1

INTRODUCTION AND OVERVIEW

All men by nature desire knowledge.
Aristotle

Man has an intense desire for assured knowledge.
Albert Einstein

This chapter describes the motivation for predictive learning from data and the connection between predictive learning, philosophy of science and various ways of handling uncertainty.

Section 1.1 informally describes multiple facets of 'learning from data', including statistical, philosophical and psychological aspects of learning. It also explores the connection between statistical learning and induction in philosophy.

Section 1.2 provides a historical account of handling uncertainty and risk. It is interesting to note that modern probabilistic treatment of uncertainty is very recent, even though humans had to deal with uncertainty throughout thousands of years.

Predictive Learning
By Vladimir Cherkassky

Section 1.3 describes different types of human knowledge, and the growing importance of empirical knowledge in today's data-rich society.

Section 1.4 shows how the problem of learning dependencies from data can be reduced to the problem of function estimation from noisy samples. Such a mathematical formalization allows one to quantify the notions of explanation and prediction. We also point out that the problem of learning from data is just one step in the general experimental procedure used in different fields of science and engineering. Various steps of this procedure are described, with emphasis on the importance of other steps preceding learning.

Section 1.5 presents summary and discussion. This book is mainly concerned with estimation of *predictive* data-analytic models. This framework is called Predictive Learning. The task of predictive learning is essential to many diverse fields such as pattern recognition, statistics, data mining, machine learning, signal processing etc. These data-analytic methodologies are also briefly discussed in Section 1.5.

1.1 OVERVIEW

We live in a data-rich world. With the advent of computer technology, most information now is digital, and its amount doubles every few years. This information, however, is useful to humans only if there are associations and relationships present within the data. For example, it is easy to memorize a song or a poem, because it rhymes and has meaning, but it is very difficult to memorize 500 random unrelated words. The process of inferring stable relationships from data (or observed events) is called 'learning'. Learning, or making sense of observed data, is central to human intelligence. Likewise, learning is necessary for adaptation of all living organisms to a changing unknown environment. Humans and animals are natural experts on learning. Human babies can effortlessly acquire a language without being explicitly taught any grammar or linguistic rules. This is a remarkable example of learning that cannot be matched by any computer.

Much of human knowledge is based on observations of repeatable events. For example, we 'know' that the Sun rises in the East every morning. People knew this fact thousands of years ago, before the advent of astronomy and physics. In essence, such knowledge is a result of generalization from many observed instances (of the Sun rising in the East every morning). This process of making a general statement from several observations/facts is called 'induction' or 'inductive inference'.

It is also an example of 'empirical knowledge' that is purely data-driven. In contrast, 'scientific knowledge' provides much deeper insights into observed empirical data. For example, Kepler's laws can be used to predict planets' movement, and Newton's law of gravity can be used to derive Kepler's laws and also to describe the trajectory of a thrown stone. Scientific laws *explain and predict* many seemingly unrelated events, such as the motion of planetary bodies, and the motion of a falling object, in the case of Newton's law.

For humans, it is not sufficient just to detect regularities from observations of repeatable events; these regularities need to be 'explained' in terms of a small number of basic concepts and causal relationships. These explanations constitute modern science. However, present scientific understanding of the world is fairly recent (just a few hundred years old), and prior to that people used other 'non-scientific' explanations. For example, in the Middle Ages people believed that the Sun moves in the sky because it is pushed by little angels flapping their wings. Such explanations are now considered 'beliefs'. Note that scientific theories and beliefs are inductive as they both explain observed repeatable events. So when a new theory is proposed, it is difficult to classify it as true scientific theory, or just a belief. The Philosophy of Science is concerned with general conditions (principles) for distinguishing between scientific theories, and non-scientific explanations or beliefs. Such a criterion is known as the 'demarcation principle' in philosophy. The demarcation principle is not just an obscure philosophical notion; it is actually very relevant in everyday life, where beliefs and opinions supported by observed correlations in the data are often presented as scientific findings. This is done routinely by politicians, lawyers and advertisers in mass media. Consumers of information are bombarded by such data-driven opinions disguised as 'scientific explanations', and it is up to each individual to evaluate these 'explanations' and possibly act upon them.

This book attempts to cover various aspects of inductive learning from data, and demonstrates the connection between the statistical view of learning, common-sense psychological induction and the Philosophy of Science. For example, the motion of planets has been described throughout human history using

(a) non-scientific beliefs (angels pushing planets),
(b) statistical analysis of empirical data (leading to Kepler's laws) and
(c) scientific theory (Newton's law of gravity).

Another example is the following statement: *Many old men are bald*. This statement is a result of *psychological induction*, i.e. common-sense generalization based on observed past data. Note that it also has a predictive aspect, as it implies that from past observations of many bald men one may infer that future old men are likely to be bald. The classical *statistical approach* to describing this knowledge in quantitative terms may go as follows:

(1) The amount of hair on a man's head is considered a random variable *HAIR*.
(2) Past (known) observations of a large number of men (say, 500 men of different age) are called training samples, or training data. Each data sample has two values, the amount of hair and person's age.
(3) From this training data one can estimate a histogram of the random variable *HAIR*, as a function of *AGE*.

This estimated distribution can be used to *predict* the likelihood that a future man (of a particular age) will be bald. Note that this statistical model has two aspects:

- Descriptive, i.e. it describes (or explains) previously observed (training) data.
- Predictive, as it allows making predictions for new data samples.

The predictive (aka inferential) aspect is usually trickier, and this book describes various methodologies for estimating predictive models from past data. Finally, one can propose a *scientific* theory that relates the lack of hair in older men to a particular gene, or to some hormonal changes in older males. This will be an example of a first-principle scientific explanation for the same data. In order to qualify for a true scientific explanation, it has to explain an existing phenomenon and also explain some other phenomena (i.e., explain why older women or older Asian men do not lose much hair).

As evident from these examples, there are three different ways of explaining empirical data: non-scientific explanations (beliefs), statistical models and first-principle models. Modern science and engineering are based on using *first-principle* models to describe physical, biological, and social systems. Such an approach starts with a basic scientific model (e.g., Newton's laws of mechanics or Maxwell's theory of electromagnetism) and then builds upon it various applications in mechanical engineering or electrical engineering. Here the experimental data (measurements) are used to verify the underlying first-principle models and to estimate some of the model parameters that are difficult to measure directly. However, in many applications the underlying first principles are unknown or the systems under study are too complex to be mathematically described. With the growing use of computers and low-cost sensors for data collection, there is a great amount of data being generated by such systems. In the absence of first-principle models, such readily available data can be used to derive models by estimating useful relationships between system's inputs and outputs. Practical utility of these data-analytic models is usually related to their prediction capability. Currently, there is a paradigm shift from the classical modeling based on first principles to developing empirical data-driven models.

However, the proliferation of data-driven modeling also poses a number of new challenging questions:

- how to measure the quality of explanation and prediction?
- under what conditions good prediction is possible?
- if two models explain past data equally well, which one is 'better' (more plausible)?
- how to distinguish between true scientific theories (models) and pseudo-scientific theories (or beliefs)?

