Statistical Machine Learning Basic Theory

Lecture Notes 1: Basic Theory

Professor: Zhihua Zhang

0 Introduction

What's statistical machine learning? Here 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."

In machine learning, data is typically expressed in a matrix form. Suppose we have n samples, p variables (or features). Then we have

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

The ith sample can be denoted as $X_i = (X_{11}, X_{12}, \ldots, X_{1p})^T$.

Machine learning is mainly to solve the following problems:

(1) **Dimension Reduction:** Dimension reduction is the process of reducing the number of random variables(or features) under consideration. Formally, let $X_i \in \mathbb{R}^p$, we want to find $Z_i \in \mathbb{R}^q (q < p)$ to present X_i.

If we use linear transformation, then we need to find a matrix A such that $Z_i = AX_i$. Note that A should be full row rank.

If we use nonlinear transformation, then we need to find a nonlinear function f such that $Z_i = f(X_i)$.

(2) **Clustering:** Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).We can view n samples as n points, and our object is to cluster them into k clusters.

(3) **Classification:** Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Formally, in the training set, we have a label Y_i for each X_i, where $Y_i \in C$, C is a non-empty finite set. If $Y_i \in \{-1, 1\}$ or $\{0, 1\}$, it's a binary classification problem. If $Y_i \in \{1, 2, \ldots, k\}$, it's a multi-class classification problem. There are also problems that one observation belongs to more than one category and they are called multi-label or multi-output classification.

(4) **Regression:** Regression is a particular classification problem in which the label $Y_i \in \mathbb{R}$.

(5) **Ranking:** also called isotonic regression(IR). Isotonic regression 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 $x_i \geq x_j$.

Note that (1),(2) are unsupervised learning, (3),(4),(5) are supervised learning. Unsupervised learning is that of trying to find hidden structure in unlabelled data. Supervised learning is the machine learning task of inferring a function from labelled training data.

For supervised learning, the data is usually split into two or three parts.

(1) **Training data:** A set of examples used for learning, that is to fit the parameters (e.g., weights for neural networks) of the model.

(2) **Validation data:** Sometimes, we also need a validation set to tune the model, for example to choose the number of hidden units in a neural network or for pruning a decision tree. It is usually used to prevent overfitting and enhance the generalization ability.

(3) **Test data:** This data set is used only for testing the final solution in order to confirm the actual performance.

0.1 Frequentist's view vs. Bayesian view

0.1.1 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 0.1 *Suppose we have n pairs of samples* $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ *and we want to fit a linear function* $x_i^T a$*(More strictly, it should be* $x_i^T a + b$ *or include a constant variable 1 in* x_i*) to predict* y_j.

Using least squares, we have loss function $L = \sum_{i=1}^n (y_i - x_i^T a)^2$, *where a is an unknown fixed parameter. We can solve a by minimizing the loss function.*

Using maximum likelihood estimation, let $y_i \sim \mathcal{N}(x_i^T a, \sigma^2)$, *namely,*

$$p(y_i \mid x_i) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{(y_i - x_i^T a)^2}{2\sigma^2}}.$$

So the log likelihood is (assuming the samples are independent)

$$l = log \prod_{i=1}^n p(y_i \mid x_i).$$

We can solve a by maximizing the joint likelihood.

Under the above conditions, you can prove that maximum likelihood estimation is the same as least squares.

0.1.2 Bayesian view

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

Example 0.2 *Let's continue example 1.1, let $y_i \sim \mathcal{N}(x_i^T a, \sigma^2)$ again. Here a and σ are random variables, not constants. Let $a \sim \mathcal{N}(0, \lambda^2)$, $\sigma^2 \sim \Gamma(\alpha, \beta)$. Our interest is the posterior probability $P(a|x_i, y_i) \propto P(x_i, y_i|a)P(a)$. We can use maximum posterior estimation or Bayesian estimation to solve a.*

0.2 Generative vs. Discriminant

A generative model (e.g., naive Bayes) explicitly models the joint probability distribution $P(x, y)$ and then uses the Bayes rule to compute $P(y|x)$. On the other hand, **a discriminative model** (e.g., logistic regression) directly models $P(y|x)$.

