Statistical Machine Learning Introduction

Lecture Notes 0: Introduction

Professor: Zhihua Zhang

What's statistical machine learning? There is a quote from Jordan, "A field that bridges computation and statistics with ties to information theory, signal processing, algorithms, control theory and optimization theory."

Many data can be denoted as matrix. Suppose we have n samples, p variables (features). Then we have

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times p}$$

We can denote the ith sample as $X_i = \begin{pmatrix} X_{11} \\ X_{12} \\ \vdots \\ X_{1p} \end{pmatrix}$.

Statistical machine learning is mainly to solve the following problems:

Dimension Reduction: $X_i \in \mathbb{R}^p$, we want to find $Z_i \in \mathbb{R}^q (p > q)$ to present X_i. If we want to use linear transformation, then our purpose is to find a matrix A so that $Z_i = AX_i$. If we want to use nonlinear transformation, then we need to find a nonlinear function f so that $f(X_i) = Z_i$.

Clustering: We can view the n samples as n points, and cluster them to k clusters, where k is determined by us.

Classification: We have a label Y_i for each X_i. And $Y = (Y_1, Y_2, \cdots, Y_n)^T$. And $Y_i \in C$, where C is a finite set.

Regression: Like classification, but $Y_i \in \mathbb{R}$.

Ranking: Isotonic regression (IR) involves finding a weighted least-squares fit $x \in \mathbb{R}^n$ to a vector $a \in \mathbb{R}^n$ with weights vector $w \in \mathbb{R}^n$ subject to a set of non-contradictory constraints of kind $x_i \geq x_j$.

Multi-task: Like Classification, but Y_i can have multi value.

1 Frequentist's view vs. Bayesian view

1.1 Frequentist's view

The frequentistic approach views the model parameters as unknown constants and estimates them by matching the model to the training data using an appropriate metric.

Example 1.1 *Assume we have n pairs of samples $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. In the future, we want to predict y_j's according to new x_j's.*

Using least mean square, we will get object $L = \sum_{i=1}^{n}(y_i - x_i^T a)^2$, where a is an unknown fixed parameter. Denote the object form as $L = l(y, f(x, a))$. Here $f(x, a) = x_i^T a$, $l(y, f(x, a)) = (y - f(x, a))^2$.

We can add some restricts to the parameters, so we can update $L = l(y, f(x, a)) + \lambda r(a)$, where $r(a)$ is a restrict to a.

1.2 Bayesian view

The Bayesian approach views the model parameters as a random variable and estimates them by using Bayes' theorem.

Example 1.2 *In example 1.1, assume $y \sim \mathcal{N}(x^T a, \sigma^2)$. a and σ are random variables, let $a \sim \mathcal{N}(0, \lambda^2)$, $\sigma^2 \sim \Gamma(\alpha, \beta)$. Our interest is posterior probability $P(a|(x_i, y_i))$*

2 Parametrical view vs. Nonparametrical view

In a parametrical model, the number of parameters is fixed once and for all, independent to the number of the training data. In a nonparametrical model, the number of parameters can change according to the number of training data.

Example 2.1 *In **Nearest Neighbor**, the number of parameters is the number of training samples. So this model is nonparametrical model.*

*In **Logistic Regression**, the number of parameters is the dimension of the training samples. So this model is parametrical model.*