These questions have been posed and discussed by philosophers over many centuries. However, only recently with the advent of computer technology, have these issues become increasingly relevant for engineers, biologists and scientists estimating predictive models from data.

1.2 HISTORICAL PERSPECTIVE

In ancient human societies people assumed a rather passive role towards risk-taking, because their social status was pre-determined from birth. So the prevailing attitude was that the future is determined by gods and (sometimes) revealed by oracles and priests. Modern capitalist society is based on the notions of personal freedom and risk-taking, when individuals (or institutions) try to improve their status by making rational decisions in anticipation of unknown future events. This management of uncertainty and risk-taking is at the core of all modern economic institutions, such as banking, insurance industry and the stock market. It is also critical for understanding and coping with complex engineering and social systems. These systems range from popular electronic devices (cell phones and laptops) to large systems (e.g., electric power grid) – all of which occasionally fail, sometimes with catastrophic consequences.

The quantitative approach to managing uncertainty dates back to the Renaissance when Italian and French mathematicians (Cardano, Pascal and Fermat) applied their mathematical skills to gambling. Gambling, however, had been popular pastime in all ancient human civilizations, and many ancient societies had sophisticated mathematical culture. So why hadn't probability theory and statistics been developed in ancient societies? There are several possible explanations, and they relate to both the lack of experimental science and cultural attitudes towards uncertainty in these societies. In the case of Ancient Greece, there was a clear separation between pure mathematics (prized for its logical derivations) and practical applications. In fact, Greeks placed little value on practical applications of math, and the concept of experimental science was unknown to them. In contrast, they had a high appreciation for 'absolute truth' logically derived from axioms, as in Euclidean geometry. Hence, they felt that uncertain events are of somewhat lower intellectual value. Socrates defines probability or plausibility as 'likeness to truth'. According to Greeks, real truth can be only established by pure logic and reasoning, whereas any study of probability requires observation of real-world events and active participation of a scientist collecting and measuring the data.

In the age of the Renaissance, people started to explore the world through active investigation and experimentation. For the first time in history, individuals tried to take control of their personal destiny. This was an age of major advances in science, art, engineering, geography and business. During this time, two-thirds of the world was discovered (by Europeans), firearms and the printing process were invented. It also marked the beginning of modern experimental science, when Galileo challenged Aristotle's theory of motion, leading to clear separation between theology and science. It was common, at that time, for a scientist to be also a brilliant writer, an engineer and an artist. A talented 16-th century Italian physician and mathematician Girolamo Cardano was the quintessential Renaissance man. He was also a gambling addict. Not surprisingly, he tried to develop a mathematical theory to describe the frequency of outcomes (of observable past events). The idea that uncertainty (probability) can be *measured* empirically, by observing frequencies of past events, was a revolutionary breakthrough. Prior to Cardano, philosophers and mathematicians viewed probability as the degree of belief or the strength of personal opinion. Cardano's frequentist view of probability is more objective, as it allows experimental verification of statistical models estimated from past observations. We will further explore subtle relationships between human understanding of uncertainty, prevailing cultural attitudes and the philosophy of science, in Chapter 3. At this point, we note that human understanding of uncertainty and risk management is constantly evolving, rather than postulated by mathematical theories, such as modern probability theory and statistics. So this book advocates a critical approach to understanding statistical, machine learning and data mining techniques, currently used for various applications.

Modern mathematical treatment of uncertainty was finally fully developed only in the 20th century by Andrei Kolmogorov, who proposed axiomatic definition of probability theory, and Ronald Fisher, the father of classical statistics. These developments occurred long after monumental advances in natural sciences (physics, chemistry, biology) and their practical applications that have totally transformed the human society. These advances are based on the first-principles scientific knowledge, such as Newton's Laws or Maxwell's equations that completely define the state of a physical system. Due to success of such deterministic first-principles laws, the classical science adopted a philosophical view called *causal determinism*. For example, the future state of a mechanical system (consisting of moving objects) is fully predictable if the current coordinates and velocities of each object are known. So assuming the current state of a system can be observed or measured, we can always predict its future state, and there is no room for uncertainty. Causal determinism expresses a philosophical view that *every effect has a cause*, so science can explain, in principle, all natural phenomena. Accordingly, uncertainty simply reflects our ignorance (about Nature) and inability to perform accurate measurements. However, uncertainty is not an *intrinsic property* of natural or social systems. This view remains deeply rooted in modern science and society as a whole. Many famous physicists criticized quantum mechanics (based on probabilistic models) purely on philosophical grounds. For example, Albert Einstein said: 'I am convinced that He (God) does not play dice'. In modern financial markets, public companies report quarterly earnings, and if such earnings are missed by 1-2 pennies relative to 'predictions' (i.e., analysts' estimate), the company's stock price may lose 10-15 % in one day. Such an expectation (of predictable earnings) is unrealistic, because a company's earnings are affected by many uncertain factors, such as industry trends, the state of economy, political developments, and natural disasters.

In classical statistics, the principle of causal determinism has led to the goal of estimating a true statistical model of unknown system. That is, the unknown system is described via a probabilistic model or statistical distribution. Then, given past observations of data samples, the goal of statistics is to estimate this unknown distribution. Assuming this distribution can be accurately estimated from past data, it can be used for both explaining the unknown system and making future predictions. This classical probabilistic approach can be used for risk management in gambling. Here risk management can be achieved by (1) estimating first the probabilities (of various future events), and then (2) taking risks and making decisions so as to minimize the expected loss (or, equivalently, to maximize gain). This approach will be referred to as *probabilistic* or *system identification* approach to handling uncertainty throughout this book. The statistical approach, developed by R. Fisher in early 20th century, was paralleled by the invention of quantum mechanics in physics, at approximately the same time. Quantum mechanics assumes probabilistic description of nature at the microscopic level of atoms. In quantum theory, the future is not precisely determined by the past, as alleged in classical Newtonian physics. Instead, multiple future states of a system are allowed with different probabilities. The time evolution of these probabilities (called wave functions) is quantifiable, in a sense that these distributions are governed by the laws of quantum physics. The philosophical implication is that the physical universe itself is probabilistic rather than deterministic. Note that the quantum physics still adopts the system identification view where the future state of a system can be described, in principle, via some probabilistic model.

The statistical *system identification* approach, however, may not work well in practice, because the probabilities depend on too many unknown factors and cannot be reliably estimated from available limited data. So a more practical approach is to make decisions based on the known risk associated with past events. For instance, a professional stock trader may consider several promising trading strategies (i.e., 'rules' when to buy or sell, and associated price limits). Based on past trading experience, these strategies are evaluated, and the 'best' one is selected for future trading. For the sake of discussion, assume that a trader adopted the following strategy: Buy a broad stock market index S&P 500 on Friday and sell it on the following Monday. This strategy was selected because it worked successfully 70% of the time during the past 52 weeks, and made 50% profit during this time period (whereas the S&P 500 index gained just 15%). While this approach appears reasonable, it is philosophically different from the system identification approach in classical statistics. In our example, a stock trader does not attempt to come up with a 'true statistical model' of the stock market. All he/she is concerned with is choosing a good trading strategy for making money. This selection is based on the minimization of some well-defined measure of risk for past data. Moreover, such a *risk-minimization* approach does not explain why the strategy makes money, as there is no goal of system identification. The main goal is to imitate certain properties of the unknown statistical system that are useful for successful predictions. This leads to the *system imitation* approach to risk management under the framework of predictive learning.

