Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Foundations of Predictive Analytics

James Wu
Stephen Coggeshall

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Foundations of Predictive Analytics

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

SERIES EDITOR

Vipin Kumar

University of Minnesota

Department of Computer Science and Engineering

Minneapolis, Minnesota, U.S.A

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

UNDERSTANDING COMPLEX DATASETS:
DATA MINING WITH MATRIX DECOMPOSITIONS
David Skillicorn

COMPUTATIONAL METHODS OF FEATURE SELECTION
Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: ADVANCES IN ALGORITHMS, THEORY, AND APPLICATIONS
Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT
David Skillicorn

MULTIMEDIA DATA MINING: A SYSTEMATIC INTRODUCTION TO CONCEPTS AND THEORY
Zhongfei Zhang and Ruofei Zhang

NEXT GENERATION OF DATA MINING
Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

DATA MINING FOR DESIGN AND MARKETING
Yukio Ohsawa and Katsutoshi Yada

THE TOP TEN ALGORITHMS IN DATA MINING
Xindong Wu and Vipin Kumar

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, SECOND EDITION
Harvey J. Miller and Jiawei Han

TEXT MINING: CLASSIFICATION, CLUSTERING, AND APPLICATIONS
Ashok N. Srivastava and Mehran Sahami

BIOLOGICAL DATA MINING
Jake Y. Chen and Stefano Lonardi

INFORMATION DISCOVERY ON ELECTRONIC HEALTH RECORDS
Vagelis Hristidis

TEMPORAL DATA MINING
Theophano Mitsa

RELATIONAL DATA CLUSTERING: MODELS, ALGORITHMS, AND APPLICATIONS
Bo Long, Zhongfei Zhang, and Philip S. Yu

KNOWLEDGE DISCOVERY FROM DATA STREAMS
João Gama

STATISTICAL DATA MINING USING SAS APPLICATIONS, SECOND EDITION
George Fernandez

INTRODUCTION TO PRIVACY-PRESERVING DATA PUBLISHING:
CONCEPTS AND TECHNIQUES
Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu

HANDBOOK OF EDUCATIONAL DATA MINING
Cristóbal Romero, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S.J.d. Baker

DATA MINING WITH R: LEARNING WITH CASE STUDIES
Luís Torgo

MINING SOFTWARE SPECIFICATIONS: METHODOLOGIES AND APPLICATIONS
David Lo, Siau-Cheng Khoo, Jiawei Han, and Chao Liu

DATA CLUSTERING IN C++: AN OBJECT-ORIENTED APPROACH
Guojun Gan

MUSIC DATA MINING
Tao Li, Mitsunori Ogihara, and George Tzanetakis

MACHINE LEARNING AND KNOWLEDGE DISCOVERY FOR
ENGINEERING SYSTEMS HEALTH MANAGEMENT
Ashok N. Srivastava and Jiawei Han

SPECTRAL FEATURE SELECTION FOR DATA MINING
Zheng Alan Zhao and Huan Liu

ADVANCES IN MACHINE LEARNING AND DATA MINING FOR ASTRONOMY
Michael J. Way, Jeffrey D. Scargle, Kamal M. Ali, and Ashok N. Srivastava

FOUNDATIONS OF PREDICTIVE ANALYTICS
James Wu and Stephen Coggeshall

This page intentionally left blank

Foundations of Predictive Analytics

James Wu
Stephen Coggeshall

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20120119

International Standard Book Number-13: 978-1-4398-6948-2 (eBook - PDF)

Contents

This page intentionally left blank

List of Figures

This page intentionally left blank

List of Tables

This page intentionally left blank

Preface

This text is a summary of techniques of data analysis and modeling that the authors have encountered and used in our two decades of experience practicing the art of applied data mining across many different fields. The authors have worked in this field together and separately for many large and small companies, including the Los Alamos National Laboratory, Bank One (JPMorgan Chase), Morgan Stanley, and the startups of the Center for Adaptive Systems Applications (CASA), the Los Alamos Computational Group, and ID Analytics. We have applied these techniques to traditional and nontraditional problems in a wide range of areas including consumer behavior modeling (credit, fraud, marketing), consumer products, stock forecasting, fund analysis, asset allocation, and equity and fixed income options pricing.