0.3 Parametrics vs. Nonparametrics

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 0.3 *In **Nearest Neighbor** method, 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.*

1 Probability Theory Basics

1.1 Sample Space and Events

Definition 1.1 *The sample space Ω is the set of possible outcomes of an experiment, $\omega \in \Omega$ are called sample outcomes, realizations or elements. The subsets of Ω are called events.*

Definition 1.2 *Given an event, $A \subset \Omega$, let $A^c = \{\omega \in \Omega, \omega \notin A\}$ denote the complement of A.*

Definition 1.3 *A sequence of sets $A_1, A_2, \cdots$ is* monotone increasing, *if $A_1 \subset A_2 \subset \cdots$, we define $\lim\limits_{n\to\infty} A_n = \bigcup\limits_{i=1}^{\infty} A_i$.*

Definition 1.4 *A sequence of sets $A_1, A_2, \cdots$ is* monotone decreasing, *if $A_1 \supset A_2 \supset \cdots$, we define $\lim\limits_{n\to\infty} A_n = \bigcap\limits_{i=1}^{\infty} A_i$.*

Example 1.1 *Let $\Omega = \mathbf{R}$ and $A_i = [0, 1/i)$ for $i = 1, 2 \cdots$,then*
$\bigcup\limits_{i=1}^{\infty} A_i = [0, 1)$, $\bigcap\limits_{i=1}^{\infty} A_i = \{0\}$. If $A_i = (0, 1/i)$, then $\bigcup\limits_{i=1}^{\infty} A_i = (0, 1)$, $\bigcap\limits_{i=1}^{\infty} A_i = \emptyset$.

1.2 σ-field and Measures

Definition 1.5 *Let $\mathcal{A}$ be a collection of subsets of a sample space Ω. $\mathcal{A}$ is called σ-field (or σ-algebra). iff*

1. *The empty set $\emptyset \in \mathcal{A}$.*
2. *If $A \in \mathcal{A}$, $A^c \in \mathcal{A}$.*
3. *If $A_i \in \mathcal{A}$, $i \in \{1, 2, ..., k\}$, then $\bigcup_{i=1}^{k} A_i \in \mathcal{A}$.*

Definition 1.6 *A pair $(\Omega, \mathcal{A})$ is called a measurable space.*

Example 1.2 *Let A be a nonempty proper subset of Ω, i.e. $A \neq \emptyset$, $A \neq \Omega$, the smallest $\mathcal{A} = \{\emptyset, \Omega, A, A^c\}$.*

Example 1.3 $\Omega = \mathbb{R}$. *The smallest σ-field that contains all the finite open sets of $\mathbb{R}$ is called Borel σ-field.*

Definition 1.7 *Let $(\Omega, \mathcal{A})$ be a measurable space. A set function ν defined on $\mathcal{A}$ is called a measure iff*

1. $0 \leq \nu(A) \leq \infty$ *for any $A \in \mathcal{A}$.*
2. $\nu(\emptyset) = 0$.
3. *If $A \in \mathcal{A}$, and A_i are disjoint, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, then $\nu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$.*

Definition 1.8 *Tripe $(\Omega, \mathcal{A}, \nu)$ is called a measure space.*

If $\nu(\Omega) = 1$, then ν is called a probability measure and denote it by P. $(\Omega, \mathcal{A}, P)$ is called a probability space.

Example 1.4 *Let Ω be a sample space, $\mathcal{A}$ is a collection of all subsets, and $\nu(A)$ is the number of elements in A.*

Lemma 1.1 *For any two events A and B. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*

Theorem 1.1 (Continuity of Probability) *If $A_n \to A$, then*

$$P(A_n) \to P(A) \text{ as } n \to \infty.$$

Proof: We first consider the case where A_n is monotone increasing.
Recall that $A_1 \subset A_2 \dots$ and let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$.
Define $B_1 = A_1$, $B_2 = \{\omega \in \Omega : \omega \in A_2, \omega \notin A_1\}$, $B_3 = \{\omega \in \Omega : \omega \in A_3, \omega \notin A_2\} \dots$. Then for each n, we have $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$.
Thus, $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$. So that,

$$P(A_n) = \sum_{i=1}^{n} P(B_i)$$

Hence, we have,

$$\begin{aligned}\lim_{n\to\infty} P(A_n) &= \lim_{n\to\infty}\sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i)\\ &= P(\bigcup_{i=1}^{\infty} B_i) = P(A)\end{aligned}$$

For arbitrary sequence $\{A_i\}$, we can define $\{C_i\}$ to construct a monotone increasing sequence. Specifically, $C_1 = A_1 \cap A$, $C_2 = (A_1 \cup A_2) \cap A$, $C_3 = (A_1 \cup A_2 \cup A_3) \cap A$,. . .

1.3 Independent Events

Definition 1.9 *Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.*

We write $A \perp\!\!\!\perp B$ to denote independence. For a set of events $\{A_i, i \in I\}$ A, it is independent if $P(\bigcap_{i\in J} A_i) = \prod_{i\in J} P(A_i)$, for every finite subset J of I.

