Theory of Statistical Mathematics: Guidance

- Lecturer: T. Kanamori (e-mail: kanamori@c.titech.ac.jp)
  http://www.kana-lab.c.titech.ac.jp/2019-statmath.html

- Course Schedule
  - Guidance. A brief review of Probability
  - Regression and Classification: Kernel methods
  - Statistical Learning theory
  - Deep Learning

- Assessment criteria and methods
  - Evaluated by report submission

- Reference books, course materials, etc.

  - Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.
  - smoe handouts

— Framework of Machine Learning —

Purpose of Data Analysis: extract useful information from observed data.

- In this course, mainly we learn some statistical methods for regression and classification problems.

Problem Setup

Input: x, Output: y.

$$\text{ex.} \quad x \longrightarrow \boxed{??} \longrightarrow y$$

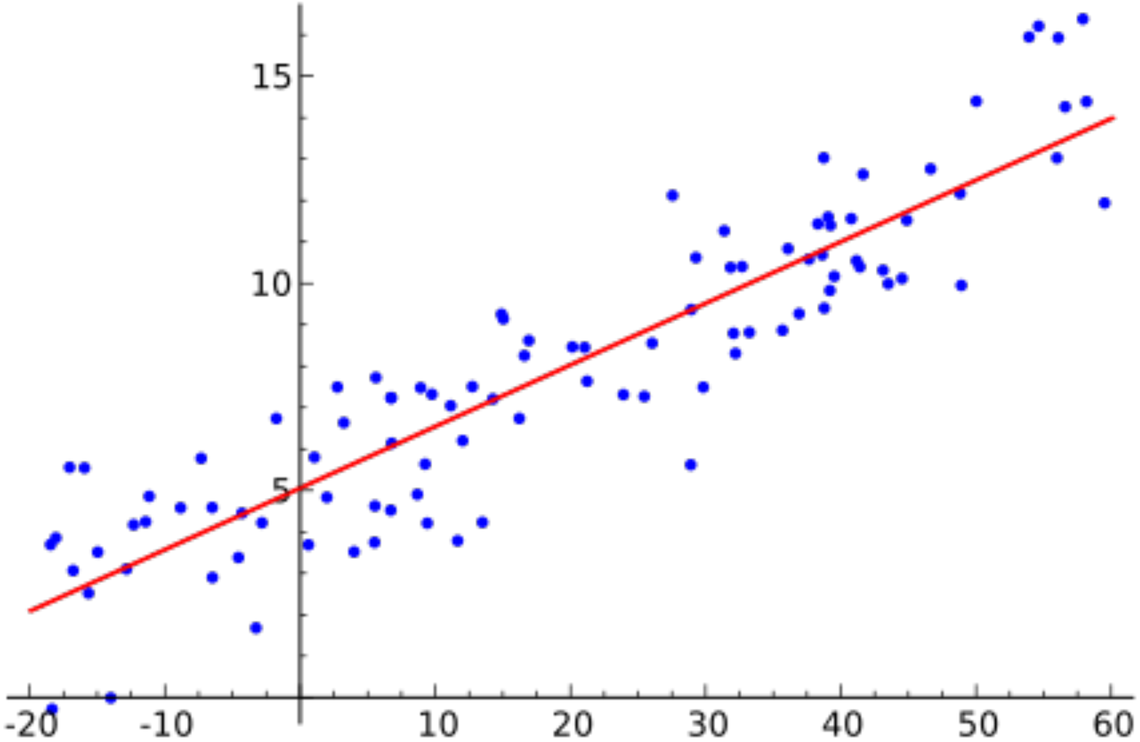
- training samples $(x_1, y_1), \ldots, (x_n, y_n)$ are observed.

- predict the output y of a new input point x.

  * Regression: y is continuous.
  * Classification: y is discrete finite.

Regression

y can take a real number, i.e., y is a continuous random variable.