This book provides the necessary information for understanding the common techniques for exploratory data analysis and modeling. It also explains the details of the algorithms behind these techniques, including underlying assumptions and mathematical formulations. It is the authors' opinion that in order to apply different techniques to different problems appropriately, it is essential to understand the assumptions and theory behind each technique.

It is recognized that this work is far from a complete treatise on the subject. Many excellent additional texts exist on the popular subjects and it was not a goal for this present text to be a complete compilation. Rather, this text contains various discussions on many practical subjects that are frequently missing from other texts, as well as details on some subjects that are not often or easily found. Thus this text makes an excellent supplemental and referential resource for the practitioners of these subjects. This text is self-contained in that it provides a step-by-step derivation explicitly for each topic from the underlying assumptions to the final conclusions.

We hope the readers will enjoy reading and using this text, and will find it to be helpful to better understand the various subjects treated here. You can find the software package (DataMinerXL library for building predictive models) and more information on the topics of modeling at the author's website: www.DataMinerXL.com.

We would appreciate receiving your suggestions and comments at the contact at www.DataMinerXL.com.

This page intentionally left blank

Chapter 1

Introduction

In this book we cover the essentials of the foundations of statistical modeling. We begin with the concepts and properties of statistical distributions, and describe important properties of the various frequently encountered distributions as well as some of the more exotic but useful ones. We have an extensive section on matrix properties that are fundamental to many of the algorithmic techniques we use in modeling. This includes sections on unusual yet useful methods to split complex calculations and to iterate with the inclusion/exclusion of specific data points. We describe the characteristics of the most common modeling frameworks, both linear and nonlinear. Time series or forecasting models have certain idiosyncrasies that are sufficiently special so we devote a chapter to these methods.

There is description of how and why to prepare the data for statistical modeling, including much discussion on practical approaches. We discuss the variety of goodness measures and how and when they are applied. There is a section on optimization methods which are required for many of the training algorithms. We also have description on some of the less well-known topics such as multidimensional scaling, Dempster–Shafer Theory, and some practical discussion about simulation and reject inference.

1.1 What Is a Model?

In its most general definition, a model is a representation of something. It could be a static model that just "sits there" or it could be a dynamic model that represents some kind of process. Examples of static models include a woman walking down a fashion runway wearing some new designer clothes, and she represents how a person would look wearing these clothes. A model of the atom is a conceptual representation of a nucleus with tightly bound protons and neutrons surrounded by a cloud of electrons in various complex orbital shells. A model airplane is typically a smaller and simpler representation of the real object. Examples of dynamic models might be a set of equations that describe how fluid flows through passageways, traffic flowing in cities, stock values rising and falling in time.

Models are usually simpler representations of more complicated systems.

We build models to help us understand what is going on, how things interact, or to predict what may happen as things change or evolve. Models could be physical objects or algorithms (a set of rules and/or equations). For algorithmic models they could be first principles or statistical. For first principle models we examine a system and write down a set of rules/equations that describe the essence of the system, ignoring complicating details that are less important. The first principles models could be differential equations that describe a dynamical system or they could be a set of evolutional rules that are written down from expert knowledge of the system's processes.

We are guided in modeling by observing the real system to be modeled. We do our best to understand the important elements of the system, what features characterize the system, what changes, what are the possible dependencies, and what are the important characteristics we are interested in. This observation guides us in constructing the model, either through the first principles or in building a statistical model. Sometimes the process is so complex or unknown that we cannot begin to write down first principles equations, but our observations are sufficient that we have many data examples of the system to be modeled. In these cases we can build a statistical model. The remainder of this book will deal with issues around such statistical modeling.

1.2 What Is a Statistical Model?

A statistical model is typically a set of equations with adjustable parameters that we "train" using the many data observations. Invariably we have an overdetermined system, where we have many more data examples than adjustable parameters, and we desire the "best" set of parameters that allow our equations to fit the data as best as possible. The structure of the equations can be very simple or very complex. It could be a simple linear combination of parameters times system characteristics, or it could be a highly complex nonlinear combination of characteristics and parameters. Our modeling task has several components: (1) a set of equations/rules with adjustable parameters, (2) a set of data that entails examples of the system that we are trying to model, (3) a concept of goodness of fit to the data, and (4) a set of rules that tells us how to adjust the parameters to increase the goodness of the model fit.