1.4 Conditional Probability

Definition 1.10 *If $P(B) > 0$, the conditional probability of A given B is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Lemma 1.2 *If A and B are independent events, then $P(A|B) = P(A)$. Also, for any events A, B*

$$P(AB) = P(A|B)P(B) = P(B|A)P(A).$$

1.5 Bayes Theorem

Theorem 1.2 (The Law of Total Probability) *Let $A_1, A_2, \ldots, A_k$ be partition of Ω. Then for any event B,*

$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Proof: Define $C_j = B \cap A_j$ for $j = 1, \ldots, k$. Then we have $C_j \cap C_i = \emptyset$ and $B = \bigcup_{i=1}^{k} C_i$. Thus,

$$P(B) = \sum P(C_j) = \sum P(B \cap A_j) = \sum P(B|A_j)P(A_j)$$

Theorem 1.3 (Bayes Theorem) *Let $A_1, \ldots, A_k$ be a partition of Ω, such that $P(A_i) > 0$ for each i. If $P(B) > 0$, then for each $i = 1, \ldots, k$*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_j)P(A_j)}$$

Remarks: We usually call those probabilities as

- $P(A_i)$ - prior probability of A_i
- $P(A_i|B)$ - posterior probability of A_i
- $P(B|A_i)$ - likelihood

2 Random Variables

Definition 2.1 *A random variable X is a measure map $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each out come Ω and "measurable" means that for every X, $\{\omega : X(\omega) \leq x\} \in \mathcal{A}$.*

Example 2.1 *Flip a coin ten times. Let $X(\omega)$ be a number of heads in the sequence ω. If $w = HHHTTTHHTT$, $X(\omega)$=5.*

Example 2.2 *Let $\Omega = \{(x,y)|x^2 + y^2 \leq 1\}$. Consider drawing a point at random from Ω. $\omega = (x,y) \in \Omega$, $X(\omega) = x, X(\omega) = y, X(\omega) = x + y$ are possible random variables.*

Definition 2.2 *Let $A \subset \mathbb{R}$, $X^{-1} = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{A}$. $P(X \in A) \triangleq P(X^{-1}(A)) = P(\{\omega \in \Omega | X(\omega) \in A\})$. $P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega | X(\omega) = x\})$*

Note for simplicity, we will use $\{X > 0\}$ to denote $\{\omega \in \Omega : X(\omega) > 0\}$, $P(X > 0)$ to denote $P(\{X > 0\})$.

Example 2.3 *Flip a coin twice and let X be the number of heads.*

ω	$P(\{\omega\})$	$X(\omega)$
TT	*1/4*	*0*
TH	*1/4*	*1*
HT	*1/4*	*1*
HH	*1/4*	*2*

X	*P(X)*
0	*1/4*
1	*1/2*
2	*1/4*

2.1 Distribution Function

Cumulative distribution function (or distribution function). CDF is the function $F_X : \mathbb{R} \to [0,1]$.

$F_X(x) = P(X \leq x)$.

Example 2.4 *From example 1.3, we can get*

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

Theorem 2.1 *Let X have CDF F, Y have CDF G. If $F(x) = G(x)$ for all x, then $P(X \in A) = P(Y \in A)$ for all measurable A.*

Theorem 2.2 *A function F mapping $\mathbb{R} \to [0,1]$ is a CDF for probability iff*

1. *F is non-deceasing, $x_1 < x_2 \implies F(x_1) \leq F(x_2)$.*
2. *F is normalized, i.e. $\lim\limits_{x \to -\infty} F(x) = 0$, $\lim\limits_{x \to +\infty} F(x) = 1$.*
3. *F is right-continuous. $F(x) = F(x^+)$, where $F(x^+) = \lim\limits_{y \to x, y > x} F(y)$.*

Now we will get the proof of right-continuous.
Proof: Let $y_1 > y_2 > \cdots$, and $\lim_{n \to +\infty} y_n = x$.
Then $F(y_1) = P(Y \leq y_1)$, $F(y_2) = P(Y \leq y_2)$, ...
Let $A_i = (-\infty, y_i]$ and $A = (-\infty, x]$.
Note that $A = \cap_{i=1}^{\infty} A_i$ and $A_1 \supset A_2 \supset \cdots$
$\lim_{i \to \infty} P(A_i) = P(\cap_{i=1}^{\infty} A_i)$.
$F(x) = P(A) = P(\cap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i) = \lim_{i \to \infty} F(y_i) = F(x^+)$.