CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

1

Introduction

Two roads diverged in a wood, and I –
I took the one less traveled by,
And that has made all the difference.

Robert Frost, *The Road Not Taken*

1.1 Statistics and Machine Learning $\boxed{\text{A}}$☺

Back in the early 1970s when the author was starting his undergraduate studies, freshmen interested in studying data analysis would pursue statistics in a mathematics department or a statistics department. In contrast, today's freshmen would most likely study machine learning in a computer science department, though they still have the option of majoring in statistics. Once there was one, now there are two options. Or is machine learning (ML) merely statistics with a fancy new wrapping? In this section, we will try to answer this question by first following the evolution of the two fields.

ML and statistics have very different origins, with statistics being the much older science. Statistics came from the German word *Statistik*, which appeared in 1749, meaning 'collection of data about the State', that is, government data on demographics and economics, useful for running the government. The collected data were analysed and the new science of *probability* was found to provide a solid mathematical foundation for data analysis. Probability itself began in 1654 when two famous French mathematicians, Blaise Pascal and Pierre de Fermat, solved a gambling problem brought to their attention by Antoine Gombaud. Christian Huygens wrote the first book on probability in 1657, followed by contributions from Jakob Bernoulli in 1713 and Abraham de Moivre in 1718. In 1812, Pierre de Laplace published *Théorie analytique des probabilités*, greatly expanding probability from games of chance to many scientific and practical

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

problems. By the time of World War II, statistics was already a well-established field.

The birth of the electronic digital programmable computer by the end of World War II led to the growth of computer science and engineering, and the first successful numerical weather prediction in 1950 (Charney et al., 1950). While computers could compute with enormous speed and solve many problems, it soon became clear that they were very poor at performing simple tasks humans could do easily, such as recognizing a face, understanding speech, and so on.

The term *artificial intelligence* (AI) was invented by John McCarthy when he organized the Dartmouth Summer Research Project on Artificial Intelligence in 1956, an 8-week summer school held at Dartmouth College in Hanover, New Hampshire with about 20 invited attendees. This seminal workshop was considered by many to spark the field of AI research, with AI research mainly pursued by computer scientists/engineers and psychologists.

Machine learning (ML) is a major branch of AI[1] that allows computers to learn from data without being explicitly programmed. As for the origin of the term "machine learning", Turing (1950) raised the question 'can computers think?' and introduced the concept of "learning machines". In the 1955 Western Joint Computer Conference in Los Angeles, there was a session on "Learning Machines" (Nilsson, 2009), while the term 'machine learning' appeared later in Samuel (1959).

Meanwhile, the general public has become fascinated with the new genre of science fiction, depicting machines with human intelligence. Under this intoxicating atmosphere, some AI researchers became unrealistically optimistic about how soon it would take to produce intelligent machines; thus, a backlash against overpromises became inevitable. The UK Science Research Council asked the Cambridge Lucasian professor Sir James Lighthill to evaluate the academic research in AI with an outsider perspective, as Lighthill was a fluid dynamicist. The 1973 report was largely negative, stating that 'Most workers in AI research and in related fields confess to a pronounced feeling of disappointment in what has been achieved in the past twenty-five years. Workers entered the field around 1950, and even around 1960, with high hopes that are very far from having been realised in 1972. In no part of the field have the discoveries made so far produced the major impact that was then promised' (Lighthill, 1973). AI was devastated, as the UK closed all academic AI research except at three universities. Around the same time, the US Defense Advanced Research Projects Agency (then known as 'ARPA', now 'DARPA'), which had been the main source of AI funding in the US, also lost faith in AI, leading to drastic funding cuts. There were two major 'AI winters', periods of poor funding in AI, lasting around 1974–1980 and 1987–1993 (Crevier, 1993; Nilsson, 2009).

As AI suffered a tarnished reputation during the long AI winters, many researchers in AI in the mid-2000s referred to their work using other names, such as machine learning, informatics, computational intelligence, soft computing, data

[1] AI has other branches besides ML; for example *expert systems* were once very popular but have almost completely disappeared.

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

driven modelling, data mining and so on, partly to focus on a more specific aspect and partly to avoid the stigma of overpromises and science fiction overtones associated with the name 'artificial intelligence'. Google Trends reveals how terminology usage changes over time (see Fig. A1 in Appendix A), with 'machine learning' having been searched more often than 'artificial intelligence' on Google since 2015.

