DDA5001 Machine Learning

Basic Math & Concepts of Leanring

Xiao Li

School of Data Science
The Chinese University of Hong Kong, Shenzhen

Basic Mathematics

Concepts of Learning

Basic Notions of Linear Algebra

Basic Notions of Linear Algebra

- Vector. $\boldsymbol{x} \in \mathbb{R}^n$ is a real-valued n-dimensional column vector; i.e.,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i \in \mathbb{R} \; \forall i.$$

- You can regard the vector $\boldsymbol{x} \in \mathbb{R}^n$ as a point in the n-dimensional linear space $\mathbb{R}^n$ (Think of $n = 2$ and $n = 3$).
- Addition of vectors. The addition of two vectors is defined by adding corresponding coordinates, i.e.,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix},$$

Basic Notions of Linear Algebra

- Multiplication. The multiplication of a scalar with a vector is defined by performing multiplication in each coordinate:

$$a \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix},$$

  where $a \in \mathbb{R}$.
- Commutativity. $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
- Distributive properties $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$ and $(a+b)\boldsymbol{x} = a\boldsymbol{x} + b\boldsymbol{x}$ for all $a, b \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.
- Transpose of vector. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$. The notation $\boldsymbol{x}^\top$ means that

$$\boldsymbol{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}.$$

Basic Notions of Linear Algebra

- Linear independence. We say that a finite collection $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ of vectors in $\mathbb{R}^n$ is linearly dependent if there exist scalars $a_1, \ldots, a_m \in \mathbb{R}$, not all of them are zero, such that

$$\sum_{i=1}^{m} a_i \boldsymbol{x}_i = 0.$$

  The collection $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ is said to be linearly independent if it is not linearly dependent.

- Span. The set of all linear combinations of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ is called the span of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, i.e., $\mathrm{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} := \{\sum_{i=1}^{m} a_i \boldsymbol{x}_i : \boldsymbol{a} \in \mathbb{R}^m\}$
- Basis. A basis of the n-dimensional space $\mathbb{R}^n$ is a collection of vectors in $\mathbb{R}^n$ that is linearly independent and spans $\mathbb{R}^n$. For example,

$$\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix} \right\} \quad \text{and} \quad \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

  are bases of $\mathbb{R}^2$.

Basic Notions of Linear Algebra

- Inner product. Given two vectors $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^n$, their inner product is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

  We say that $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ are orthogonal if $\boldsymbol{x}^\top \boldsymbol{y} = 0$.

- (Euclidean) ℓ_2-norm. For vector $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$,

$$\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

  which measures the length of $\boldsymbol{x}$. For simplicity, we often only write $\|\boldsymbol{x}\|$ to represent $\|\boldsymbol{x}\|_2$.

- More generally, a norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a function that satisfies
  - $\|\boldsymbol{x}\| > 0$ for all $\boldsymbol{x} \neq 0$ and $\|\boldsymbol{x}\| = 0$ only if $\boldsymbol{x} = 0$;
  - $\|a\boldsymbol{x}\| = |a|\|\boldsymbol{x}\|$ for $\boldsymbol{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}$;
  - $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ (triangle inequality)

Basic Notions of Linear Algebra

- Hölder p-norm. We now introduce common norms in $\mathbb{R}^n$—the Hölder p-norm, $1 \leq p \leq \infty$, which are defined as follows:

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

- Special cases. When $p = 2$, it reduces to the ℓ_2-norm. When $p = 1$, it reduces to the ℓ_1-norm, i.e.,

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

- Cauchy-Schwarz inequality.

$$\boldsymbol{x}^\top \boldsymbol{y} \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Basic Notions of Linear Algebra

- Matrix. We use $\mathbb{R}^{m \times n}$ to denote the set of $m \times n$ arrays whose components are from $\mathbb{R}$. We can write a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad a_{i,j} \in \mathbb{R} \ \forall i, j.$$

- Transpose of Matrix. Given an $m \times n$ matrix $\boldsymbol{A}$, its transpose $\boldsymbol{A}^\top$ is defined as the following $n \times m$ matrix:

$$\boldsymbol{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix},$$

- Symmetric matrix. An $m \times m$ real matrix A is said to be symmetric if $\boldsymbol{A} = \boldsymbol{A}^\top$.