The set of parameter-adjusting rules are usually obtained from the goodness of fit measure, often differentiating an objective function with respect to the fitting parameters and setting it to zero, thus finding the set of parameters that are at least locally optimal. One can consider the objective function as a model error surface that floats above a hyperplane of the parameter space, and our desire is to find the minimum of this error surface in this hyperplane. If the error surface is simple we may have an easy-to-find location with the global

minimum over the set of all possible parameter values. This is the case with linear regression using a mean square error objective, where the error surface is a paraboloid over the hyperplane of the set of parameters. Often, however, this fitting surface is complex and has multiple local valleys in which we can be trapped as we traverse the hyperplane of possible parameters in search of the minimum. Many techniques exist to help us avoid getting trapped in these local minima regions, such as gradient search with momentum, simulated annealing where the step size starts off large and decreases, or genetic algorithms. The more complex the model, the more complex the fitting surface, and the more likely we need to be careful about local versus global minima.

1.3 The Modeling Process

The process of building a model is a blend of both science and art, as are most endeavors. There is a straightforward formulaic process that we follow, and at each step we are guided by intuition and experience. Here are the basic steps in building a statistical model:

1. **Define the goals.** What are we trying to achieve? What is the outcome we are trying to predict? Under what situations will the model be used and for what purposes? Understand what will determine a good model.

2. **Gather data.** What data is available, in what form, with what quality? How many records (we use the term record interchangeably with data point) will we have? What fields (or model inputs) might be available? How far back in time does the data go, and how relevant are the older records? Generally, modelers want as much data as possible.

3. **Decide the model structure.** Should we do a linear regression, logistic regression, or a nonlinear model, and if so, which kind? What time frames and/or populations should be used for model development and validation? What are the inputs, the output(s), and the measure of goodness? A good approach might be to try a simple linear model as a baseline and then try various nonlinear models to see if one can do better. Choices of model structure require experience and deep knowledge of the strengths and weaknesses of each technique. Important characteristics to consider in the choice of technique include continuous or categorical (classification) output, number of records, likely dimensionality, and amount of data noise.

4. **Prepare the data.** Assemble the data records into the appropriate form for the model. Encode the data into inputs, using expert knowledge as much as possible. Appropriately normalize the numerical data and encode

the categorical data fields. Examine distributions for appropriateness and take care of outliers. Separate data into the desired training, testing, and validation sets.

5. **Variable selection and elimination**. In this step, which may or may not be associated with candidate models, variables are examined for the importance to the models and selected or eliminated. Some variable selection methods are stand alone (filters) and others (wrappers) are integrated with a particular model methodology. Generally in this step a list of candidate good variables rank ordered by importance is decided.

6. **Build candidate models.** Begin building models and assessing the model goodness. We may begin with baseline linear models and then try to improve using more complex nonlinear models. In this and all phases it is important to keep in mind the environment in which the model will be implemented.

7. **Finalize the model.** Select among the candidates the most appropriate model to be implemented. Document the model as needed.

8. **Implementation and monitoring.** Here the model is embedded into the necessary system process, and monitoring steps are built to examine the model performance in ongoing use.

1.4 Modeling Pitfalls

There are many possible traps and difficulties in the model building process. Here we list a few in each of the above-mentioned modeling steps:

1. **Define the goals—pitfalls.** Lack of clarity around problem definition. Problems being defined too narrowly. Sometimes a larger, more important problem can be solved, broader than the one specifically being targeted. Lack of understanding how and where the model will be used. Sometimes key data will not be available for the model when implemented.

2. **Gather data—pitfalls.** Using data too old or otherwise not relevant going forward. Not excluding records of types that will not be seen by the model when implemented. Not considering additional key data sources (features/inputs) or data sets (records) that might be available.

3. **Decide the model structure—pitfalls.** Using too simple a model (rarely a problem). Using too complex a model (more often a problem). Using

a modeling methodology that is not appropriate for the nature of the data (sizes, dimensions, noise...).

4. **Prepare the data—pitfalls.** Not cleaning or considering outliers (some techniques are robust to outliers; many are not). Not properly scaling data. Inefficient encoding of categorical variables. Not eliminating fields that will not be available or will be different going forward. Not giving enough thought to building special expert variables. Not recognizing inherent data bias, including not having data from important categories of records.