The general goal of having computers learning from data without being explicitly programmed was achieved in 1986 with an artificial neural network model called the multi-layer perceptron (Rumelhart et al., 1986a).[2] The rise of the Internet in the mid-1990s meant the connection of numerous computers and datasets, thereby introducing a huge amount of data for ML to extract useful information from. The commercial potential was quickly recognized, leading to the spectacular growth of many high technology companies, which in turn poured massive amounts of funding into ML and AI research.

By the late 1990s, statisticians introduced the new term '*data science*' to broaden statistics by including contributions by computer scientists (C. Hayashi, 1998; Cleveland, 2001). Data science is an interdisciplinary field that tries to extract knowledge from data using techniques from statistics, mathematics, computer science and information science. As such, one could consider statistics and ML as components within data science.

Did the separate paths of evolution taken by statistics and ML bring them to more or less the same domain within data science? Certainly there is partial overlap between the two, as it is not uncommon to have similar methods developed independently by statisticians and by ML scientists. Nevertheless, ML and statistics have their own distinct characteristics or cultures (Breiman, 2001b; D. R. Cox, Efron, et al., 2001). In fact, the two cultures are sufficiently different that it would be very difficult for ML to germinate from within statistics. For instance, one would expect counterculture art or music movements to germinate from societies with liberal laws rather with rigorous laws. Statistics, rooted in mathematics, requires a high standard in mathematical rigour for publications. While rigorous proofs can usually be derived for linear models, they may not even exist for the non-linear models used in ML. Thus, ML can only germinate within a culture that supports a more liberal, heuristic approach to research. Not surprisingly, ML germinated mainly from computer science, psychology, engineering and commerce.

When fitting a curve to a dataset, a statistician would ensure the number of adjustable model parameters is small compared to the sample size (that is, the number of observations) to avoid *overfitting*, that is, the model fitting to the noise in the data as the model becomes too flexible with abundant adjustable parameters (see Section 1.3). This prudent practice in statistics is not strictly followed in ML, as the number of parameters can be greater, sometimes much greater, than the sample size, as ML has developed ways to avoid overfitting while using a large number of parameters. The relatively large number of pa-

[2] While multi-layer perceptron neural network models became very popular after the appearance of Rumelhart et al. (1986a), there had been important contributions made by earlier researchers (Schmidhuber, 2015, section 5.5).

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

rameters renders ML models more difficult to interpret than statistical models; thus, ML models are often regarded somewhat dismissively as 'black boxes'.

The rationale of using large numbers of model parameters in ML is based on AI's desire to develop models following the architecture of the human brain. The following argument is attributed to Geoffrey Hinton, who often used it in his lectures: In the brain, there are more than 10^{14} synapses, that is, connections between nerve cells; thus, there are more than 10^{14} adjustable parameters in the brain. A human lifetime is of the order of 10^9 seconds, and learning say 10 data points per second implies a total sample size of 10^{10} in a lifetime. Thus, the number of parameters greatly exceeds the sample size for the human brain. In other words, if AI is to model the human brain function it has to explore the domain where the number of model parameters exceeds the sample size. For instance, in the ILSVRC-2012 image classification competition, the winning entry from Hinton's team used 60 million parameters trained with about 1.2 million images (Krizhevsky et al., 2012).

Dualism in nature was noted by the ancient Greek philosopher Heraclitus and in the Chinese philosophy of *yin and yang*, where opposite properties in nature may actually be complementary and interconnected and may give rise to each other as a wave trough gives rise to a wave crest. Yin is the shady or dark side and yang the sunny or bright side. Examples of traditional yin–yang pairs are night–day, moon–sun, feminine–masculine, soft–hard, and so on, and we can now add ML–statistics to the list as the yin and yang sides of data science (Fig. 1.1).

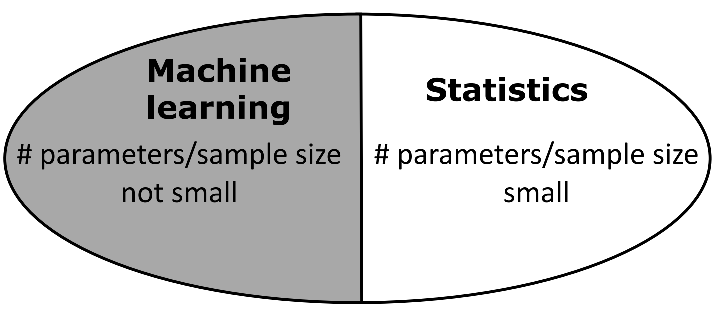

Figure 1.1 ML and statistics tend to occupy different parts of the data science space, as characterized by the number of model parameters to the sample size. In reality, there is overlap and more gradual transition between the two than the sharp boundary shown (see the Venn diagram in Fig. 15.1).