Basic Notions of Linear Algebra

- Matrix-matrix multiplication. The matrix-matrix multiplication between $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$ is defined as

$$\mathbb{R}^{m \times p} \ni \boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} \quad \text{where} \quad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Illustration:

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

The matrix-vector multiplication can be viewed as a special case of matrix-matrix multiplication, i.e., with $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^{n}$, we have

$$\mathbb{R}^{m} \ni \boldsymbol{c} = \boldsymbol{A}\boldsymbol{b} \quad \text{where} \quad c_{i} = \sum_{k=1}^{n} a_{ik} b_{k}.$$

Basic Notions of Linear Algebra

- Three perspectives for matrix-matrix multiplication. There are three (equivalent) important ways for interpreting $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$:
  - The first one is by definition

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \quad \forall i = 1, 2, \ldots, m. \ j = 1, 2, \ldots, p.$$

  - The second one is by outer product

$$\boldsymbol{C} = \sum_{k=1}^{n} \boldsymbol{a}_k \boldsymbol{b}_k^\top,$$

    where $\boldsymbol{a}_k$ and $\boldsymbol{b}_k^\top$ are the k-th column and row of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.
  - The third one is by matrix-vector product

$$\boldsymbol{c}_j = \boldsymbol{A}\boldsymbol{b}_j, \quad \forall j = 1, 2, \ldots, p.$$

Basic Notions of Linear Algebra

- Rank. The rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, denoted by $\operatorname{rank}(\boldsymbol{A})$, is defined as the number of elements of a maximal linearly independent subset of its columns or rows. Some facts about the rank of a matrix:
  - $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}^\top)$;
  - $\operatorname{rank}(\boldsymbol{A} + \boldsymbol{B}) \leq \operatorname{rank}(\boldsymbol{A}) + \operatorname{rank}(\boldsymbol{B})$;
  - $\operatorname{rank}(\boldsymbol{AB}) \leq \min\{\operatorname{rank}(\boldsymbol{A}), \operatorname{rank}(\boldsymbol{B})\}$.
- Matrix inverse. An $n \times n$ square matrix $\boldsymbol{A}$ is said to be invertible if the columns of $\boldsymbol{A}$ has full-rank. The inverse of the matrix $\boldsymbol{A}$ is denoted as $\boldsymbol{A}^{-1}$, and we have

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbf{I}.$$

  Facts:
  - $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.
  - $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$, where $\boldsymbol{A}, \boldsymbol{B}$ are square and invertible.

Basic Notions of Linear Algebra

- Orthogonal matrix. An $n \times n$ square matrix $\boldsymbol{A}$ is said to be orthogonal, or orthonormal, is a real square matrix whose columns and rows are orthonormal vectors. That is,

$$\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \mathbf{I}$$

  In another word, for orthogonal matrix $\boldsymbol{A}$, we have

$$\boldsymbol{A}^\top = \boldsymbol{A}^{-1}.$$

- Positive semi-definite (definite), abbrev. PSD (PD), matrix. An $n \times n$ real matrix $\boldsymbol{A}$ is said to be PSD (PD) if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ (> 0) for all $\boldsymbol{x} \in \mathbb{R}^n$ (for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{0\}$).

Basic Notions of Multivariate Calculus

Basic Notions of Multivariate Calculus

- Gradient. It is a generalization of derivative to multi-dimensional functions. Assume $f(\boldsymbol{x}) = f(x_1, x_2, ..., x_n)$ is continuously differentiable. Then, we denote the gradient of f by (an $n \times 1$ vector):

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

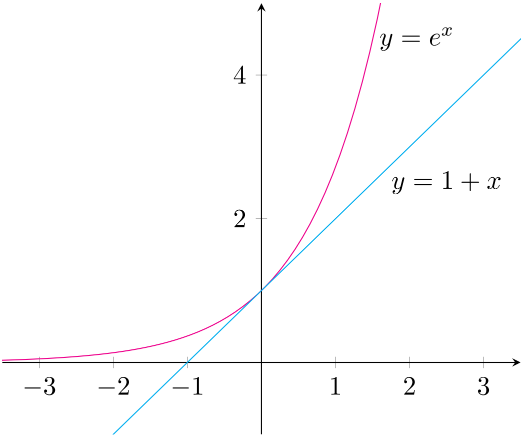
  Facts:
  - If $f(\boldsymbol{x}) = \boldsymbol{c}^\top \boldsymbol{x}$, then $\nabla f(\boldsymbol{x}) = \boldsymbol{c}$.
  - If $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{x}$ ($\boldsymbol{M}$ is symmetric), then: $\nabla f(\boldsymbol{x}) = 2\boldsymbol{M}\boldsymbol{x}$.