5. **Variable selection and elimination—pitfalls**. Relying on only linear variable selection or compression techniques (e.g., PCA). Too much reliance on simply eliminating correlated variables. Not examining all model inputs (distributions, outliers). Keeping too many variables, making it hard for modeling, interpretation, implementation, or model maintenance. Target leakage, where the output value to be predicted is inadvertently and inappropriately hiding in one of the input variables. (Important note—Look carefully and skeptically at any variables that seem to be TOO good in the model.)

6. **Build candidate models—pitfalls.** Going deep on a single specialized technique instead of at first trying a broad spectrum of methods. Not doing a simple linear model as a baseline. Not doing proper training/testing as one examines candidate models. Overfitting.

7. **Finalize the model—pitfalls.** Not rebuilding the final model optimally using all the appropriate data. Improperly selecting the final model without consideration to some implementation constraints. Lack of proper model documentation.

8. **Implementation and monitoring—pitfalls.** Errors in the implementation process: data input streams, variable encodings, algorithm mistakes, simple bugs around special cases. Not monitoring model ongoing performance (inputs, outputs, predictive power).

1.5 Characteristics of Good Modelers

Good modelers can come from a wide range of technical backgrounds. They typically have an undergraduate and a graduate degree in any one of a handful of scientific disciplines, including computer science, mathematics, statistics, engineering, and physics. Physicists make some of the best modelers because of the rigorous applied mathematical exposure, and the trained ability to eliminate all but the essential nature of a problem. Applied mathematicians

have built up a repertoire of the tools needed for modeling and are not afraid of jumping into any new algorithmic technology. Engineers, like physicists, have a keen sense of what is essential to the system being modeled. Statisticians are awash in the foundations of the tools underlying most of the modeling techniques. Anyone with any of these backgrounds has the essential baseline skills to become a good modeler.

What are the characteristics that are important to be a good modeler?

1. **Technical competence**. Mastery of the advanced math and statistics fundamentals, such as those found in this book. Very good programming skills since everything is done in software, and one needs the ability to quickly and efficiently code one's new ideas. These are the basic tools of the trade. A bonus is experience with very large, messy data sets. Given these as backgrounds, good modelers will have complete in-depth understanding of every modeling technique they are using rather than just running some commercial software. This deep understanding is the only way one can make good choices between possible techniques and allow the possible modifications of existing algorithms.

2. **Curiosity**. The best modelers are those who are innately curious. They drill into something that just does not look or feel right. If there is something that they do not completely understand they dig deeper to find out what is going on. Frequently the best results come from inadvertent discoveries while chasing a vague feeling around a mystery.

3. **Common sense and perspective**. One needs the ability to drill deep without losing sight of the big picture. A good modeler is constantly making appropriate judgments about what is more and less important, focusing the pursuit in the directions that provide the most overall value without plunging down a rabbit hole.

4. **Passion for data**. The best modelers have an unbridled passion for everything about data—lots and lots of data. One can never get enough data. Never say no to data. One needs to understand everything about the data: where it came from, where it will come from, the quality, robustness, and stability of the data. A lot of effort needs to go into data cleansing, standardizing and organizing. A good data steward will build libraries of data for present and future uses. Data is the lifeblood of statistical model builders. It is the foundation of all our models and everything we do. It is better to use basic techniques on good data sets than to use complex, sophisticated techniques on poor quality data.

5. **Tenacity**. There are many processes involved in good model building and it is important to do each step well without cutting corners. It is tempting to jump around in the model searching process while sometimes a more rigorous, organized search method is appropriate. Tenacity is also an important part in the curiosity characteristic in the chasing down of loose ends.

6. **Creativity**. Always be questioning your modeling technique. Is it the most appropriate for your particular problem and data set? Look for possible modifications of the algorithms that might help, for example, model training rules, data weighting, fuzzifying rather than using sharp boundaries. Don't be afraid to modify algorithms based on need or hunches. Watch for the possibility of broadening of the problem objective to some higher-order goal that may be of more use.

7. **Communication skills**. A useful characteristic for modelers is the ability to explain the essence of the model to nontechnical people. This would include aspects such as what inputs are used, what is the output modeled, what are the strengths, weaknesses, and performance of the model, the nature of the algorithmic structure, the training and testing process, and opinions about robustness and stability based on various statistical tests. A desired skill is to be able to explain complex processes in clear, concise nontechnical language. This requires complete understanding of all the underlying aspects and details of the model, data, and process.