In journeys of discovery, the obscure yin side is often explored after the yang side. For instance, the European maritime exploration to India first proceeded eastward, and only westward from the time of Columbus, as sailing far into the obscure western ocean was not considered sensible nor profitable. In cosmology, the search had originally focused on visible, ordinary matter, but later it was

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

found that ordinary matter accounts for only 4.9% of the Universe, while the rest is the dark universe, containing dark matter (26.8%) and dark energy (68.3%) (Hodson, 2016). Similarly, in data science, the domain where the number of parameters is small relative to sample size was explored first by statisticians, and the seemingly meaningless domain of a large number of parameters was only explored much later by ML scientists, driven by their interest in building models that simulate the human brain. The old constraint requiring the number of parameters to be no larger than the sample size turned out to be breakable, much like the 'sound barrier' preventing supersonic flight. In Fig. 1.1, the yin and yang domains are drawn to be equal in size – in reality, the domain where the number of parameters is restricted to be small is much smaller than the domain without this restriction. The reason ML enjoyed much faster growth than statistics in recent decades is that the solutions of many problems in image and speech recognition, self-driving cars and so on lie in the domain of a large number of parameters.

Another major difference between the statistics and ML cultures lies in their treatment of predictor variables (Breiman, 2001b). Predictor selection, that is, choosing only the relevant predictor variables from a pool of predictors, is commonly practiced in statistics but not often in ML. ML generally does not consider throwing away information a good practice. Furthermore, first selecting predictors based on having high correlation with the response variable then building statistical/ML models leads to overestimation of the prediction skill (DelSole and Shukla, 2009).

In summary, the main tradeoff between statistics and ML is *interpretability* versus *accuracy*. With relatively few parameters and few predictors, statistical models are much more interpretable than ML models. For instance, the parameters in a linear regression model give useful information on how each predictor variable influences the response variable, whereas ML methods such as artificial neural networks and random forests are run as an ensemble of models initialized with different random numbers, leading to a huge number of parameters that are uninterpretable in practical problems. However, as datasets become increasingly larger and more complex, interpretability becomes harder and harder to achieve even with statistical models, while the advantage in prediction accuracy attained by ML models makes them increasingly attractive.

In physics, a similar transition occurred between classical mechanics and quantum mechanics in the 1920s. The clear deterministic view of classical mechanics was replaced by a fuzzy, random picture for atomic particles, thanks to revolutionary concepts like the Heisenberg uncertainty principle, wave-particle duality, and so on. Uncomfortable with the apparent randomness of nature in quantum mechanics, Einstein protested with the famous quote 'God does not play dice with the world' (Hermanns, 1983). A modern physicist learns both classical mechanics and quantum mechanics, using the former on everyday problems and the latter on atomic-scale problems. Similarly, a modern data scientist learns both statistics and machine learning, choosing the appropriate statistical or ML method based on the particular data problem.

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

1.2 Environmental Data Science A ☺

Environmental data science is the intersection between environmental science and data science. *Environmental science* (ES) is composed of many branches – atmospheric science, hydrology, oceanography, cryospheric science, ecology, agricultural science, remote sensing, climate science, environmental engineering, and so on, with the data from each branch having their own characteristics. Often, the plural term 'environmental sciences' is used to denote these branches. Statistical methods have long been popular in the environmental sciences, with numerous textbooks covering their applications in climate science (von Storch and Zwiers, 1999), atmospheric science (Wilks, 2011), oceanography (Thomson and Emery, 2014) and hydrology (Naghettini, 2007).

Environmental data tend to have different characteristics from non-environmental data. Most non-environmental datasets in ML applications contain *discrete data* (e.g. intensity of colour pixels in an image) and/or *categorical data* (e.g. alphabets and numbers in texts),[3] whereas most environmental datasets contain *continuous data* (e.g. temperature, air pressure, wind speed, precipitation amount, pollutant concentration, sea level, salinity, streamflow, crop yield, etc.). The discrete/categorical data from ML problems are in general bounded, that is, having a finite domain – for example, a colour pixel normally has intensity values ranging from 0 to 255, while texts are typically composed of 26 alphabets and 10 digits (plus upper cases and some special symbols). In contrast, continuous data are in general not bounded; for example, there are no guaranteed upper limits for variables such as wind speed, precipitation amount and pollutant concentration.