The risk minimization approach is similar to learning in biological systems. In fact, many machine learning methods have been inspired by the learning capabilities of biological systems and, in particular, those of humans. Biological systems learn to cope with the unknown statistical nature of the environment in a data-driven fashion. Babies are not aware of the laws of mechanics when they learn how to walk, and most adults drive a car without knowledge of the underlying laws of physics. Humans as well as animals have excellent pattern recognition capabilities for such tasks as face and voice recognition. People are not born fully equipped with such capabilities, but learn them through data-driven interaction with the environment. Usually humans cannot articulate the rules they use to recognize, for example, a face in a complex picture. The field of pattern recognition has the goal of building artificial systems that mimic human recognition capabilities. These systems are based on the principles of statistical learning rather than biology. Further, risk-minimization learning does not even require knowledge of probability theory or statistics. Our hypothetical stock trader (in the above example) does not need to be an expert on probability, and in fact many successful traders have no mathematical background but they all had plenty of past trading experience. This does not imply that mathematical analysis is not useful for understanding risk-minimization learning, but suggests that developing successful data-driven strategies can be done based on extensive experience and common sense.

The system imitation (or risk minimization) approach also appears to be the main practical alternative for many challenging applications where the number of input variables is larger than the number of data samples used for model estimation. For such applications, the classical system identification approach falls apart. Much of the book describes this system imitation approach, including:

- its mathematical treatment provided by Vapnik-Chervonenkis (VC)-theory;
- conceptual and methodological principles of predictive learning;
- practical machine learning methods introduced under predictive learning framework.

1.3 GROWTH OF EMPIRICAL KNOWLEDGE

As noted in previous sections, the growing use of digital technology in modern society has affected the nature of human knowledge. Knowledge can be broadly defined as a relationship between facts and ideas. Classical science aims at describing many facts (observations) using a few fundamental ideas. Typically, this *first-principle* knowledge is in the form of deterministic relationships between a few basic concepts. Such knowledge has several characteristic properties:

- it describes simple deterministic systems. Here 'simple' refers to conceptual simplicity, rather than actual complexity. For instance, a mechanical system may have many moving objects interacting with each other; however each object is described by simple equations.
- The number of facts (data samples, experimental observations) used to derive or support such knowledge is small.
- The cost of collecting or generating these facts is usually high.

In the modern world the classical balance between facts and ideas has totally shifted. Today, we are flooded with data and are expected to act upon it. With the advent of the Internet, the cost of acquiring, generating and transmitting information has become negligible (practically zero). However, this fast accrual of new information (facts) has not translated into rapid growth of knowledge. Classical statistical estimation approaches may be inadequate for description of complex data-rich systems. Estimation of useful dependencies in such systems requires new methodologies, where the goal of modeling is to act (or predict) well rather than accurate system identification. These approaches are based on the risk-minimization strategies developed in machine learning.

We loosely define *empirical knowledge* as useful dependencies estimated from data or derived from experience. In contrast to first-principle knowledge, empirical knowledge typically:

- describes certain properties of complex systems. Here 'complex' refers to a large number of observed parameters (variables) and to the lack of credible first-principle models;
- is statistical in nature, i.e., allows to make non-deterministic predictions, at best;
- has a quantifiable practical utility for a given application.

We emphasize that our definition requires such knowledge to be *useful* or *actionable*. This is consistent with general notion of learning, which implies accomplishing a specific task, i.e. learning to drive, learning to play piano or learning table manners. Hence, empirical knowledge is synonymous with *instrumental knowledge*, and both terms will be used

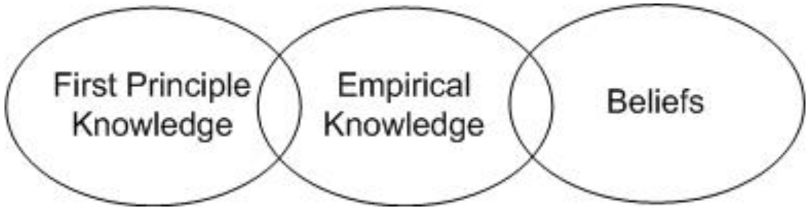


FIGURE 1.1 Three types of knowledge.

interchangeably throughout the book. Empirical knowledge has been used by humans in medicine for centuries (i.e., herbal folk medicine); however, its role has dramatically increased in the digital age.

It is important to differentiate between three types of knowledge: first-principle, empirical and beliefs, as shown in Fig. 1.1. Here *beliefs* refer to empirical models that have no predictive value. The distinction between first-principle knowledge and beliefs is usually easy to make. Examples of true scientific theories versus pseudo-scientific beliefs are: chemistry vs. alchemy and astronomy vs. astrology. The distinction between empirical knowledge and beliefs is not so clear, as both are usually supported by statistical correlations in observed data. Yet this distinction is of great practical importance, because rational humans prefer to act upon knowledge rather than beliefs.

Next we consider a few real-life examples illustrating three types of knowledge.

Example 1.1: Stock Market.

Let us recall a hypothetical stock trader who empirically discovered successful trading strategy for S&P 500 index (i.e., *Buy* on Friday, *Sell* on Monday) by analyzing past data. In order to qualify for empirical knowledge, this strategy should pass a test for 'quantifiable practical utility'. That is, we need to define the objectives of trading *a priori*. For instance, a set of reasonable objectives may be:

(1) to avoid losing money during any 3-month period, under most market conditions;
(2) trading strategy should exceed return of the 'buy-and-hold' strategy for S&P 500 index.

If both objectives have been met under most market conditions, as determined by analysis of past data, then this strategy is likely to have practical utility and can be regarded as empirical knowledge.

(a)

(b)

FIGURE 1.2 Time series of daily closing prices of S&P500 index:
(a) for 1-year period, (b) intraday prices for 3-day period August 5 - 8, 2008.

On the other hand, consider the following statement: the stock market, most of the time, is either in the UP trend, DOWN trend or FLAT trend, implying that this knowledge can be useful for short-term trading. This statement seems to be true and supported by the data. For example, see the graph of daily closing prices of S&P 500 over a one year period in Fig 1.2a. Moreover, the statement holds true even for intraday prices, as shown in Fig. 1.2b. Yet it does not constitute empirical knowledge, because it is not clear how to make practical use of this assertion (i.e. how to make money from it). So this statement should be regarded as belief. In fact, there is a whole field called 'technical analysis' that tries to analyze trends based on past historical prices. However, technical analysis cannot demonstrate *consistently accurate predictions* under varying market conditions, and it has questionable objective value. So, according to our criteria, technical analysis should be regarded as belief, rather than empirical knowledge.