8. **Ability to work in teams**. Modelers are frequently required to work in teams, where a complex modeling task is divided up into subtasks. Ability to work well together with good cooperation, communication, and helping each other is very important.

1.6 The Future of Predictive Analytics

Statistical modeling is a booming field. There is an explosion of data everywhere, being collected more and more around ever-increasing and ubiquitous monitoring of many processes. Data continues to grow and hardware is successfully struggling to keep up, both in processing power and large volume data repositories. New data repository structures have evolved, for example the MapReduce/Hadoop paradigm. Cloud data storage and computing is growing, particularly in problems that can be separable, using distributed processes. Data sizes of terabytes are now the norm. Predictive Analytics is an exciting and growing field!

Data can be broadly categorized as either structured or unstructured. Structured data has well-defined and delimited fields, and this is the usual type of data used in statistical modeling. More often now we have the need to incorporate unstructured data, such as free text, speech, sound, pictures and video. These formats require sophisticated encoding techniques, and many currently exists and more continue to emerge. Generally we reduce these unstructured data sets to structured fields through these specialized data encoding methodologies that are particular to the characteristics of the unstructured

data and the particular problem to be modeled. It is likely that we will need to deal more and more with these unstructured data in the future.

Statistical modeling can also be divided into two broad categories: search and prediction. For search problems we try to identify categories of data and then match search requests with the appropriate data records. In prediction problems we try to estimate a functional relationship, so we can provide an output to a set of inputs. These prediction statistical models are in general the types of modeling problems that are considered in this text.

Chapter 2

Properties of Statistical Distributions

This chapter presents common distributions widely used in data analysis and modeling. Fully understanding these distributions is key to understanding the underlying assumptions about the distributions of data. The next chapter will discuss various aspects of matrix theory, which is a convenient vehicle to formulate many of these types of problems.

The basic distributions found in data analysis begin with the simplest, the uniform distribution, then the most common, the normal distribution, then a wide variety of special distributions that are commonly found in many aspects of data analysis from consumer modeling and finance to operations research. Many of these distributions are fundamental to standard statistical tests of data relevance (chi-squared, Student's t...). For example, the non-central chi-squared distribution is found to be important in the solution of interest rate models in finance.

The central limit theorem is used explicitly or implicitly throughout statistical analysis and is core to examining statistical properties of compound phenomenon.

First a word about naming conventions. To make a clear differentiation between density functions, distribution functions, and cumulative density/distribution functions we will use the convention of a probability distribution function (pdf) as the density function, and when we integrate over some domain we have the cumulative distribution function (cdf).

2.1 Fundamental Distributions

2.1.1 Uniform Distribution

The simplest distribution is the uniform distribution over the interval from zero to one, $U(0, 1)$. Its pdf is

$$u(x) = \begin{cases} 1 & \text{if } x \in (0,\, 1) \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Most modern computer languages provide random number generators for the uniform distribution. It can be used to generate random variables with other

distributions and we will discuss this in a later section of this chapter, when we discuss random number generators.

2.1.2 Details of the Normal (Gaussian) Distribution

The most common distribution is the normal distribution, also known as the Gaussian distribution or a bell-shaped curve. Because of its central importance in all statistical modeling we describe here a fairly extensive set of fundamental properties and characteristics. The normal distribution $N[\mu, \sigma^2]$, with mean μ and variance σ^2, is

$$\varphi\left(x, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.2}$$

The standard normal distribution, $N[0, 1]$, has zero mean and unit standard deviation,

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \tag{2.3}$$

Its cumulative distribution function (cdf) is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{2.4}$$

It can be expressed in terms of the complementary error function

$$\Phi(x) = 1 - \tfrac{1}{2}\text{erfc}\left(x/\sqrt{2}\right), \tag{2.5}$$

or in terms of the error function

$$\Phi(x) = \tfrac{1}{2} + \tfrac{1}{2}\text{erf}\left(x/\sqrt{2}\right), \tag{2.6}$$

where

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-t^2} dt \quad \text{and} \quad \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt \tag{2.7}$$

are the complementary error function and the error function, respectively, and they satisfy $\text{erf}(x) + \text{erfc}(x) = 1$.

