Based on the Appendix of the textbook

A Probabilistic Theory of Pattern Recognition

Luc Devroye   László Györfi   Gábor Lugosi

Edited by András Antos

[We handle only discrete random variables and joint distributions of finitely many of them.]

Appendix

In this appendix we summarize some basic definitions and results from the theory of probability. Most proofs are omitted as they may be found in standard textbooks on probability, such as Feller [1], Ash [2], Shiryayev [3], Chow and Teicher [4], Durrett [5], Grimmett and Stirzaker [6], and Zygmund [7]. We also give a list of useful inequalities that are used in the text.

1 Basics of Measure Theory

Definition 1 *Let S be a set, and let $\mathcal{F}$ be the family of all subsets of S. Then $(S, \mathcal{F})$ is called a measurable space. The subsets of S are called measurable sets.*

Definition 2 *Let $(S, \mathcal{F})$ be a measurable space and let $\nu : \mathcal{F} \to [0, \infty)$ be a function. ν is a measure on $\mathcal{F}$ if*

(i) $\nu(\emptyset) = 0$,

(ii) *ν is σ-additive, that is, $A_1, A_2, \ldots \in \mathcal{F}$, and $A_i \cap A_j = 0$, $i \neq j$ imply that $\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i)$.*

In other words, a measure is a nonnegative, σ-additive set function.

Definition 3 *The triple $(S, \mathcal{F}, \nu)$ is a measure space if $(S, \mathcal{F})$ is a measurable space and ν is a measure on $\mathcal{F}$.*

Definition 4 *Let ν_1 and ν_2 be measures on the measurable spaces $(S_1, \mathcal{F}_1)$ and $(S_2, \mathcal{F}_2)$, respectively. Let $(S, \mathcal{F})$ be a measurable space such that $S = S_1 \times S_2$, and $F_1 \times F_2 \in \mathcal{F}$ whenever $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$. ν is called the product measure of ν_1 and ν_2 on $\mathcal{F}$ if for $F_1 \in \mathcal{F}_1$ and $F_2 \in \mathcal{F}_2$, $\nu(F_1 \times F_2) = \nu_1(F_1)\nu_2(F_2)$. The product of more than two measures can be defined similarly.*

2 Probability

Definition 5 *A (countable) measure space* $(\Omega, \mathcal{F}, \mathbf{P})$ *is called a probability space if* $\mathbf{P}\{\Omega\} = 1$. Ω *is the sample space or sure event, the measurable sets are called events, and the* $\Omega \mapsto \mathcal{R}$ *functions are called (discrete) random variables. If* $X_1, \ldots, X_n$ *are random variables then* $X = (X_1, \ldots, X_n)$ *is a vector-valued random variable.*

Definition 6 *Let* X *be a random variable, then* X *induces the measure* μ *on the subsets of* $\mathcal{R}$ *by*

$$\mu(B) = \mathbf{P}\left\{\{\omega : X(\omega) \in B\}\right\} = \mathbf{P}\{X \in B\}, \qquad B \subseteq \mathcal{R}.$$

The probability measure μ *is called the distribution of the random variable* X.

Definition 7 *Let* X *be a random variable. The expectation of* X *is*

$$\mathbf{E}\{X\} = \sum_x x\mathbf{P}\{X = x\} = \sum_{x>0} x\mathbf{P}\{X = x\} + \sum_{x<0} x\mathbf{P}\{X = x\},$$

if at least one term on the right-hand side is finite.

Definition 8 *Let* X *be a random variable. The variance of* X *is*

$$\mathbf{Var}\{X\} = \mathbf{E}\left\{(X - \mathbf{E}\{X\})^2\right\}$$

if $\mathbf{E}\{X\}$ *is finite, and* ∞ *if* $\mathbf{E}\{X\}$ *is not finite or does not exist.*

Definition 9 *Let* $X_1, \ldots, X_n$ *be random variables. They induce the measure* $\mu^{(n)}$ *on the subsets of* $\mathcal{R}^n$ *with the property*

$$\mu^{(n)}(B) = \mathbf{P}\left\{\{\omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\}\right\}, \qquad B \subseteq \mathcal{R}^n.$$