Example 1.2: Kepler's Laws of Planetary Motion.

History of science provides many examples of scientific discoveries based on observations of empirical data. Often, such an empirical knowledge leads to first-principle knowledge. Perhaps the most famous example is Kepler's laws that describe motion of planets in the solar system. Kepler discovered these laws in early 17th century, using comprehensive observational data of planetary motions collected by Danish astronomer Tycho Brahe in late 16th century. Tycho collected very accurate and comprehensive astronomical observations of planets, in order to validate one of the two competing systems of the universe, Ptolemaic (geocentric) or Copernican (heliocentric). Tycho himself was not a Copernican, and he proposed a combined system in which the Sun orbited the Earth while the other planets orbited the Sun. Fortunately, in late years of his life, Tycho had an assistant, Johannes Kepler, who also happened to be a brilliant mathematician. So shortly after Tycho's death, Kepler analyzed volumes of Tycho's data, and extracted three simple laws, known as Kepler's laws:

1. The orbit of every planet is an ellipse with the sun at a focus
2. The line joining a planet to the sun sweeps out equal areas during the same time intervals. See Fig. 1.3.
3. The square of the orbit period of a planet is proportional to the cube of the orbit size (or the length of the major axis).

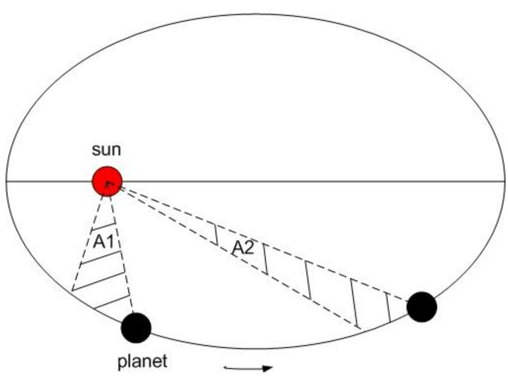


FIGURE 1.3 Illustration of Kepler's laws.

TABLE 1.1. Illustration of third Kepler's law. For any planet, the ratio P^2/D^3 is constant, where P is the orbit period and D is the orbit size.

	P	D	P^2	D^3
Mercury	0.24	0.39	0.058	0.059
Venus	0.62	0.72	0.38	0.39
Earth	1.00	1.00	1.00	1.00
Mars	1.88	1.53	3.53	3.58
Jupiter	11.90	5.31	142.0	141.00
Saturn	29.30	9.55	870.0	871.00

Kepler's laws are illustrated in Fig. 1.3 and Table 1.1. Kepler's laws obviously constitute instrumental empirical knowledge, as they can accurately describe motion of planets. Moreover, these laws can also predict motion of previously unknown heavenly bodies, such as outer planets and asteroids orbiting the Sun. Later, in the end of 17^{th} century, Newton showed that Kepler's laws can be derived from more general first-principle laws of Newtonian mechanics and the law of gravity. So in this historical example, empirical knowledge (Kepler's laws) has eventually led to first-principle laws discovered by Newton.

Example 1.3: Spread of AIDS in Africa.

The HIV epidemic represents a growing problem in sub-Saharan Africa. Early statistical and epidemiological studies in 1990's suggest lower rate of HIV spread in Muslim African countries, relative to non-Muslim countries. One possible explanation is that strict religious constraints on sexuality in Islam reduce the sexual transmission of HIV in Muslim countries. There is strong statistical evidence to support this claim. However, this does not qualify as empirical knowledge, because it does not provide any practical mechanism to reduce the spread of HIV virus, other than forceful conversion of all Africans to Islam (which is not feasible). Later, scientists suggested another explanation, noting that all males in Muslim countries undergo circumcision, which may affect transmission of HIV. This hypothesis has led to more medical research showing that the cells on the underside of the foreskin are prime targets for the virus and that abrasions in the foreskin can invite the infection. The negative correlation between circumcision and the spread of AIDS is truly empirical knowledge, as it leads to a practical strategy to battle

AIDS pandemic in Africa. This knowledge has resulted in many international programs for male circumcision in African countries.

As evident from last example, distinction between instrumental empirical knowledge and beliefs may be subtle, but it is very important. Further, knowledge qualified as 'beliefs' is not necessarily useless, as it may lead to empirical knowledge, and even to first-principle knowledge.

Much of this book describes quantitative methods for estimating empirical knowledge, or learning dependencies, from data. These methods, known as learning algorithms or statistical methods constitute the fields of pattern recognition, machine learning and data mining. However, we also take a broader view, in order to show the connections between these mathematical approaches and philosophical/cultural aspects of inductive learning and risk management. This non-technical perspective, discussed in Chapter 3, is fascinating in its own right, but also becomes important for understanding critical concepts underlying many technical approaches.

1.4 MATHEMATICAL REPRESENTATION OF EMPIRICAL KNOWLEDGE

Earlier sections described empirical knowledge as useful dependencies estimated from past data in qualitative terms. This section explains how these concepts can be formalized mathematically, so that learning from data can be viewed as estimation of a function from noisy samples. Here we only try to present general ideas of such a formal representation, mainly via examples. More formal mathematical descriptions and terminology will be given later in Chapters 2 and 4. In this book, a (scalar) variable is denoted by script letters, such as x or y. Multivariate inputs (or vectors) are indicated by lower-case bold symbols, such as $\mathbf{x}$. Matrices are denoted as upper-case bold letters, such as $\mathbf{X}$.

Let us consider unknown system in Fig. 1.4 that has observable inputs (shown as vector $\mathbf{x}$ an output y), and unobserved inputs (vector $\mathbf{z}$). Then past data (the training set) is in the form $(\mathbf{x}_i, y_i), i = 1, 2, ..., n$, where n is the number of training samples. Unobserved and unknown inputs $\mathbf{z}$ are responsible for the random nature of the unknown system, so that it may produce different output values y, for the same values of observed inputs $\mathbf{x}$. Then the goal of predictive learning is to estimate unknown dependency between the input ($\mathbf{x}$) and output (y) variables,

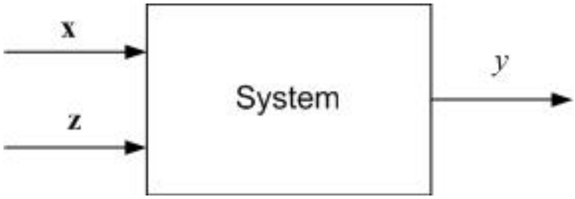
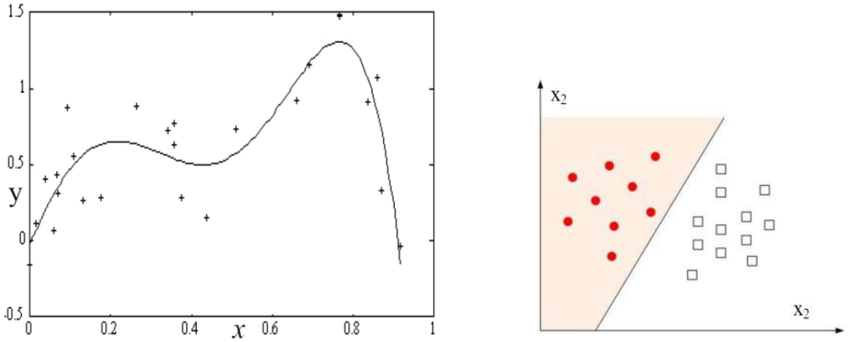


FIGURE 1.4 Unknown system with observed inputs **x** and unobserved inputs **z**.