Asymptotic Expansion of the Normal Distribution

The asymptotic expansion of $\Phi(x)$ is

$$\Phi(x) \sim -\varphi(x)\left[\frac{1}{x} - \frac{1}{x^3} + \frac{1 \cdot 3}{x^5} + ... + (-1)^k \frac{(2k-1)!!}{x^{2k+1}}\right] \qquad x < 0 \tag{2.8}$$

and

$$1 - \Phi(x) \sim \varphi(x)\left[\frac{1}{x} - \frac{1}{x^3} + \frac{1 \cdot 3}{x^5} + ... + (-1)^k \frac{(2k-1)!!}{x^{2k+1}}\right] \qquad x > 0. \tag{2.9}$$

The right-hand side overestimates if k is even and underestimates if k is odd. The leading order approximation is

$$\begin{aligned} \Phi(x) &\sim -\tfrac{1}{x}\varphi(x) \qquad x \to -\infty \\ \Phi(x) &\sim 1 - \tfrac{1}{x}\varphi(x) \quad x \to \infty. \end{aligned} \tag{2.10}$$

We can directly work on the integration to obtain the asymptotic expansion. For $x < 0$ we have

$$\begin{aligned} \Phi(x) &= \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt \\ &= \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x+t)^2/2} dt \\ &= \tfrac{e^{-x^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(xt+t^2/2)} dt \\ &= \tfrac{e^{-x^2/2}}{\sqrt{2\pi}} \tfrac{1}{-x} \int_{0}^{\infty} e^{-\left(t+\frac{t^2}{2x^2}\right)} dt \\ &= -\tfrac{1}{x}\varphi(x)\, I(x), \end{aligned} \tag{2.11}$$

where

$$I(x) = \int_0^\infty e^{-\left(t+\frac{t^2}{2x^2}\right)} dt. \tag{2.12}$$

Clearly it is easy to see that $I(-\infty) = \int_0^\infty e^{-t} dt = 1$. The higher-order terms are

$$\begin{aligned} I(x) &= \sum_{k=0}^{\infty} \int_0^\infty e^{-t} \tfrac{1}{k!} \left(\tfrac{-t^2}{2x^2}\right)^k dt \\ &= \sum_{k=0}^{\infty} \tfrac{(-1)^k}{k!} \left(\tfrac{1}{2x^2}\right)^k \int_0^\infty e^{-t} t^{2k} dt \\ &= \sum_{k=0}^{\infty} \tfrac{(-1)^k}{k!} \left(\tfrac{1}{2x^2}\right)^k (2k)! \\ &= \sum_{k=0}^{\infty} (-1)^k \tfrac{(2k-1)!!}{x^{2k}}. \end{aligned} \tag{2.13}$$

Therefore we have

$$\Phi(x) = -\frac{1}{x}\varphi(x) \sum_{k=0}^{\infty} (-1)^k \frac{(2k-1)!!}{x^{2k}}, \qquad x < 0. \tag{2.14}$$

The Multivariate Normal Distribution

Some integrals related to the Gaussian distribution are useful. The first one is the normalization

$$\int_{-\infty}^{\infty} dx\, e^{-x^2} = \sqrt{\pi}. \tag{2.15}$$

Let a be a positive real number. Through transformation of the x variable we have

$$\int_{-\infty}^{\infty} dx\, e^{-ax^2} = (\pi/a)^{1/2} \tag{2.16}$$

and

$$\int_{-\infty}^{\infty} dx\, e^{-(ax^2+bx)} = (\pi/a)^{1/2}\, e^{b^2/(4a)}. \tag{2.17}$$

Also handy is

$$\int_{-\infty}^{\infty} dx\, x e^{-(ax^2+bx)} = -\frac{b}{2a}\,(\pi/a)^{1/2}\, e^{b^2/(4a)}. \tag{2.18}$$

Another useful related integral that is not easy to find is

$$I_n(a) = \int_0^{\infty} dx\, x^n e^{-ax^2}, \quad a > 0. \tag{2.19}$$

This branches depending on whether or not n is even or odd:

$$I_{2n}(a) = \frac{(2n-1)!\sqrt{\pi}}{2^{2n}(n-1)!}a^{-(n+1/2)} \quad \text{and} \quad I_{2n+1}(a) = \frac{1}{2}n!\, a^{-(n+1)}. \tag{2.20}$$