$\mu^{(n)}$ *is called the joint distribution of the random variables* $X_1, \ldots, X_n$. *Let* μ_i *be the distribution of* X_i $(i = 1, \ldots, n)$. *The random variables* $X_1, \ldots, X_n$ *are independent if their joint distribution* $\mu^{(n)}$ *is the product measure of* $\mu_1, \ldots, \mu_n$. *The events* $A_1, \ldots, A_n \in \mathcal{F}$ *are independent if the random variables* $I_{A_1}, \ldots, I_{A_n}$ *are independent.*

Theorem 1 *If the random variables* $X_1, \ldots, X_n$ *are independent and have finite expectations then*

$$\mathbf{E}\{X_1 X_2 \ldots X_n\} = \mathbf{E}\{X_1\}\mathbf{E}\{X_2\} \cdots \mathbf{E}\{X_n\}.$$

3 Inequalities

Theorem 2 (Cauchy-Schwarz inequality). *If the random variables* X *and* Y *have finite second moments* ($\mathbf{E}\{X^2\} < \infty$ *and* $\mathbf{E}\{Y^2\} < \infty$), *then*

$$|\mathbf{E}\{XY\}| \leq \sqrt{\mathbf{E}\{X^2\}\mathbf{E}\{Y^2\}}.$$

Theorem 3 (Markov's inequality). *Let* X *be a nonnegative-valued random variable. Then for each* $t > 0$,

$$\mathbf{P}\{X \geq t\} \leq \frac{\mathbf{E}\{X\}}{t}.$$

Theorem 4 (Chebyshev's inequality). *Let X be a random variable. Then for each $t > 0$,*

$$\mathbf{P}\{|X - \mathbf{E}\{X\}| \geq t\} \leq \frac{\mathbf{Var}\{X\}}{t^2}.$$

Theorem 5 (Jensen's inequality). *If f is a real-valued convex function on a finite or infinite interval of $\mathcal{R}$, and X is a random variable with finite expectation, taking its values in this interval, then*

$$f(\mathbf{E}\{X\}) \leq \mathbf{E}\{f(X)\}.$$

4 Convergence of Random Variables

Definition 10 *Let $\{X_n\}$, $n = 1, 2, \ldots$, be a sequence of random variables. We say that*

$$\lim_{n\to\infty} X_n = X \qquad \textit{in probability}$$

if for each $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\{|X_n - X| \geq \epsilon\} = 0.$$

We say that

$$\lim_{n\to\infty} X_n = X \qquad \textit{with probability one (or almost surely),}$$

if

$$\mathbf{P}\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = 1.$$

For a fixed number $p \geq 1$ we say that

$$\lim_{n\to\infty} X_n = X \quad \textit{in } L_p,$$

if

$$\lim_{n\to\infty} \mathbf{E}\left\{|X_n - X|^p\right\} = 0.$$

Theorem 6 *Convergence in L_p implies convergence in probability.*

Theorem 7 $\lim_{n\to\infty} X_n = X$ *with probability one if and only if*

$$\lim_{n\to\infty} \sup_{n\leq m} |X_m - X| = 0$$

in probability. Thus, convergence with probability one implies convergence in probability.

5 Conditional Expectation

If Y is a random variable with finite expectation and A is an event with positive probability, then the conditional expectation of Y given A is defined by

$$\mathbf{E}\{Y|A\} = \frac{\mathbf{E}\{YI_A\}}{\mathbf{P}\{A\}}.$$

The conditional probability of an event B given A is

$$\mathbf{P}\{B|A\} = \mathbf{E}\{I_B|A\} = \frac{\mathbf{P}\{A \cap B\}}{\mathbf{P}\{A\}}.$$

Definition 11 *Let Y be a random variable with finite expectation and X be a d-dimensional vector-valued random variable. For $x \in \mathcal{R}^d$ such that $\mathbf{P}\{X = x\} > 0$, let*

$$g(x) = \mathbf{E}\{Y|X = x\} = \frac{\mathbf{E}\{Y I_{\{X=x\}}\}}{\mathbf{P}\{X = x\}}.$$

The conditional expectation $\mathbf{E}\{Y|X\}$ of Y given X is a random variable with the property that $\mathbf{E}\{Y|X\} = g(X)$ with probability one.