(a) estimation of real-valued function (b) estimation of indicator function

FIGURE 1.5 Learning as function estimation from noisy samples.

from a set of past observations of $(\mathbf{x}, y)$ values. In other words, learning amounts to estimation of a function $f(\mathbf{x})$ that will be later used for predicting the output for new input samples, called test data or test set. For example, if the output is real-valued, the learning can be viewed as an estimation of a real-valued function from samples. Fig. 1.5a shows an example of a univariate function that can be estimated from noisy samples (training data). Another common situation is when the output is categorical or binary, corresponding to two output classes. In this case, the *indicator function* $f(\mathbf{x})$ is a boundary separating an input space **x** into two halves. For example, two-dimensional training samples and estimated linear indicator function are shown in Fig 1.5b.

Example 1.4: Univariate Regression Estimation.

Consider 10 noisy data samples (x, y) generated according to $y = x^2 + 0.1 \cdot x + \xi$, where additive noise ξ is Gaussian with zero mean and variance $\sigma^2 = 0.25$, and input x is uniformly distributed on

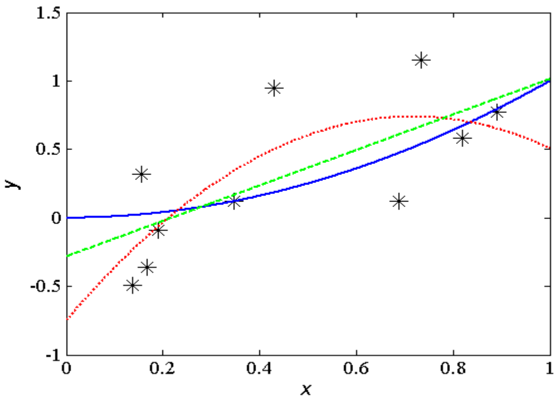


FIGURE 1.6 Fitting linear and second order polynomial model to training data. True target function is shown in blue.

[0,1] interval. Further, unknown function is estimated from training data by least squares fitting of the first-order and second-order polynomial models. Figure 1.6 shows the training data, two estimated models and the true function $g(x) = x^2 + 0.1 \cdot x$. Visual comparison of the two estimated models suggests that the linear model is closer to the true function, and thus will yield better prediction for new test samples. It is interesting to note that using a set of linear models for function estimation yields a better solution than a set of second-order polynomials, even though the 'true model' is a second-order polynomial.

This example illustrates two important observations:

– goal of learning from data is estimation of a good predictive model, rather than true model;
– prediction accuracy (generalization) is related to *complexity* of a set of admissible models. That is, in our example, a class of linear models is simpler than second-order polynomials.

The first point has interesting philosophical implications, that is, the goal of learning is search for instrumental (useful) knowledge, rather than search for the truth. The second point suggests the importance of specifying and controlling model complexity in learning with finite data. Both points will be fully explored later in this book.

Note that representation of learning as a function estimation problem is very powerful. The concepts of 'input variables' and 'function' are quite general. The process of function estimation from noisy samples can be interpreted as a mathematical model for 'inductive inference' or 'inductive learning'. That is, finite number of training samples represents examples, an estimated function is *inductive model*, and using this model to predict outputs (for new inputs) is *deduction*. This is similar to a classical *inductive-deductive* model of reasoning (for logical inference), except that we are using it for noisy data. Most important, this formalization allows us to quantify rather vague notions of explanation and prediction. Recall from Section 1.1, that any data-driven model has two aspects, i.e. *explanation* (of past or known data) and *prediction* (of future data). For classification problems, one can use the fraction of samples misclassified by the decision boundary, as a measure for quality of explanation and prediction, called training error and test error, respectively. For example, decision boundary for data set in Fig 1.5b has zero training error, i.e. it explains perfectly available data. Note that training error measures the quality of 'explanation' of training data. However, the goal of modeling is prediction for unknown test data, i.e. minimization of test error. For regression problems, a common choice for discrepancy between the model estimate $f(\mathbf{x})$ and observed output (y) is squared error $(f(\mathbf{x})-y)^2$. So the quantitative measure for explanation and prediction will be average squared error for training and test data, respectively.

Let us briefly comment on the assumptions behind this formalism:

1. Prediction is possible only if past (training) data and future (test) data have some similarity. In mathematical terms, it is stated as an assumption that both past and future data are generated from the *same unknown distribution*.
2. The two goals of learning, explanation and prediction, usually cannot be achieved perfectly for data-analytic models. This is in contrast to first-principle scientific knowledge which has perfect explanatory and prediction power. For predictive empirical models, the main goal is prediction accuracy.
3. The goal of prediction usually implies that a model performs (predicts) well for many future inputs. That is, the *test set is large*.
4. The problem of function estimation from finite data may have many valid solutions. For example, there may be many linear decision boundaries, separating (explaining) the data set shown in Fig. 1.5b. Such problems are called *ill-posed* in mathematics and are known to be hard. Their difficulty stems from the fact that we attempt to estimate a continuous function from finite noisy data. In this case, there are infinitely many possible solutions (functions) that can

explain available data. Later chapters of this book describe various technical issues for ill-posed learning problems, and corresponding learning algorithms that provide robust performance and good prediction capability.

In a broader picture, it is important to realize that the problem of learning/estimation of dependencies from data is only one part of the general experimental procedure used by scientists, engineers, medical doctors, and others who apply statistical (machine learning, data mining etc.) methods to make inferences from the data. There are many important informal decisions and approximations that lead to a successful formalization of a real-life application as function estimation problem. For instance, one needs to decide which variables are inputs or outputs, how to measure (define) the quality of estimated model etc. These application-specific issues are very important, and they will be illustrated in examples throughout the book.

The general experimental procedure for data-driven modeling can be described as follows:

1. *Understand application goals and requirements.*
2. *Formulate the hypothesis.*
3. *Obtain the data.*
4. *Data cleaning, encoding and preprocessing.*
5. *Model estimation or learning.*
6. *Interpret the model/ draw the conclusions.*

Note that *Step 5* (model estimation) is just one step in the procedure. Good understanding of the whole procedure is important for any successful application. No matter how powerful the learning method used in *Step 5* is, the resulting model would not be valid if the data are not informative, or the problem formulation is not statistically meaningful. For example, poor choice of the input and output variables (*Steps 1 and 2*) and poor data encoding (*Step 4*) may adversely affect model estimation (in *Step 5*), or even make it impossible.

Next we briefly discuss each step in this procedure.

Step1. Understand application goals and requirements. Practical data modeling studies are performed for a given application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement.

