Probability Primer

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 790-784, Korea
seungjin@postech.ac.kr

Sets, Fields, Events

- A set is a collection of objects. The objects are called elements of the set.
- A sample space Ω is the set of all outcomes and subsets of Ω are called events.
- Consider a universal set Ω and a collection of subsects of Ω. Let $E, F, \ldots$ denote subsets in this collection. This collection of subsets of Ω forms a field $\mathcal{M}$ if
  1. $\phi \in \mathcal{M}$, $\Omega \in \mathcal{M}$.
  2. If $E \in \mathcal{M}$ and $F \in \mathcal{M}$, then $E \cup F \in \mathcal{M}$. and $E \cap F \in \mathcal{M}$.
  3. If $E \in \mathcal{M}$, then $E^c \in \mathcal{M}$.
- A σ field $\mathcal{F}$ is a field that is closed under any countable set of unions, intersections, and combinations.

Probability Measure

Given a sample space Ω, a function P defined on the subsets of Ω is a probability measure if the following four axioms are satisfied:

1. $P(A) \geq 0$ for any event $A \in \mathcal{F}$.
2. $P(\emptyset) = 0$.
3. $P(\Omega) = 1$.
4. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ if $A_1, A_2, \ldots$ are events that are mutually exclusive or pairwise disjoint.

The probability measure $P : \mathcal{F} \mapsto [0, 1]$ is a function on $\mathcal{F}$ that assigns to an event $A \in \mathcal{F}$ a number in [0,1], such that above axioms are satisfied.

Joint, Marginal, Conditional, Total Probabilities, Independence

- Joint probability: $P(A, B)$
- Marginal probability: $P(A) = \sum_B P(A, B)$
- Conditional probability: $P(A|B) = \frac{P(AB)}{P(B)}$
- Total probability: $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$ where $\cup_{i=1}^{n} A_i = \Omega$ ($A_i's$ are mutually exclusive)
- Independence: $P(A, B) = P(A)P(B)$

Bayes Theorem

Theorem (Bayes' theorem)

Let A_i, $i = 1, \ldots, n$ be a set of disjoint and exhaustive events. Then $\cup_{i=1}^{n} A_i = \Omega$, $A_i \cap A_j = \phi$, $i \neq j$. For any event B with $P(B) > 0$ and $P(A_i) \neq 0$ $\forall i$,

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

Random Variables and Ensembles

- A random variable $x(\zeta)$ (for a shorthand notation, x) is a correspondence rule between a random outcome ζ of an experiment $\mathcal{H}$ and the real line $\mathbb{R}$.
- An ensemble X is a random variable x with a set of possible outcomes, $\mathcal{A}_x = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$, having probabilities $\{p_1, p_2, \ldots, p_n\}$, with $P(x = \zeta_i) = p_i$, $p_i > 0$ and $\sum_{x \in \mathcal{A}_x} P(x) = 1$.
- A joint ensemble XY is an ensemble in which each outcome is an ordered pair x, y with $x \in \mathcal{A}_x = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$ and $y \in \mathcal{A}_y = \{\nu_1, \nu_2, \ldots, \nu_n\}$.

Marginal and Conditional Probability

- Marginal probability

$$\begin{aligned} P(x = \zeta_i) &= \sum_{y \in \mathcal{A}_y} P(x = \zeta_i, y), \\ P(y) &= \sum_{x \in \mathcal{A}_x} P(x, y). \end{aligned}$$

- Conditional probability

$$P(x = \zeta_i | y = \nu_j) = \frac{P(x = \zeta_i, y = \nu_j)}{P(y = \nu_j)} \quad \text{if} \quad P(y = \nu_j) \neq 0.$$

Product and Sum Rules

$\mathcal{H}$ denotes background assumptions.

- Product rule

$$P(x, y|\mathcal{H}) = P(x|y, \mathcal{H})P(y|\mathcal{H}).$$

- Sum rule

$$\begin{aligned} P(x|\mathcal{H}) &= \sum_y P(x, y|\mathcal{H}) \\ &= \sum_y P(x|y, \mathcal{H})P(y|\mathcal{H}). \end{aligned}$$

Mean, Variance, Moments

- Mean (ensemble average, statistical average, expected value)

$$\mu_x = E\{x\} = \sum_{x \in \mathcal{A}_x} xP(x).$$

- Variance

$$\begin{aligned} \sigma_x^2 &= E\{(x - \mu_x)^2\} \\ &= E\{x^2\} - \mu_x^2. \end{aligned}$$

- Moment

$$m_k = E\{x^k\}.$$

Random Vector and Covariance

A random vector $x \in \mathbb{R}^n$ is a collection of n random variables, $\{x_i\}_{i=1}^n$. The probability density function of the random vector x is defined by the joint density function, i.e.,

$$P(x) = P(x_1, \ldots, x_n)$$

A mean vector and a covariance matrix are defined by

$$\begin{aligned}
\boldsymbol{\mu} &= E\{x\} \\
&= \sum x P(x), \\
\boldsymbol{\Sigma} &= E\left\{(x - \boldsymbol{\mu})(x - \boldsymbol{\mu})^T\right\} \\
&= \sum (x - \boldsymbol{\mu})(x - \boldsymbol{\mu})^T P(x).
\end{aligned}$$

Gaussian Distribution

- Univariate

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

- Multivariate

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right\}.$$

Why Gaussian Distribution?

- Simple analytical properties.
- Completely characterized by mean and covariance.
- Central limit theorem.
- The distribution is again normal after a nonsingular linear transform.
- Marginal density is also normal and conditional density is also normal.
- There exists a linear transform which diagonalizes the covariance matrix (whitening, data sphering).
- Has maximum entropy, given values of the mean and the covariance matrix.