The most common data problem consists of predicting the value of an output variable (also known as a *response* variable or *dependent* variable) given the values of some input variables (a.k.a. *predictors* or *features*).[4] If the output variable is discrete or categorical, this is a *classification* problem, whereas if the output is continuous, it is a *regression* problem. Again, classification is much more common in non-environmental datasets, and many ML methods were developed first for classification and later modified for regression, such as support vector machines (Cortes and V. Vapnik, 1995; V. Vapnik et al., 1997).

After a model has been built or trained with a *training* dataset, its performance is usually evaluated with a separate *test* dataset. If the test input data lie outside the domain of the input data used to train the model, the model will be forced to do *extrapolation*, yielding inaccurate or even nonsensical predictions. Figure 1.2 illustrates why the outlier problem can be much worse with unbounded continuous input data than with finite-domain discrete/categorical data. Thus, making accurate predictions using environmental data could be a much harder problem than typical non-environmental data problems.

[3] The difference between categorical data (e.g. water, land, snow, ice) and discrete data (e.g. 1, 2, 3) is that categorical data normally have no natural ordering, though some categorical data (namely *ordinal data*) do have natural ordering (e.g. sunny, cloudy, rainy). See Section 2.1.

[4] "Predictors" are used in the statistics literature while "features" are used in ML.

© in this web service Cambridge University Press & Assessment www.cambridge.org

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

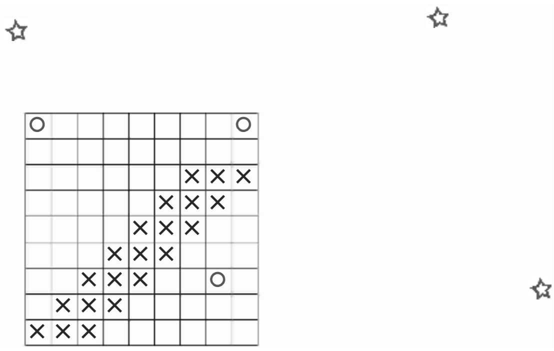

Figure 1.2 Schematic diagram illustrating the problem of outliers in the input data in 2-D. The grid illustrates a finite-domain discrete input data space with crosses indicating training data and circles marking outliers in the test data. For unbounded continuous input variables, the test data can lie well outside the grid and much farther from the training data, as illustrated by the stars.

Let us look at an example of input outliers. A common air quality measure of fine inhalable particles with diameters $\leq 2.5\ \mu$m is the $PM_{2.5}$ concentration. For predicting the hourly $PM_{2.5}$ concentration in Beijing, an important predictor is the cumulated precipitation (X. Liang et al., 2016), as the pollutant concentration drops after precipitation. When data from 2013 to 2015 were used for training non-linear regression models and data from 2010 to 2012 were used for testing, Hsieh (2020) noticed that the cumulated precipitation of an intense precipitation event reached 223.0 mm in the test data in July 2012, whereas the maximum value in the three years of training data was only 51.1 mm, that is, this input in the test data was over four times the maximum value in the training data, which led to wild extrapolation (Section 16.9).

From the old saying 'climate is what you expect; weather is what you get' (a similar version originated from Mark Twain), it follows that environmental problems also tend to group into 'weather' and 'climate' problems, with the former concerned with short-term variations and the latter concerned with the expected values from long-term records or with longer-term variations. For instance, by averaging daily weather data over three months, one obtains seasonal data and can build models to predict seasonal variations. Farmers, utility companies, and so on have great interest in seasonal forecasts, for example on whether next season will be warm or cool, dry or wet.

The averaging of weather data to form climate data changes the nature of the data through the central limit theorem from statistics. To illustrate this effect, consider the synthetic dataset

$$y = x + x^2 + \epsilon, \tag{1.1}$$

where x is a random variable obeying the Gaussian probability distribution with zero mean and unit standard deviation (see Section 3.4) and ϵ is Gaussian noise

© in this web service Cambridge University Press & Assessment

CAMBRIDGE

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

with a standard deviation of 0.5. Averaging these 'daily' data over 30 days reveals a dramatic weakening of the non-linear relation in the original daily data (Fig. 1.3). Thus, in this example, a non-linear regression model will greatly outperform a linear regression model in the daily data but not in the 30-day averaged data. In the real world, tomorrow's weather is not independent from today's weather (i.e. if it is rainy today, then tomorrow will also have higher odds of being rainy). Thus, the monthly data will be effectively averaging over far fewer than 30 independent observations as done in this synthetic dataset, so the weakening of the non-linear relation will not be as dramatic as in Fig. 1.3(c). Nevertheless, using non-linear regression models from ML on climate data will generally be less successful than using them on weather data due to the effects of the central limit theorem (Yuval and Hsieh, 2002).