The generalization to p-dimensional integration is straightforward. If we let A be an $n \times n$ positive definite and symmetric square matrix and B a p-dimensional vector, we have

$$\int_{-\infty}^{\infty} d^p x\, e^{-\frac{1}{2}x^T A\, x + B^T\, x} = \left(\frac{(2\pi)^p}{\det A}\right)^{1/2} e^{\frac{1}{2}B^T A^{-1} B} \tag{2.21}$$

and

$$\int_{-\infty}^{\infty} d^p x\, x_i x_j\, e^{-\frac{1}{2}x^T A\, x} = \left(\frac{(2\pi)^p}{\det A}\right)^{1/2} (A^{-1})_{ij}. \tag{2.22}$$

Distributions can easily be extended into higher dimensions. In a p-dimensional space, the multi-dimensional normal distribution function, $N[\mu, \Sigma]$, is

$$f(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \tag{2.23}$$

where the mean μ now is a vector and the standard deviation now is the covariance matrix Σ. If $x \sim N[\mu, \Sigma]$ (read this as "x is a random variable distributed as $N[\mu, \Sigma]$"), then any linear combinations of x, $Ax + b$, are normally distributed. Note that

$$E[Ax] = A\, E[x] = A\mu \tag{2.24}$$

and

$$\text{var}[Ax] = A\, \text{var}[x] A^T = A\Sigma A^T. \tag{2.25}$$

Here $E[x]$ is the expectation of x and $\mathrm{var}[x]$ is the variance of x. We have

$$E[Ax + b] = A\mu + b \tag{2.26}$$

and

$$\mathrm{var}[Ax + b] = A\Sigma A^T. \tag{2.27}$$

We therefore have

$$Ax + b \sim N[A\mu + b,\, A\Sigma A^T]. \tag{2.28}$$

Here we give a proof that the linear combination of the normally distributed variables is normally distributed. Since the characteristic function (discussed in Section 2.10) uniquely determines a distribution function, we can use it to prove that variables have a normal distribution. Letting $y = Ax + b$, its characteristic function is

$$\begin{aligned} E\left[e^{ik^T y}\right] &= E\left[e^{ik^T(Ax+b)}\right] = E\left[e^{i(A^T k)^T x + ik^T b}\right] \\ &= e^{i(A^T k)^T \mu - \frac{1}{2}(A^T k)^T \Sigma (A^T k) + ik^T b} \\ &= e^{ik^T(A\mu+b) - \frac{1}{2}k^T(A\Sigma A^T)k}. \end{aligned} \tag{2.29}$$

Therefore y is normally distributed:

$$y \sim N[A\mu + b, A\Sigma A^T]. \tag{2.30}$$

As a special case, if $x \sim N[0,\, I]$ and A is a square matrix, and $AA^T = 1$, then $Ax \sim N[0,\, I]$.

Given a collection of $x_i \sim N[\mu_i,\, \sigma_i^2]$, $i = 1,\, 2,\, ...,\, n$ and they are independent of each other, then $(x_1, x_2, ..., x_n) \sim N[\mu, \Sigma]$, where $\mu = (\mu_1, \mu_2, ..., \mu_n)^T$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$. Let $y = \sum_{i=1}^{n} x_i$, i.e., $A = (1, 1, ..., 1)$ and $b = 0$, then we have

$$y = \sum_{i=1}^{n} x_i \sim N\left[\sum_{i=1}^{n} \mu_i,\, \sum_{i=1}^{n} \sigma_i^2\right]. \tag{2.31}$$

If we partition the x-space into separate regions we can identify several properties. Let the partitions of x and the covariance matrix be

$$x - \mu = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{2.32}$$

with dimensions p_1 for x_1 and p_2 for x_2. By direct integration we can prove that the marginal distribution of a multiple normal distribution is normal, namely,

$$\int d^{p_2} x_2 f(x_1,\, x_2) = \frac{1}{(2\pi)^{p_1/2} (\det \Sigma_{11})^{1/2}} e^{-\frac{1}{2}(x_1 - \mu_1)^T \Sigma_{11}^{-1} (x_1 - \mu_1)}. \tag{2.33}$$