Lecture Notes 1: Basic Theory

Professor: Zhihua Zhang *Scribe:Cheng Chen, Luo Luo, Cong Xie*

1 Introduction

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.

1.1 Frequentist's view vs. Bayesian view

1.1.1 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 *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.

1.1.2 Bayesian view

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

Example 1.2 *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* $P(a \mid x_i, y_i) \propto P(x_i, y_i \mid a)P(a)$. *We can use maximum posterior estimation or Bayesian estimation to solve a.*

1.2 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 1.3 *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.*

2 Random Variables and Basic Properties

2.1 Random Variables

Definition 2.1 *Let $X = (X_1, \cdots, X_p)^T$ be a random vector. The cumulative distribution function (C.D.F.) associated with X is: F_X: $F_X(x) = Pr(X \leq x) = Pr(X_1 \leq x_1, \cdots, X_p \leq x_p)$*

There are two kinds of random variables:

- Absolutely continuous
- Discrete

2.1.1 Absolutely Continuous Random Variables

Definition 2.2 *X is a absolutely continuous random variable if there exists a probability density function (P.D.F.) $f_X(x)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ for absolutely continuous variables where $F_X(\infty) = \int_{-\infty}^{+\infty} f(u)du = 1$.*

2.1.2 Discrete Random Variables

Definition 2.3 *X is a discrete random variable if there exists a probability mass function (P.M.F.) such that X takes countable (or finite) set of points $\{X_j, j = 1, \cdots\}$ in which P.M.F. is*

$$\begin{cases} f_X(x_j) = P(X_j = x_j) \\ f_X(x) = 0, otherwise \end{cases}$$

where $P(x \in D) = \int_D f(u)du$ and $D \subseteq \mathbb{R}^p$.

2.1.3 Support Set

Definition 2.4 *Support set $S = \{x \in \mathbb{R}^p, f_X(x) > 0\}$*

2.1.4 Marginal and Conditional Distribution

For $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ where $x_1 \in \mathbb{R}^k$, $x_2 \in \mathbb{R}^{p-k}$,

Definition 2.5 *Marginal distribution: If there exists* $Pr(X_1 \leq x_1) = F_X(x_1, \cdots, x_k, \infty_{k+1}, \cdots, \infty_p)$ *where* $X \sim P, D.F. f(x)$, *then* $f_1(x_1) = \int_{-\infty}^{+\infty} f(x)dx_2$ *is the marginal distribution function.*

Definition 2.6 *Conditional distribution: By Bayes' Theory,* $f_{x_2}(x_2|X_1 = x_1) = \frac{f(x_1,x_2)}{f_1(x_1)}$ *is the conditional distribution function.*

Example 2.1 *For* $f(x_1, x_2) = 1$, *the marginal distribution functions are* $f_1(x_1) = 1$, $f_2(x_2) = 1$ *where* $0 < x_1, x_2 < 1$.

Example 2.2 *For* $f(x_1, x_2) = 1 + \alpha(2x_1 - 1)(2x_2 - 1)$, *the marginal distribution functions are* $f_1(x_1) = 1$, $f_2(x_2) = 1$ *where* $0 < x_1, x_2 < 1$ *and* $-1 \leq \alpha \leq 1$.

2.1.5 Independence

If x_1 and x_2 are independent, then the conditional P.D.F. $f_2(x_2|x_1) = f_2(x_2)$.

Theorem 2.1 *If* X_1 *and* X_2 *are statistically independent, then* $f(x) = f(x_1, x_2) = f_1(x_1)f_2(x_2)$.

2.2 Population Moments

2.2.1 Expectation

Definition 2.7 X *is a random variable with* $F_X(x)$, *then the expectation or mean of a scalar-valued function g(X) is*

$$\begin{aligned} &\mathbb{E}(g(x)) \\ &= \int_{-\infty}^{+\infty} g(x)dF(x) \\ &= \begin{cases} \sum\limits_x g(x)f(x) \text{ if } x \text{ is discrete} \\ \int_{-\infty}^{+\infty} g(x)f(x)dx \text{ if } x \text{ is continuous} \end{cases} \end{aligned}$$

Proposition 2.1 *Basic Properties of Expectation*

- *Linearity:*
  $\mathbb{E}(\alpha_1 g_1(x) + \alpha_2 g_2(x)) = \alpha_1 \mathbb{E}(g_1(x)) + \alpha_2 \mathbb{E}(g_2(x))$
- *Partition:*

For $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$,

$$\begin{aligned} &\mathbb{E}\{g_1(x_1)\} \\ &= \int_{-\infty}^{+\infty} g_1(x_1) f(x) dx \\ &= \int_{-\infty}^{+\infty} g_1(x_1) f(x_1, x_2) dx \\ &= \int_{-\infty}^{+\infty} g_1(x_1) f_1(x_1) dx_1 \end{aligned}$$

- *Independence:*
  If $g_i(x_i), i = 1, 2$, *then* $\mathbb{E}(g_1(x_1)g_2(x_2)) = \mathbb{E}(g_1(x_1))\mathbb{E}(g_2(x_2))$

Definition 2.8 *In general we denote* $G = g_{ij}(X)$ *as the matrix-value function of* X. *And* $\mathbb{E}(G(x)) = (\mathbb{E}g_{ij}(x))$.

2.2.2 Population Mean and Covariance Matrix

Definition 2.9 *We denote* $\mathbb{E}(x) = \mu = \begin{pmatrix} \mu_1 \\ \cdots \\ \mu_p \end{pmatrix}$ *where* $\mu_i = \int_{-\infty}^{+\infty} x_i f_i(x) dx_i$. *Thus the covariance matrix is* $\mathbb{E}((x-\mu)(x-\mu)^T) = \Sigma = V(x)$ *where* $\Sigma = (\sigma_{ij})$

Proposition 2.2 *Basic Properties of Covariance Matrix*

- $\sigma_{ij} = C(x_i, x_j) = \mathbb{E}((x_i - \mu_i)(x_j - \mu_j))$, $\sigma_{ii} = V(x_i)$
- $\Sigma = \mathbb{E}(xx^T) - \mu\mu^T$
- $V(a^T x) = a^T V(x) a = \sum_{i,j=1}^{p} a_i a_j \sigma_{ij}$
- $\Sigma \succcurlyeq 0$ *and it is symmetric.*
- $V(Ax + b) = AV(x)A^T$
- $\Delta = diag(\sigma_{ij})$, $\rho = \Delta^{-\frac{1}{2}} \Sigma \Delta^{-\frac{1}{2}}$ *which is called the correlation matrix.*
  $\rho = (\rho_{ij})$, $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}^{\frac{1}{2}} \sigma_{jj}^{\frac{1}{2}}} = \frac{x_i^T x_j}{||x_i||_2 ||x_j||_2}$

2.2.3 Sample Mean and Covariance Matrix

Definition 2.10 *Sample mean is* $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ *in which* $\mathbb{E}(\overline{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(x_j) = \mu$

Definition 2.11 *Sample covariance is*

$$\begin{aligned} S &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^T \\ &= \frac{1}{n-1}x^T(I_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)(I_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)x \\ &= \frac{1}{n-1}x^T(I_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)x \\ &= \frac{1}{n-1}xH_nx \end{aligned}$$

where H_n is the centering matrix. And we can prove that $\mathbb{E}(S)=\Sigma$.