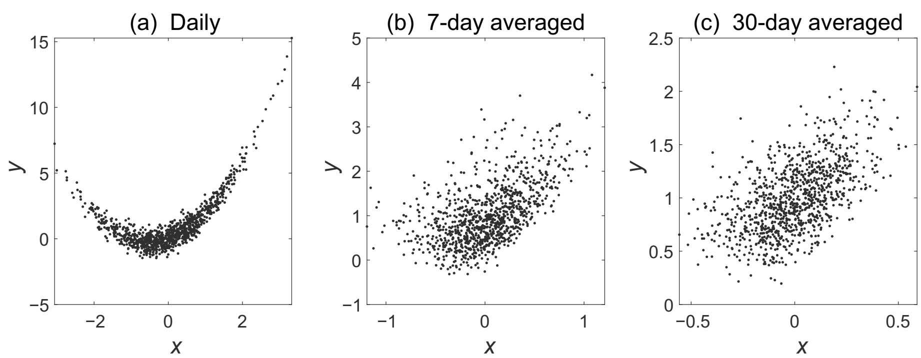

Figure 1.3 Effects of time-averaging on the non-linear relation (1.1). (a) Synthetic 'daily' data from a quadratic relation between x and y. The data time-averaged over (b) 7 observations and (c) 30 observations. [Follows Hsieh and Cannon (2008).]

Obtaining climate data from daily weather data by taking the time mean or average is no longer the only statistic used. In the last couple of decades, there has been a growing interest in the climate of extreme weather events (simply called 'climate extremes'), as global climate change may affect the extremes even more than the means. There is now a long list of such climate extreme variables derived from daily data, for example, the annual number of frost days, the maximum number of consecutive days when precipitation is < 1 mm, and so on (X. B. Zhang et al., 2011) and ML methods have been used to study climate extremes (Gaitan, Hsieh et al., 2014).

Like seeds broadly dispersed by the wind, ML models landing in numerous environmental fields germinated at different rates depending on the local conditions. If a field already had successful physics-based models, ML models tended to suffer from neglect and slow growth. Meteorology, where dynamical (a.k.a. numerical) models have been routinely used for weather forecasting, has been

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE

Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

slower to embrace ML models than hydrology, where by the year 2000 there were already 43 hydrological papers using neural network models (Maier and Dandy, 2000). ML models were readily accepted in hydrology because physical-based hydrological models were not skillful in forecasting streamflow.[5] Remote sensing is another field where ML was quickly adopted (Benediktsson et al., 1990; Atkinson and Tatnall, 1997).

Compared to linear statistical models, non-linear ML models require larger sample sizes to excel. Oceanography, a field where collecting *in situ* observations is far more difficult than in meteorology or hydrology, and climate science, where the long timescales involved preclude large effective sample size, are fields where the adoption of ML have been relatively slow among the environmental sciences. Zwiers and Von Storch (2004) noted: 'much of the work that has had a large impact on climate research has used relatively simple techniques that allow transparent interpretation of the underlying physics'. Nevertheless, in the last few years, ML has grown rapidly even in fields such as oceanography and climate science. Perhaps even more unexpectedly, divergent approaches such as ML and physics have been merging in recent years within environmental science (Chapter 17). The history and practice of AI/ML in the environmental sciences have been reviewed by S. E. Haupt, Gagne et al. (2022) and Hsieh (2022).

1.3 A Simple Example of Curve Fitting $\boxed{\text{A}}$☺

In this section, we will illustrate some basic concepts in data science by a simple example of curve fitting, using one independent variable x and one dependent variable y. Assume the true signal is a quadratic relation

$$y_{\text{signal}} = x - 0.25\,x^2. \tag{1.2}$$

The y data are composed of the signal plus random noise,

$$y = y_{\text{signal}} + \epsilon, \tag{1.3}$$

where the noise ϵ obeys a Gaussian probability distribution with zero mean and standard deviation being half that of y_{signal}. The advantage of using synthetic data in this example is that we know what the true signal is.