Step2. Hypothesis Formulation[*]. The hypothesis in this step postulates the existence of unknown dependency between system's inputs and an output(s), which is to be estimated from experimental data. At this step a modeler usually specifies a set of input and output variables for the unknown dependency and (if possible) a general form of this dependency. There may be several hypothesis formulated for a single problem. Step 2 requires combined expertise of an application domain and of statistical modeling. In practice, it usually means close interaction between a modeler and application experts. This step effectively results in a mathematical formalization of application domain requirements, known as *learning problem setting*. Most descriptions in this book assume standard inductive learning setting. Several novel non-standard learning formulations are discussed in Chapter 9.

Step3. Obtain the Data. This step is concerned with how the data are generated. There are two distinct possibilities. The first is when the data generation process is under control of a modeler - it is known as the *designed experiment* setting. The second is when the modeler cannot influence the data generation process - this is known as the *observational* setting. An observational setting, where the data is generated randomly, is assumed in this book. A random distribution used to generate data is also called a *sampling distribution*. In some applications, the sampling distribution may not be completely unknown and such knowledge, of course, should be reflected in the data collection procedure. Another important modeling assumption is that past (training) data used for model estimation, and the future (test) data used for prediction, originate from the same sampling distribution. Otherwise, predictive models estimated from the training data alone cannot be used for prediction.

Step4. Data Cleaning, Encoding and Preprocessing. This step has to do with both data collection and its subsequent preprocessing. In the observational setting, data is usually "collected" from the existing databases. Data preprocessing includes several common tasks:

---

[*] Here, 'hypothesis' refers to the existence of some (unknown) predictive model that can be estimated from data. That is, hypothesis refers to specification of input/output variables and formalization of application goals as the problem of function estimation from data. This is different from the interpretation in classical science, where 'hypothesis' usually refers to a theory (or well-defined mathematical relationship between variables), and experimental data is then used to support or disprove this theory.

- *Detection of missing data*. Data samples with missing values are common in real-life, and (if undetected) can negatively affect model estimation. For example, if the data comes from a survey, some respondents may not answer selected questions. When the survey data is entered into a database, missing inputs are assigned zero value, but this information may not be properly disclosed to a data modeler, who uses zero values as true input values. The simplest strategy for dealing with missing data is simply to remove these samples from the training set. Alternatively, there are statistical techniques that estimate 'likely' values for missing data, and use these samples for modeling.
- *Outlier Detection.* Outliers are unusual data values that are not consistent with most observations. Commonly outliers are due to gross measurement errors, coding/recording errors, and abnormal cases. Such non-representative samples can seriously affect the model estimated later in step 5. There are two strategies for dealing with outliers: outlier detection and removal as a part of preprocessing, and development of robust estimation methods that are (by design) insensitive to outliers.
- *Data Preprocessing and Scaling.* This includes several steps such as scaling of inputs and different types of encoding techniques. Examples will be discussed throughout the book.
- *Feature Selection/ Dimensionality Reduction.* This step amounts to selection of a small number of informative features from a high-dimensional data. In many real-life applications, good feature selection is the most important part in the whole procedure because it makes the task of estimating dependency much simpler.

Step5. Model Estimation. Each hypothesis in step 2 corresponds to unknown dependency between the input and output features representing appropriately encoded variables. These dependencies are quantified using available data and a priori knowledge about the problem. The main goal is to construct models for accurate prediction of future outputs from the (known) input values. The prediction accuracy is also known as *generalization* capability in biologically inspired methods (i.e., neural networks). Traditional statistical methods typically use fixed parametric functions (usually *linear in parameters*) for modeling the dependencies. In contrast, more recent methods described in this book are based on much more flexible modeling assumptions which, in principle, allow estimation of arbitrary nonlinear dependencies.

Step6. Interpretation of the Model and Drawing Conclusions. In many cases, predictive models developed in step 5 are used for human decision-making. Hence such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex 'black-box' models. Note that the goals of accurate prediction and interpretation are rather different, since interpretable models would be (necessarily) simple but accurate predictive models may be quite complex. Modern machine learning approaches favor methods providing high prediction accuracy, and then view interpretation as a separate task.

Most of this book is on formal methods for estimating dependencies from data (i.e., step 5). However, other steps are equally important for an overall application success. Note that the steps preceding model estimation strongly depend on the application-domain knowledge. Hence practical applications of learning methods require a *combination* of modeling expertise with application domain knowledge. Also, we emphasize that for real-life applications, meaningful interpretation of the predictive learning model in Step 6 usually requires a good understanding of the issues and trade-offs in steps 1 through 4 (preceding actual learning or model estimation).

1.5 SUMMARY AND DISCUSSION

The growing importance of discovering regularities in application data has led to many diverse data-analytic methodologies, each with its own specialized terminology and practical objectives. These include: Pattern Recognition, Machine Learning, Data Mining, Artificial Neural Networks etc. Several ‘definitions’ of what constitutes each field, extracted from popular textbooks, are given below:

- The field of *Pattern Recognition* is concerned with the automatic discovery of regularities in data.
- *Data Mining* is the process of automatically discovering useful information in large data repositories.
- This book (on *Statistical Learning*) is about learning from data.
- The field of *Machine Learning* is concerned with the question of how to construct computer programs that automatically improve with experience.

- *Artificial Neural Networks* perform useful computations through the process of learning.

These definitions suggest very close similarity, yet each field uses its own specialized terminology. Moreover, each field seems to be a collection of computer algorithms that are often re-discovered or borrowed from other fields. Again, a few quotations from popular textbooks are in order:

- The science of *Statistical Learning* plays a key role in the fields of statistics, data mining and artificial intelligence, intersecting with areas of engineering and other disciplines.
- This book (on *Machine Learning*) introduces concepts from statistics, artificial intelligence, information theory and other disciplines.
- *Artificial Neural Network* (ANN) approaches overlap with other areas such as pattern recognition, computer architecture, adaptive signal processing, artificial intelligence etc.

On a closer look, it appears that such 'diversity' is mainly due to historical reasons, rather than technical substance. That is, machine learning was introduced by computer scientists, pattern recognition – by engineers, data mining – by database researchers, neural networks – by psychologists and neuroscientists. The main common theme among all these fields is estimation (learning) of good data-analytic models, while the differences are minor and terminological. For example:

- Data mining differentiates itself (from other fields) in that it is concerned with algorithms for very large data sets. Yet all books on data mining describe general-purpose learning algorithms developed in pattern recognition and statistics. So statisticians often view data mining as just another application domain for statistics, whereas data mining researchers consider statistical methods as one of many useful 'tools'.
- The field of artificial neural networks describes methods for learning from data, using special model parameterizations (called 'neural networks') inspired by a simplified model of biological brain. There is no clear mathematical advantage for using this particular parameterization, other than the original biological motivation.
- Statisticians working on data mining applications tend to adopt a probabilistic modeling approach. That is, 'statistical inference' is defined as the use of a subset of a population called a sample to draw conclusions about the entire population from which it was taken. According to this view, the goal of modeling is to estimate (unknown) probabilistic model of observed data.