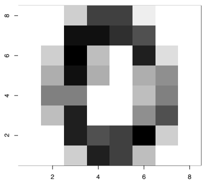
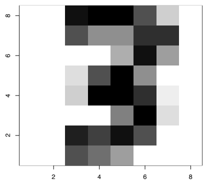
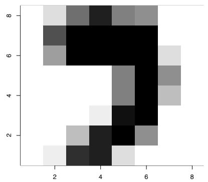


Classification

The candidate of y is finite, i.e., y is a discrete random variable.

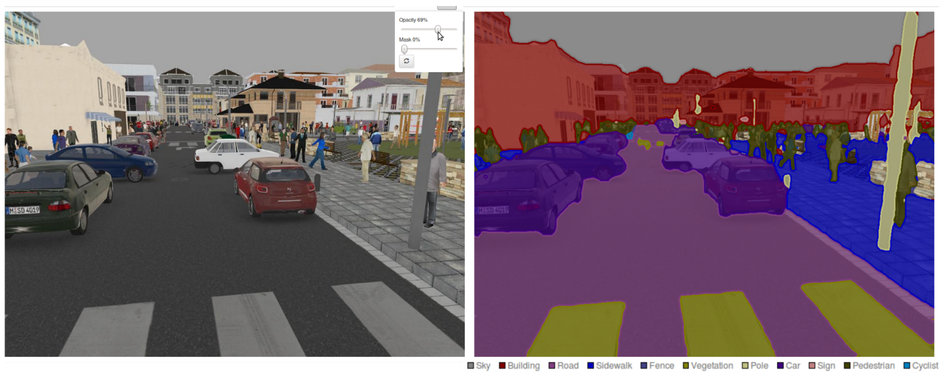
- Character Recognition:

  ex.: $\boldsymbol{x} \in \mathbb{R}^{64}$, $y \in \{0, 1, 2, \ldots, 9\}$.

   $\rightarrow$ 0,  $\rightarrow$ 3,  $\rightarrow$ 7 or 9?

  used in reading system of handwritten zip codes.

- Image segmentation:

Statistical Data Analysis and Probability

- Observed data is often contaminated by noise.

- Probability is useful for data analysis:

$$\text{model} \;=\; [\text{non-random structure}] \;+\; [\text{random noise}]$$

$*$ In this course, we do not go into details of measure theory.

— A Review of Probability: —

- random variable(r.v.): variable whose possible values are outcomes of a random phenomenon. Upper case characters such as X, Y are commonly used to denote r.v.

  e.g. coin flipping

- Let X be a r.v. taking the value in the sample space Ω. The definition of the probability $\Pr(\cdot)$ is given by the following conditions.

Axiom of Probability:

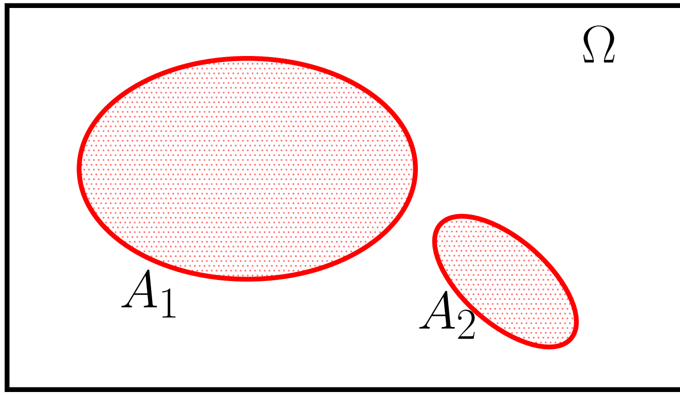
1. For the subset $A \subset \Omega$, $0 \leq \Pr(X \in A) \leq 1$.