- First-order Taylor expansion. The first-order Taylor expansion yields:

$$f(\boldsymbol{x} + \boldsymbol{d}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{d} + o(\|\boldsymbol{d}\|), \quad \|\boldsymbol{d}\| \to 0.$$

Basic Notions of Multivariate Calculus

Illustration of first-order Taylor expansion:



- Approximate the function very well around $\boldsymbol{x}$.
- Important notion for later first-order algorithm development.

Basic Notions of Probability and Statistics

Basic Notions of Probability and Statistics

- Expectation. Suppose X is a random variable, its expectation is denoted as
$$\mathbb{E}[X].$$
Suppose X takes discrete values $x_1, \ldots, x_k$ with probability $p_1, \ldots, p_k$, then
$$\mathbb{E}[X] = \sum_{i=1}^{k} p_i x_i.$$
Suppose X takes continuous values in $(-\infty, +\infty)$ with density $p(x)$, then
$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} p(x) x dx.$$
- Variance. Suppose X is a random variable, its variance is denoted as
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Basic Notions of Probability and Statistics

- Random vector. $\boldsymbol{X} = [X_1, \ldots, X_n]^\top$ is a random vector if each coordinate is a random variable.
- Expectation of random vector. Suppose $\boldsymbol{X}$ is an n-dimensional random vector, its expectation is denoted as

$$\mathbb{E}[\boldsymbol{X}] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top.$$

- Covariance matrix. Suppose $\boldsymbol{X} = [X_1, \ldots, X_n]^\top$ is an n-dimensional random vector, its covariance matrix is $n \times n$ matrix defined as

$$\mathrm{Var}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^\top].$$

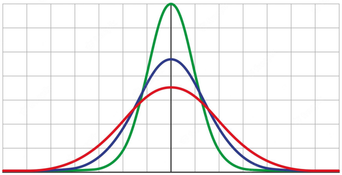
Basic Notions of Probability and Statistics

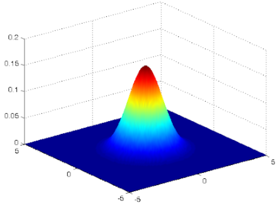
- Gaussian distribution. A random variable X is said to follow $\mathcal{N}(\mu, \sigma^2)$ (Gaussian distribution with mean μ and variance σ^2) if its probability density function (PDF) is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right)$$

- Multivariate Gaussian distribution. We say the random vector $\boldsymbol{X} \in \mathbb{R}^d$ follows Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (assumed to be PD), if its PDF is given by

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right)$$





Basic Notions of Optimization

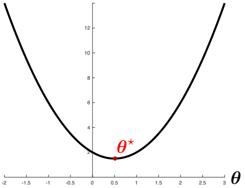
- Optimization. The seek of maximum or minimum. Formally speaking, finding the minimum value of f over $\mathbb{R}^n$ is written as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^n} f(\boldsymbol{\theta}).$$

- Global minimizer. Find the point $\boldsymbol{\theta}^\star$ (called global minimizer/global optimum/optimal solution) that achieves the minimum value of f over $\mathbb{R}^n$

$$\boldsymbol{\theta}^\star = \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^n} f(\boldsymbol{\theta}).$$

Clearly, $f(\boldsymbol{\theta}^\star) = \min_{\boldsymbol{\theta} \in \mathbb{R}^n} f(\boldsymbol{\theta})$.



Basic Mathematics

Concepts of Learning

Components of Supervised Learning: Motivation from An Example

Learning Example: Credit Approval

Task: Learning to predict if one applicant should be approved a credit card.

Applicant's info.:

age	25
gender	male
salary	100000 RMB
citizenship	CN
years in job	2 year
⋮	⋮

Question: should we approve credit card to the applicant?

How to automate such a task by using machine learning methods?

Data: Samples

- Collect a series of historical data

	age	gender	salary	citizenship	years in job
Applicant 1	2.5	1	10	3	1
Applicant 2	2.8	0	8	6	5
Applicant 3	1.6	0	0	4	0
Applicant 4	2.3	1	8	2	4
Applicant 5	3	0	4	2	1

We have the data matrix (feature matrix) $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{bmatrix} 2.5 & 1 & 10 & 3 & 1 \\ 2.8 & 0 & 8 & 6 & 5 \\ 1.6 & 0 & 0 & 4 & 0 \\ 2.3 & 1 & 8 & 2 & 4 \\ 3 & 0 & 4 & 2 & 1 \end{bmatrix}$$

- Each row $\boldsymbol{x}_i^\top$ is called a sample, representing i-th applicant's data
- Each column is called a feature, repenting all the applicant's behavior about the j-th feature.