A polynomial of order m,

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + \cdots + w_m x^m, \tag{1.4}$$

has $m + 1$ adjustable model parameters or weights w_j $(j = 0, ..., m)$, with $\hat{y}$ denoting the output value from the polynomial function as opposed to the value y from the data. Polynomials of order 1, 2, 4 and 9 are fitted to the training

[5] The difficulty lies in the subsurface flow passing through material, which is not easily observable. The subsurface flow is also complex and non-linear, thereby requiring many parameters to cover for the inexact physics, resulting also in poor model interpretability (Karpatne et al., 2017).

© in this web service Cambridge University Press & Assessment www.cambridge.org

CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-06555-0 — Introduction to Environmental Data Science
William W. Hsieh
Excerpt
More Information

dataset of 11 data points in Fig. 1.4 by minimizing the *mean squared error* (MSE) between the model output $\hat{y}$ and the data y,

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2, \tag{1.5}$$

where there are $i = 1, ..., N$ data points.

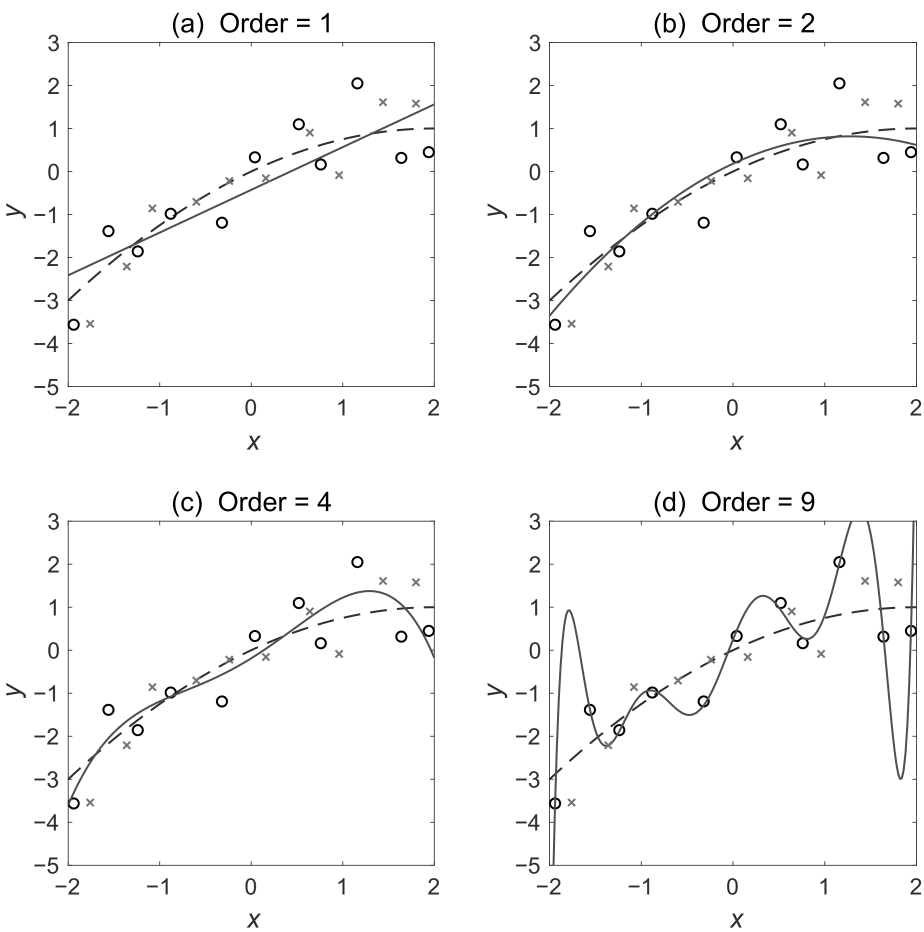

Figure 1.4 Polynomial fit to data using a polynomial of order (a) 1, (b) 2, (c) 4 and (d) 9. The circles indicate the 11 data points used for fitting (i.e. training), the solid curve the polynomial solution $\hat{y}$ and the dashed curve the true signal ($y_{\text{signal}} = x - 0.25\,x^2$). The crosses show 10 new data points used to validate the polynomial fit.

For order 1 (Fig. 1.4(a)), the polynomial reduces to a straight line and the problem is simple linear regression. As the order of the polynomial increases,

© in this web service Cambridge University Press & Assessment www.cambridge.org