical facts/concepts from: graph theory, convex analysis
  - Theorems (without proofs), lemmas (with proofs)

Taxonomies



. . . all of them incomplete

- Statistical Learning Problems
  - Unsupervised
  - Supervised
  - (Semi-supervised)
  - Reinforcement
- Statistical models
  - Parametric
  - Non-parametric
- Statistical inference paradigms
  - Bayesian
  - Maximum Likelihood (ML)
  - Penalized Likelihood
  - Maximum A-Posteriori (MAP)

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.

Taxonomies

A = action $\in \mathcal{A}$

R = reward $\in \mathbb{R}$

X = input $\in \mathcal{X}$

~~Y = output~~

. . . all of them incomplete

- Statistical Learning Problems
  - Unsupervised
  - Supervised → $P_{y|x}$
  - (Semi-supervised)
  - Reinforcement → $P_{y|x}$
- Statistical models
  - Parametric
  - Non-parametric
- Statistical inference paradigms
  - Bayesian
  - Maximum Likelihood (ML)
  - Penalized Likelihood
  - Maximum A-Posteriori (MAP)

$\mathcal{D} = \{(x^i, y^i), i = 1:n\}$

$\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U$

$\{x^{n+1}, x^{n+2}, \ldots x^{n+m}\}$

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.

RL = learning to act

Problem: $f: \mathcal{X} \to \mathcal{A}$

$f(x_t) = a_t$ the "best" action at time t

total reward

$\max \sum_{t=1}^{T} r(a_t, x_t)$

Taxonomies

. . . all of them incomplete

- Statistical Learning Problems
  - Unsupervised
  - Supervised
  - (Semi-supervised)
  - Reinforcement
- Statistical models
  - Parametric → Normal (μ, σ^2), Linear regression
  - Non-parametric → Nearest neighbors
- Statistical inference paradigms
  - Bayesian
  - Maximum Likelihood (ML)
  - Penalized Likelihood
  - Maximum A-Posteriori (MAP)

  data, prior distribution, output = posterior distribution

  frequentist ← prior

  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 535

- 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]

g graph data

- [Reinforcement Learning]

Marina Meilămp@stat.washington.edu: Lecture 0 Introduction to ML

Supervised Learning

Lecture Notes I – Examples of Predictors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

September 29, 2022

Prediction problems by the type of output

The "learning" paradigm and vocabulary

The Nearest-Neigbor and kernel predictors

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**) $(X_1, X_2, \ldots X_d) = X$. Typically we will consider that the input $X \in \mathbb{R}^d$.

d- dimensional

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*: $Y \in \mathbb{R}$
- *binary classification*: $Y \in \{-1, +1\}$
- *multiway classification*: $Y \in \{y_1, \ldots y_m\}$ 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

Y ∈ discrete, finite set

Regression – y: "stock market" "weather" – y = temperature amount precipitation

credit scores

guessing unobserved variables

Classification

$y \in \{\pm 1\}$ binary

- pass/fail
- recognition, detection
- (1 class classification)
- anomaly detection

$y \in \{0, 1, 2 \dots m-1\}$

multiway

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.

Example ((Anomaly) detection.)

This is a binary classification problem. $Y \in \{\text{normal}, \text{abnormal}\}$. For instance, Y could be "HIV positive" vs "HIV negative" (which could be abbreviated as "+", "-") and the inputs X are concentration of various reagents and lymph cells in the blood.
Anomaly detection is a problem also in artificial systems, as any device may be functioning normally or not. There are also more general detection problems, where the object detected is of scientific interest rather than an "alarm": detecting Gamma-ray bursts in astronomy, detecting meteorites in Antarctica (a robot collects rocks lying on the ice and determines if the rock is terrestrial or meteorite). More recently, *Artificial Intelligence* tasks like detecting faces/cars/people in images or video streams have become possible.