2. The probability of the whole event Ω is 1, i.e. $\Pr(X \in \Omega) = 1$

3. For mutually disjoint events A_i, $i = 1, 2, 3, \ldots$,

$$\Pr(X \in \cup_i A_i) = \sum_i \Pr(X \in A_i).$$

(mutually distjoint: $A_i \cap A_j = \phi$ for $i \neq j$)



$\Pr(X \in A)$ is often written as $\Pr(A)$ or $P(A)$.

Example 1 (coin flip). *Let* (X, Y) *be r.v. corresponding to flipping two coins and, define* $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$. *For the fair coins,*

$$\Pr((X,Y)=(x,y)) = \frac{1}{2^2}, \quad (x,y) \in \Omega,$$

$$\begin{aligned}\Pr(X=1) &= \Pr((X,Y) \in \{(1,0),(1,1)\}) \\ &= \Pr((X,Y)=(1,0)) + \Pr((X,Y)=(1,1)) = \frac{1}{2}\end{aligned}$$

calculation of probability

The following equations are derived from axioms.

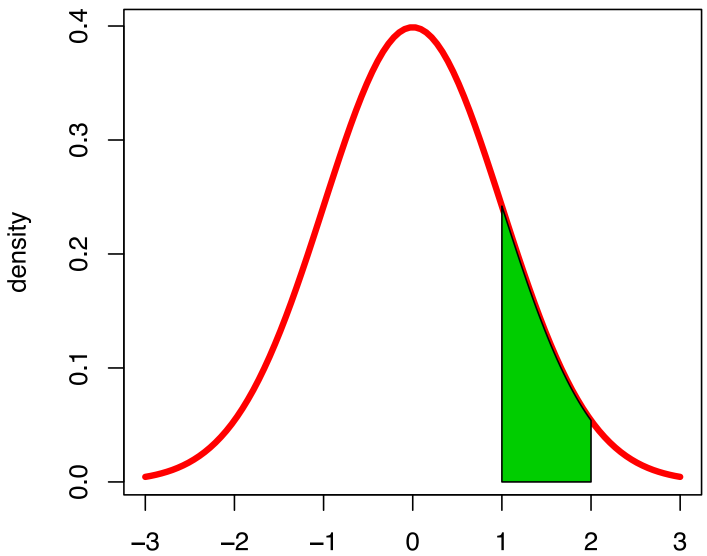
- $\Pr(A) + \Pr(A^c) = 1$, where A^c is the complement of A, i.e., $A^c = \{x \in \Omega \,|\, x \notin A\}$.
- monotonicity: $A \subset B \subset \Omega \implies \Pr(A) \le \Pr(B)$.
  Proof: if $A \subset B$, we have $B = A \cup (B \cap A^c)$, i.e., mutually disjoint.
  Thus, $\Pr(B) = \Pr(A) + \Pr(B \cap A^c) \ge \Pr(A)$.
- Addition theorem: $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
  $\Pr(\cup_i A_i) \le \sum_i \Pr(A_i)$.

Exercise 1. *Prove* $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

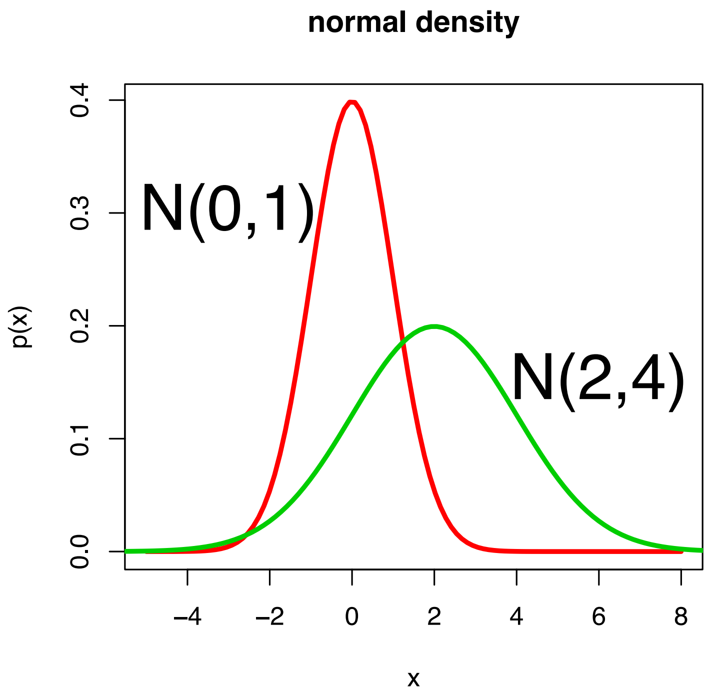
Example: 1-dim normal distribution (Gaussian distribution),

