Lecture 1. Introduction

01/04/2018

What is machine learning

Learning problems and examples

Typical machine learning problems

- prediction
- classification
- pattern recognition
- clustering
- ranking
- dimension reduction and manifold learning

Supervised learning: learning by examples

The supervised learning problem can be characterized as the following statistical inference problem.

- $X \in R^p$: real-valued input p-vector
- $Y \in R$: real-valued outcome variable The goal of a supervised learning procedure is to seek a function, $f(X)$, to predict Y. Note that, the prediction is typically imperfect, the relationship between f, X and Y is summarized as

  $$Y = f(X) + \epsilon,$$

  where ϵ denotes the prediction error by the deterministic function f. The process of supervised learning is to find f based on labeled training data.

Taking a statistical point of view, supervised learning problems can be solved as a general regression problem (note the terminology "regression" has much narrower meaning in machine learning literature), i.e., we may attempt to find $f(X) = E(Y \mid X)$, or equivalently, find $\Pr(Y \mid X)$.

One of the key differences between supervised learning and statistical inference of f lies in the aspect of evaluating f. While the (traditional) statistical theories emphasize more on the "hypothetical" correctness of f (typically in a asymptotic setting), the supervised learning puts more emphasis on its predictive performance.

The two main types of problems in supervised learning are *regression* and *classification* problems. In machine learning literature, the difference between the two is simply the nature of outcome variable Y that is of interest for prediction: if Y is continuous/quantitative, it is considered as a regression problem; if Y is categorical, it is considered as a classification problem.

Unsupervised learning: learning without examples

In unsupervised learning problems, we only deal with one data type: $X \in R^p$, and there is no labeled response variable Y. Some statisticians view unsupervised learning problems as special

cases of density estimation, i.e., the goal is to infer $\Pr(X)$ from the observed data. The main applications of unsupervised learning problems include clustering and latent variable modeling.

Exercise: formulating a clustering problem as a (mixture) density estimation problem.

Other learning scenarios

- semi-supervised learning: only part of training data are labeled. The missing data problem in statistics would be considered as a semi-supervised learning problem.
- online learning: learning with dynamically growing data, involves multiple rounds and training and testing phases are intermixed
- reinforcement learning

Model vs. algorithm (when uncertainty exists)

There is (currently) no agreement on defining model vs. algorithm when dealing with noisy data. Some (typically statisticians) think algorithm can not be defined without a statistical or probabilistic model to explicitly account for uncertainty, some (typically computer scientists) disagree.

A general view on algorithm

- algorithm: a set of computational procedures to map the input data into outcome space

The above definition works for all algorithms regardless if they deal with uncertainty, e.g., sorting algorithm.

A statistical view of model-algorithm relationship when uncertainty exists

- Model: a description of data generative mechanism
- (White box) algorithm: mathematical/computational procedures to fit a model

The distinctions between model and algorithm can be blurred, especially for the black box algorithms. But it is important to note that all algorithms operate on assumptions when dealing with uncertainties, either explicit or implicit, and those assumptions are components of a "model".

It is also important to note that in practice, it may be necessary to twist/change a model because of the complexity of fitting algorithm.

Structure of the course

We hope to cover the following machine learning topics in this course

1. computational learning theory/statistical decision theory
2. probabilistic graphical models for complex system
3. topics on supervised learning
    - linear methods and regularization
    - kernel methods
    - support vector machines
    - ensemble methods and boosting
    - neural networks and deep learning
4. topics on unsupervised learning
    - clustering
    - latent variable models

Other (non-statistical) aspects of machine learning

- Numerical optimization
- Computational complexity of learning algorithm
- Scalable computing and approximation methods
- Programming and implementation of learning algorithm

Some machine learning terminology

- training data/samples
- validation data/samples
- test data/sample
- features
- labels
- hypothesis set (in statistical terms, candidate models)
- regression

Note the different use of vocabulary in statistical literature.