Data: Labels

- Collect the corresponding label

	age	gender	salary	citizenship	years in job	approve
App. 1	2.5	1	10	3	1	+1
App. 2	2.8	0	8	6	5	+1
App. 3	1.6	0	0	4	0	-1
App. 4	2.3	1	8	2	4	+1
App. 5	3	0	4	2	1	-1

We have the label $\boldsymbol{y}$ as

$$\boldsymbol{y} = \begin{bmatrix} +1 \\ +1 \\ -1 \\ +1 \\ -1 \end{bmatrix}$$

- Each y_i represents the label of the i-th applicant.
- Label implies supervision.

Supervised Learning: Hypothesis/Model

- We have an underlying and unknown hypothesis/model $g \in \mathcal{H}$

$$g : \mathcal{X} \mapsto \mathcal{Y}$$

  where $\mathcal{X}$ is the input space (set of all possible inputs), while $\mathcal{Y}$ is the output space (label space).
  In our example, g is the target function that maps $\boldsymbol{x}_i$ to y_i.

- Learn a model f from the hypothesis/model space $\mathcal{H}$ based on the training dataset $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$.
  Ideally, f should fully capture the patterns in data, i.e., it approximates well the target function g

$$f \quad \approx \quad g.$$

- The hypothesis space $\mathcal{H}$ is one of the hardest parts to be pre-determined in a learning process. One typical instance of $\mathcal{H}$ is the set of all possible linear fit to the data (results in linear models), while another popular choice is nonlinear model (e.g., neural networks).

Supervised Learning: Hypothesis/Model

Parametrization:

$$f = f_{\boldsymbol{\theta}} \in \mathcal{H} \text{ is often parameterized by the parameters } \boldsymbol{\theta}$$

Example:

- In linear regression, $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$ is all possible linear fits and $\boldsymbol{\theta}$ is the parameters of the model. A specific $\boldsymbol{\theta}$ determines a specific model.
- In deep learning, $f_{\boldsymbol{\theta}}$ is the neural network and $\boldsymbol{\theta}$ represents weights (network parameters), respectively.

Two main categories of hypothesis space $\mathcal{H}$:

- Linear
  - Linear regression
  - Linear classification
- Nonlinear
  - Neural networks

Supervised Learning: Learning Problem and Algorithm

- Given training dataset $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.
- Choose the hypothesis $f_{\boldsymbol{\theta}}$.
- Choose the loss function $\ell : \mathbb{R} \to \mathbb{R}$.
- Learning/optimization problem

$$\boxed{\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i\right)} \tag{P}$$

$\rightsquigarrow$ Optimization algorithm $\mathcal{A}$ is designed to solve **(P)**.

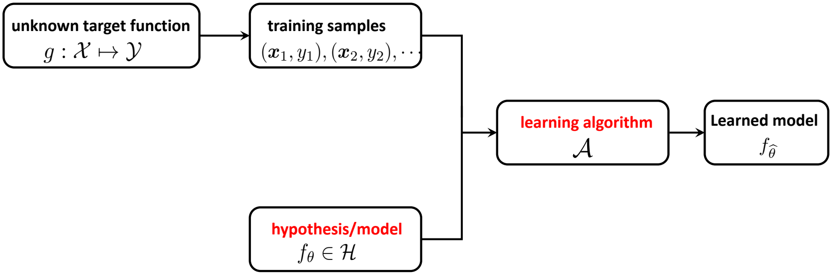
$\rightsquigarrow$ After learning to obtain $\widehat{\theta}$, we get the learned model $f_{\widehat{\boldsymbol{\theta}}}$. Then, one can use the learned $f_{\widehat{\boldsymbol{\theta}}}$ to do prediction.

Supervised Learning: Components

Formalization:

- Target function $g : \mathcal{X} \to \mathcal{Y}$ (underlying credit approval model)
- Training dataset: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ (historical records)
- Hypothesis space $\mathcal{H}$ (learning scope to approximate g)
- Hypothesis/model: $f_{\boldsymbol{\theta}}$ (model to be determined)
- Optimization algorithm: $\mathcal{A}$ (learning the model from data)

Supervised Learning: High-level View



Hopefully,

$$f_{\widehat{\boldsymbol{\theta}}} \approx g$$

Predict/decision: When a new sample data (test data) $\boldsymbol{x}$ comes, the label is predicted as:

$$y \leftarrow f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}).$$

$\rightsquigarrow$ Next lecture: Linear classification and linear regression.