$$\Pr(a \le X \le b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx, \qquad (\Omega = \mathbb{R})$$

This is expressed as $X \sim N(\mu, \sigma^2)$.



area of ■ $= \Pr(1 \le X \le 2)$



Probability Density Function (pdf)

For n random variables: $X_1, X_2, \ldots, X_n$ and a set $A \subset \mathbb{R}^n$,

Probability of $X = (X_1, X_2, \ldots, X_n) \in A$ is supposed to be given by

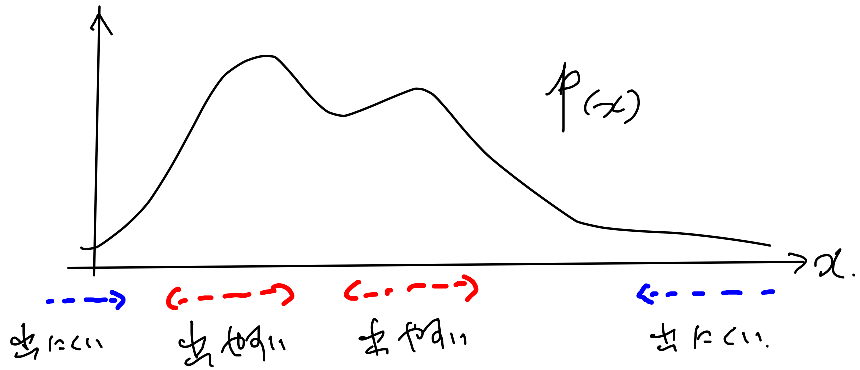
$$\Pr(X \in A) = \int_A p(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

the function $p(x_1, \ldots, x_n)$ is called the (joint) probability density function.

- (joint)probability density function:

  $p(\boldsymbol{x}) = p(x_1, \ldots, x_n) \geq 0,$

  $$\int_{\mathbb{R}^n} p(\boldsymbol{x}) d\boldsymbol{x} = 1$$



- marginal pdf: $p_1(x_1) = \displaystyle\int_{\mathbb{R}^{n-1}} p(x_1, x_2, \ldots, x_n) dx_2 \cdots dx_n$ etc.

For discrete r.v., i.e., A is a countable set, $\displaystyle\int_A \cdots d\boldsymbol{x}$ is replaced with $\displaystyle\sum_{\boldsymbol{x} \in A} \cdots$.

=== 2019-4-9(Tue): up to here ===

Expectation and Variance of r.v.

Let $p(x_1, \ldots, x_d)$ be the pdf of $X = (X_1, \ldots, X_d)$.

- Expectation of X_i: barycenter

$$\mathbb{E}[X_i] = \int_{\mathbb{R}^d} x_i \, p(x_1, \ldots, x_d) dx_1 \cdots dx_d = \int_{\mathbb{R}} x_i \, p_i(x_i) dx_i \ \in \mathbb{R}$$

$$P(X_1 = 1) = 0.5, P(X_1 = 2) = 0.2, P(X_1 = 3) = 0.3$$
$$\Rightarrow \mathbb{E}[X_1] = 1 \times 0.5 + 2 \times 0.2 + 3 \times 0.3 = 1.8$$

- Expectation of the d-dimensional r.v. $X = (X_1, \ldots, X_d)^T$:

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^T \in \mathbb{R}^d$$

$*$ $a, b \in \mathbb{R}$, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ holds.