This fragmentation and terminological confusion makes it difficult for students (in any field) to understand real research issues involved. So the scientific goal of this book is to expose students to fundamental conceptual issues and trade-off underlying most predictive learning algorithms. This is necessary for understanding and critical evaluation of hundreds of new algorithms being proposed every year in many fields concerned with estimation of models from data. From a practical point, most learning algorithms require proper tuning of (user-defined) parameters. This parameter tuning may be tricky and requires good understanding of the conceptual issues underlying adaptive learning algorithms. For practitioners, the real issue is application of readily available software algorithms, which includes parameter tuning and selection of 'good' learning algorithm appropriate for their data.

This book presents an approach for estimating predictive models from data, based on the idea of risk-minimization, i.e. estimating a model via fitting a set of possible models to available training data. This follows the framework of the Vapnik-Chervonenkis (VC) learning theory, which provides general conditions under which various estimators (implementing this approach) can generalize well. The VC-theory is a mathematical theory. This book follows VC methodological framework, and also tries to relate VC-theoretical concepts to the Philosophy of Science. Ill-posed nature of empirical inference from data has been a subject of lively discussions in philosophy over many centuries, dating back to Hume and Kant. More recently, Wittgenstein summarized the prevailing philosophical view on empirical inference as follows:

> *The process of induction is the process of assuming the simplest law that can be made to harmonize with our experience. This process, however, has no logical foundation, but only a psychological one. It is clear that there are no grounds for believing that the simplest course of events will really happen.*

This view of psychological induction lumps together the instrumental empirical knowledge and beliefs (shown in Fig.1.1) since both are based on empirical data (or experience). So this book pursues two objectives: philosophical and technical. Philosophically, the goal is to differentiate between empirical knowledge and empirical beliefs. Here empirical knowledge constitutes dependencies (estimated from empirical data) which have useful predictive value. The technical objective is to present formal concepts and learning algorithms for estimating empirical dependencies from data. Throughout the book, a strong connection is made between the VC theory (a mathematical theory that describes predictive learning) and philosophical ideas related to induction and inference.

Finally, growing acceptance of data-driven modeling raises important cultural and ethical issues. Simply put, these concerns relate to intellectual integrity of researchers who perform data modeling and institutions that 'own' the data. Not surprisingly, these ethical problems are most evident in life sciences and medical research, where financial implications of data-analytic models are very high. Recent paper by Ioannidis (2005) argues that 'most published research findings (in clinical research) are false'. This happens not because of an outright fraud, but due to self-serving data analysis. According to Ioannidis,

> *People are messing around with the data to find anything that seems significant, to show they have found something that is new and unusual.*

This over-eagerness leads to inherent bias in interpreting statistically insignificant differences and reporting them as significant findings.

Similarly, the quality of empirical studies is adversely affected by under-reporting of negative results. For example, if a drug company provides funding for a clinical study of its newly developed drug, the researchers are under obvious pressure to publish favorable findings but not to publish negative results. Turner et al (2008) provide the following statistics on review of 74 new drug studies submitted to the US Food and Drug Administration (FDA) for approval: 38 studies were judged positive by FDA, and *all but one were published*. However, most of the studies found to have negative or questionable results were not published. Clearly, such a selective publication bias leads to unrealistic estimates of drug effectiveness and distorted risk-benefit ratio for new drugs. One can only wonder about the direction of scientific progress, if similar kind of ethics had prevailed in classical science. It seems inconceivable to imagine Johannes Kepler tinkering with Tycho Brahe's data, or Galileo trying to disregard experimental evidence inconsistent with his theories.

1.6 BIBLIOGRAPHIC NOTES

Historical account of dealing with uncertainty and risk presented in Section 1.2 follows Bernstein (1998). Classical statistical paradigm for modeling uncertainty via probabilistic distribution estimated from data samples is due to Fisher (1952). Example 1.3 (spread of AIDS in Africa) is taken from Gray (2004).

Formalization of the learning problem as function estimation from noisy samples dates back to the pioneering work of Rosenblatt (1962) and Vapnik in 1960's. This formalization is commonly adopted in machine learning and statistics, i.e., see Friedman (1994), Hastie et al (2001), Cherkassky and Mulier (2007).

General experimental procedure described in Section 1.4 originates from classical statistics, and was later adopted in machine learning, pattern recognition and data mining.

Quotations describing the fields of statistical learning, data mining, machine learning, pattern recognition and neural networks, presented in Section 1.5, are taken from several textbooks, including Hastie et al (2001), Tan et al (2005), Mitchell (1997), Bishop (2006), Schalkoff (1997) and Haykin (1999). For additional reading on the relationship between statistics and data mining see Hand (1998) and Hand et al (2001).

REFERENCES

Bernstein, P.L., *Against the Gods: The Remarkable Story of Risk*, John Wiley & Sons, 1996.

Bishop, C.M., *Pattern Recognition and Machine Learning*, Springer, 2006.

Cherkassky, V., and F. Mulier, *Learning from Data: Concepts, Theory and Methods*, Wiley, 2007.

Gray, P. H., HIV and Islam: is HIV prevalence lower among Muslims?, *Social Science and Medicine*, 58, 9, 2004, 1751-1756.

Fisher, R.A., *Contributions to Mathematical Statistics*, Wiley, New York, 1952.

Friedman, J. H., An overview of predictive learning and function approximation, in *From Statistics to Neural Networks*, V. Cherkassky, J. H. Friedman, and H. Wechsler, (eds.), NATO ASI Series F, 136, New York: Springer Verlag, 1994.

Haykin, S. *Neural Networks: A Comprehensive Foundation*, New York: MacMillan, 1994.

Hand, D.J., Data mining: statistics and more?, *The American Statistician*, 52, 112-118, 1998.

Hand, D.J., Mannila, H., and P. Smyth, *Principles of Data Mining*, MIT Press, 2001.

Hastie, T. J., R. J. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, New York: Springer, 2001.

Ioannidis, J. P. A., Contradicted and initially stronger effects in highly cited clinical research, *JAMA*, Vol. 294, No. 2, 218-228, 2005.

Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.

Rosenblatt, F., *Principles of Neurodynamics,* Washington, DC: Spartan Books, 1962.

Schalkoff, R.J., *Artificial Neural Networks*, McGraw-Hill, 1997.

Tan, P.-N., Steinbach, M., and V. Kumar, *Introduction to Data Mining*, Pearson Addison Wesley, 2006.

Turner, E.H., Matthews, A.M., Linardatos, E., Tell, R.A., and Rosenthal, R**.,** Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy, *New England Journal of Medicine*, Vol. 358: 252-260, 2008.

PROBLEMS

1.1. Consider 20 i.i.d. samples from a univariate distribution shown below

1.94	0	1.21	1.23	0	0.25	2.08	0.86	1.38	1.08
5.98	4.48	5.32	5.23	5.02	3.99	4.05	4.62	3.81	3.94

(a) Calculate the sample mean and standard deviation of this data set;
(b) Generate a histogram of this data using 8 bins (using the Matlab function *hist*);
(c) Can this data set be modeled using a Gaussian p.d.f. with parameters found in (a)?