Definition 12 *Let C be an event and X be a d-dimensional vector-valued random variable. Then the conditional probability of C given X is $\mathbf{P}\{C|X\} = \mathbf{E}\{I_C|X\}$.*

Theorem 8 *Let Y be a random variable with finite expectation. Let C be an event, and let X and Z be vector-valued random variables. Then*

(i)

(ii) $\mathbf{E}\{Y\} = \mathbf{E}\{\mathbf{E}\{Y|X\}\}$, $\mathbf{P}\{C\} = \mathbf{E}\{\mathbf{P}\{C|X\}\}$.

(iii) $\mathbf{E}\{Y|X\} = \mathbf{E}\{\mathbf{E}\{Y|X, Z\}|X\}$, $\mathbf{P}\{C|X\} = \mathbf{E}\{\mathbf{P}\{C|X, Y\}|X\}$.

(iv) If Y is a function of X then $\mathbf{E}\{Y|X\} = Y$.

(v) If (Y, X) and Z are independent, then $\mathbf{E}\{Y|X, Z\} = \mathbf{E}\{Y|X\}$.

(vi) If $Y = f(X, Z)$ for a function f, and X and Z are independent, then $\mathbf{E}\{Y|X\} = g(X)$, where $g(x) = \mathbf{E}\{f(x, Z)\}$.

6 The Binomial Distribution

An integer-valued random variable X is said to be binomially distributed with parameters n and p if

$$\mathbf{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

If $A_1, \ldots, A_n$ are independent events with $\mathbf{P}\{A_i\} = p$, then $X = \sum_{i=1}^n I_{A_i}$ is binomial (n, p). I_{A_i} is called a *Bernoulli* random variable with parameter p.

7 The Multinomial Distribution

A vector $(N_1, \ldots, N_k)$ of integer-valued random variables is *multinomially distributed* with parameters $(n, p_1, \ldots, p_k)$ if

$$\mathbf{P}\{N_1 = i_1, \ldots, N_k = i_k\} = \begin{cases} \frac{n!}{i_1! \cdots i_k!} p_1^{i_1} \cdots p_k^{i_k} & \text{if } \sum_{j=1}^k i_j = k, \quad i_j \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

8 The Exponential Distribution

A nonnegative random variable has *exponential* distribution with parameter $\lambda > 0$ if it has a density

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

9 The Multivariate Normal Distribution

A d-dimensional random variable $X = \left(X^{(1)}, \ldots, X^{(d)}\right)$ has the *multivariate normal* distribution if it has a density

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)},$$

where $m \in \mathcal{R}^d$, Σ is a positive definite symmetric $d \times d$ matrix with entries σ_{ij}, and $\det(\Sigma)$ denotes the determinant of Σ. Then $\mathbf{E}X = m$, and for all $i, j = 1, \ldots, d$,

$$\mathbf{E}\left\{(X^{(i)} - \mathbf{E}X^{(i)})(X^{(j)} - \mathbf{E}X^{(j)})\right\} = \sigma_{ij}.$$

Σ is called the *covariance matrix* of X.

References

[1] W. Feller. *An Introduction to Probability Theory and its Applications, Vol.1.* John Wiley, New York, 1968.

[2] R.B. Ash. *Real Analysis and Probability.* Academic Press, New York, 1972.

[3] A. N. Shiryayev. *Probability.* Springer-Verlag, New York, 1984.

[4] Y. S. Chow and H. Teicher. *Probability Theory, Independence, Interchangeability, Martingales.* Springer-Verlag, New York, 1978.

[5] R. Durrett. *Probability: Theory and Examples.* Wadsworth and Brooks/Cole, Pacific Grove, CA, 1991.

[6] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes.* Oxford University Press, Oxford, 1992.

[7] A. Zygmund. *Trigonometric Series I.* University Press, Cambridge, 1959.