- Variance of 1-dim r.v.X: it measures how far a set of (random) numbers are spread out from their expectation.

$$\mathbb{V}[X] \overset{\textbf{def}}{=} \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

  * For $a, b \in \mathbb{R}$, $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$ holds.

Example: $X \sim N(\mu, \sigma^2) \implies \mathbb{E}[X] = \mu,\ \mathbb{V}[X] = \sigma^2$.

Exercise 2. *For 1-dim r.v.* X*, prove* $\min_{a\in\mathbb{R}} \mathbb{E}[(X-a)^2] = \mathbb{V}[X]$.

independent and identically distributed (i.i.d.) r.v.

For $X_1, X_2, \ldots, X_n$

- $X_1, \ldots, X_n$ are independnet $\iff$ joint pdf is factorized as

$$p(x_1, \ldots, x_n) = p_1(x_1)p_2(x_2) \cdots p_n(x_n)$$

- $X_1, \ldots, X_n$ are <u>independent and identically distributed</u>:

$$p(x_1, \ldots, x_n) = q(x_1)q(x_2) \cdots q(x_n), \quad (q = p_1 = \cdots = p_n)$$

For independent r.v. X, Y, the following equations hold,

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y],$$

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

note:

$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ holds even when X and Y are NOT independent.

- When $X_1, \ldots, X_n$ are i.i.d. from the probability distribution P, we write

$$X_1, \ldots, X_n \ \sim_{i.i.d.} \ P \quad \text{or} \quad (X_1, \ldots, X_n) \sim P^n$$

For example, $X_1, \ldots, X_n \sim_{i.i.d.} N(0,1)$.

In this case, clearly we have

$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n], \qquad \mathbb{V}[X_1] = \cdots = \mathbb{V}[X_n].$$

- When $X_1, \ldots, X_n \ \sim_{i.i.d.} P$, $\mu = E[X_i]$, $\sigma^2 = V[X_i]$,

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Longrightarrow \quad \mathbb{E}[Y] = \mu, \quad \mathbb{V}[Y] = \frac{\sigma^2}{n}$$

Exercise 3. *Suppose* $X_1, \ldots, X_n \sim_{i.i.d.} P$ *and* $\mu = E[X_i]$, $\sigma^2 = V[X_i]$. *For* $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$, *prove the following equations hold:*

$$\mathbb{E}[Y] = \mu, \quad \mathbb{V}[Y] = \frac{\sigma^2}{n}$$

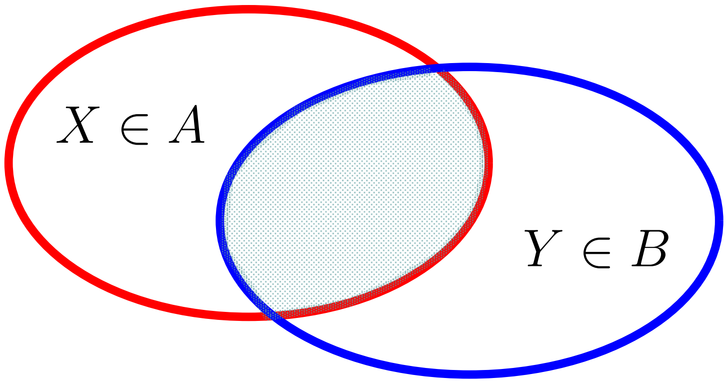
Conditional Probability & Conditional pdf

- conditional probability: the probability of $Y \in B$ under the condition of $X \in A$.

Definition of the conditional probability $\Pr(Y \in B \mid X \in A)$:

$$\Pr(Y \in B \mid X \in A) = \frac{\Pr(X \in A, Y \in B)}{\Pr(X \in A)}$$

$P(B|A) = P(A \cap B)/P(A)$.