1.2. Generate 12 data samples (x, y) such that x is uniformly distributed in the interval [0,1], and y is normally distributed y ~ $N(0, 0.5)$. Consider modeling this data as $y = f(x) + noise$, using polynomials of degree 1, 2 and 6, to estimate unknown $f(x)$. Polynomial fitting using squared loss can be performed using function POLYFIT in MATLAB. Report the fitting error (MSE) for each model, and show the estimated regression model graphically along with the data samples. Briefly discuss whether the small fitting error can be used as a good indicator for small prediction error for such polynomial models.

1.3. There are 10 guests at a beer party:

(a) What is the number of different ways to serve 10 different types of beer to 10 guests?

(b) The party host has total of 10 bottles of fancy beer. What is the number of different ways to serve these 10 bottles to 10 guests?

1.4. An e-marketing company generates mass email messages (spam) for its customers. The company is paid a percentage of sales generated within 12 hours after spam has been sent. Based on past history, the probability of at least one sale, in response to a single mass mailing, is 0.4%. The company sends 250 different spam messages each day. What is the probability that it generates (positive) daily revenue?

1.5. Show that if two random variables x_1 and x_2 are independent, then their covariance is zero.

1.6. To determine the effect of junk food consumption on the cholesterol level from a blood test, 100 students are put on a fast-food diet. After they have been on this diet for three months, their cholesterol level is taken and compared with the level before the study. A state agency running the study decides to declare the fast-food diet to be harmful if at least 65% of participants exhibit higher cholesterol count in the end of this study. What is the probability that the agency declares fast food harmful when, in reality, it has no effect on the cholesterol level?
Hint: assuming that fast food has no effect on cholesterol, each person's count will be higher, by chance, with probability 0.5. Also, use normal approximation of the binomial distribution.

1.7. Consider daily fluctuations of US stock market prices as reflected in daily closing prices of SP500 stock index. The daily percentage price change of SP500 index is defined as $X(t) = \dfrac{Z(t) - Z(t-1)}{Z(t-1)} * 100\%$, where *Z(t)* is the daily closing prices of SP500, and index *t* denotes current (trading) day. The distribution of this random variable is unknown, however it can be estimated, based on closing prices of SP500 during 2006. (SP500 daily closing prices are available at http://finance.yahoo.com).

(a) Using historical data for *Z(t)*, obtain observations of random variable *X* during year 2006. Then find the following statistical characteristics of *X*:

- Empirical distribution, in the form of the histogram of observed *X*-values, using the range (-2%, +2%), and the bin size 0.2%.
- sample mean and standard deviation of *X*.
- approximate distribution of X using the normal distribution, and show its p.d.f. in a graphical form, using the same scale as for the histogram in part (a).

(b) Using the same historical data for 2006, calculate a 5-day moving average (MA) of the daily closing prices, defined as *MA(t) = [Z(t) + Z(t-1) + Z(t-2) + Z(t-3) +Z(t-4)] / 5*. Then calculate the daily percentage change of MA as $Y(t) = \dfrac{MA(t) - MA(t-1)}{MA(t-1)} * 100\%$.

For this random variable Y, generate the histogram, and calculate its sample mean and standard deviation.

1.8. Consider the daily percentage price change of SP500 (*X*) and the daily percentage price change of a 5-day moving average (*Y*), as defined in Problem 1.7. Find an analytic relationship between:

- the mean values of X and Y
- the standard deviations of X and Y

In your derivations, assume that X –values are statistically independent (i.i.d) random samples. Validate your analytic results using empirical data for year 2006. Interpret the difference between analytic and empirical relationships as an argument for (or against) the main modeling assumption about statistical independence of daily price changes.
For additional validation, repeat your comparisons for each year from 1970 to 2006. Analyze possible differences between analytic and empirical estimates of annual standard deviations during this period. These differences may indicate possible changes in the statistical characteristics of the stock market during this period.

1.9. Consider short-term daily fluctuations of the US stock market. According to popular Random Walk Theory, these short-term market movements are statistically unpredictable. That is, today's price change (Up or Down) cannot be used to predict tomorrow's price change. As a crude test of this theory, suppose we can model the daily price changes (Up or Down) as a statistical coin-toss process. For this coin-toss process, the number of days the market continuously goes Up (or Down) in one direction is a random variable, called *trend duration*. So, one can analytically calculate an average trend duration, and then compare this analytic estimate with empirically estimated average trend duration of the stock market. For example, the random sequence of daily price changes is shown below (where + means Up, and – means Down):

Day 1	+	Recorded as a 5
Day 2	+	
.	+	
.	+	
.	+	
.	-	Recorded as a 3
.	-	
Day 8	-	
Day 9	+	

This sequence results in two (observed) values of trend duration, i.e. 5 and 3.

(1) Calculate analytically the *average trend duration* for the coin-toss process, where the Prob(+) = Prob(-) = 0.5.

(2) Estimate the average trend duration for SP500 in 1991 and in 2005, by plotting the histogram of observed values of the random variable (trend duration) for each year. Then compare the average trend duration for SP500 in 1991 and 2005 with analytic estimate (1), in order to validate the Random Walk Theory.

(3) Using your analytic and empirical results above, hypothesize about the statistical changes in the stock market dynamics during the period from 1991 to 2005. Do these results suggest that the average trend duration in 1991 is closer to random-coin-toss than in 2005? If such meaningful differences are observed, suggest a common-sense explanation.

1.10. A so-called 'January effect' in the stock market suggests that annual price movement (Up or Down) of the US stock market can be predicted based on the price movement in January. Is this statement an empirical knowledge or just a belief that has no practical value? Support your answer by analyzing the stock market (S&P 500 index) for the past 40-year period, 1970 – 2010.

1.11. Consider common racial stereotypes such as 'drunken Irish', 'Arab terrorist', 'Jewish financier'. Do these stereotypes represent:

(a) Non-scientific beliefs?
(b) Statistical inference?
(c) Scientific first-principle knowledge?

Explain your answers.

1.12. (a) Discuss an example of scientific theory discovered based on observations of empirical data. Comment on explanation and predictive capabilities of this theory.

(b) Discuss an example of non-scientific theory (belief) based on observations of empirical data.

1.13. Many ancient civilizations developed sophisticated astronomical theories for predicting periodic events, such as solar eclipses and the movement of planets. Do such theories represent first-principle scientific knowledge, empirical knowledge or pseudoscience? Answer this question using a specific example, e.g., the Mayan calendar – see http://en.wikipedia.org/wiki/ Maya_calendar.

1.14. Suggest data-analytic application of your choice, and discuss in detail each step of the general experimental procedure in Section 1.4. Comment on the difficulty or simplicity of each step in the context of this application example.

1.15. Read the paper by Ioannidis (2005) about the danger of self-serving data analysis. Explain how the general experimental procedure in Section 1.4 can help safeguard against such biased data modeling. Then give a specific example of a recent misleading research finding based on incorrect interpretation of data. Many such examples can be found in mass media (i.e. newspapers and internet).