- conditional pdf of y for given x,

$$p(y|x) := \frac{p(x,y)}{\displaystyle\int p(x,y)dy} = \frac{p(x,y)}{p_1(x)}. \quad (p_1(x)\text{: marginal pdf of } x)$$

The conditional pdf satisfies $^\forall x, y,\ p(y|x) \geq 0, \quad \displaystyle\int p(y|x)dy = 1$.

probability of $Y \in [y, y+dy]$ under $X \in [x, x+dx]$

$$= \frac{\Pr(X \in [x, x+dx],\ Y \in [y, y+dy])}{\Pr(X \in [x, x+dx])}$$

$$\approx \frac{p(x,y)dxdy}{p_1(x)dx} = p(y|x)dy$$

Bayes' theorem

$$\Pr(X \in A | Y \in B) = \frac{\Pr(Y \in B | X \in A)\Pr(X \in A)}{\Pr(Y \in B)}$$

proof:

$$\begin{aligned}\Pr(X \in A | Y \in B)\Pr(Y \in B) &= \Pr(X \in A,\ Y \in B) \\ &= \Pr(Y \in B | X \in A)\Pr(X \in A).\end{aligned}$$

Interpretation: for the cause X and the result Y,

- $\Pr(Y|X)$: For the cause X, the result Y occurs.
- $\Pr(X|Y)$: infer the cause X based on the result Y.

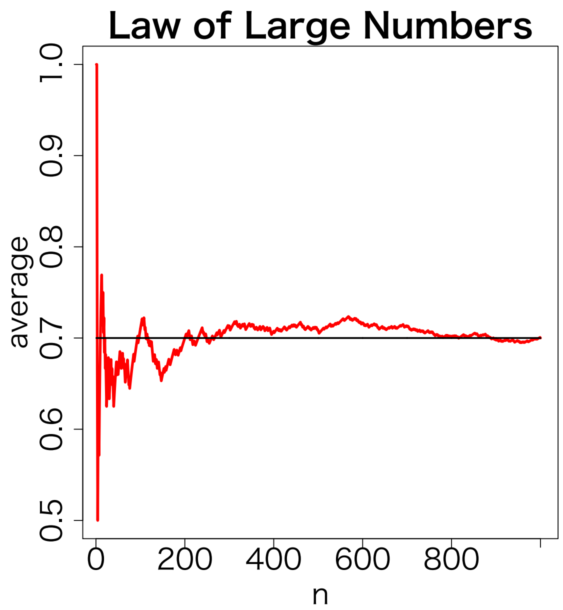
Asymptotic theory: the law of large numbers

For $X_1, \ldots, X_n \sim_{i.i.d.} P$, let $\mathbb{E}(X_i) = \mu \in \mathbb{R}$.

- The Law of Large Numbers:

  for $\bar{X}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \forall \varepsilon > 0, \ \lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \varepsilon) = 0$

  * for sufficiently large n, $\bar{X}_n$ is close to μ with high probability (w.h.p.).



$X_1, \ldots, X_n \sim_{i.i.d.} P$

$P(X_i = 1) = 0.7, \ P(X_i = 0) = 0.3.$

$\Longrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to 0.7 (in probability)

Definition

When

$$\forall \varepsilon > 0, \lim_{n\to\infty} \Pr(|Z_n - a| > \varepsilon) = 0$$

holds for the sequence of r.v. $\{Z_n\}_{n\in\mathbb{N}}$, we say,

"Z_n converges to $a \in \mathbb{R}$ in probability."

and we write $Z_n \xrightarrow{p} a$ for short.

- From the LAN, $\bar{X}_n \xrightarrow{p} \mu$.
- [Slutsky's theorem] For any continuous function $f(z)$,
  $Z_n \xrightarrow{p} a \implies f(Z_n) \xrightarrow{p} f(a)$.