# 2. Measurement and data: variables
## 2.1 Outline of measurement

In the following basic concepts and means for handling data and variables are considered, i.e., measurement and variable types that have to be known in order to ”understand” data and what operations can be made for different variable types in later course sections in association with machine learning algorithms used for data mining tasks.

Variables are also called attributes, measurements, features or parameters, sometimes also properties, indicators or factors. Mostly, this depends on the context. (We try to keep variable.)

# Measuring the world

The world is a place of unbelieveable complexity. No matter how closely we look at or somehow measure a facet of the world, there is an infinite depth of detail. Yet our brains and minds construct meaningful simplicities from the stunning complexity surrounding us.

The data itself from which information is to be discovered, however rich and copious, is but a pale reflection of the real world. Data do not provide more than a pale and hazy shadow, a murky outline, of the true workings of the world. Yet this thin gossamer wisp is just enough for us to grasp at the edges of understanding and to attempt analyse phenomena and objects in the world.

# Capturing measurements

Often measurements or variable values all are taken at the same time in order to compare some matters. For instance, in Fig. 1.3. the data used are (mostly) from the same year. In effect, the world state was ”frozen” by the validating circumstance and the measurements taken yielding a particular value. The validating feature was a timestamp. In Fig. 1.2 there are two signals or time series measured with a constant interval.

Measurement implies that there are some quantity to measure and some device or way to calibrate the measurement against. A simple illustration of such a physical measurement is measuring a distance with a ruler. A nonphysical measurement is an opinion poll calibrated in percentage points.

# Errors of measurements

There are several ways in which a measurement may be in error. Calibration may be poor, e.g., the ruler might slip out of position, leading to inaccurate distance value. The device itself may be inaccurate, e.g., a ruler being longer or shorter than the standard lenght. There are inevitably errors of precision. For example, measurements of distance must be truncated at some point, such as 1 mm for a ruler.

Since there are likely to be as many measurements short as there are long, such errors also tend to cluster about the ”correct” point. If the calibration is in error – say wrong ruler lenght - this leads to a systematic error, since all measurements made with the given ruler tend to be ”off” of the mark by the same amount. This is described as *bias*.

Fig. 2.1 shows what unbiased error might do to a measurement. Fig. 2.2 shows what bias added to unbiased error might look like.

Environmental errors are rather different in nature, but of particular importance in data mining. They express the uncertainty due to the nature of the world.

The ”frozen” state is often not possible, this may yield some errors between sequential measurements.

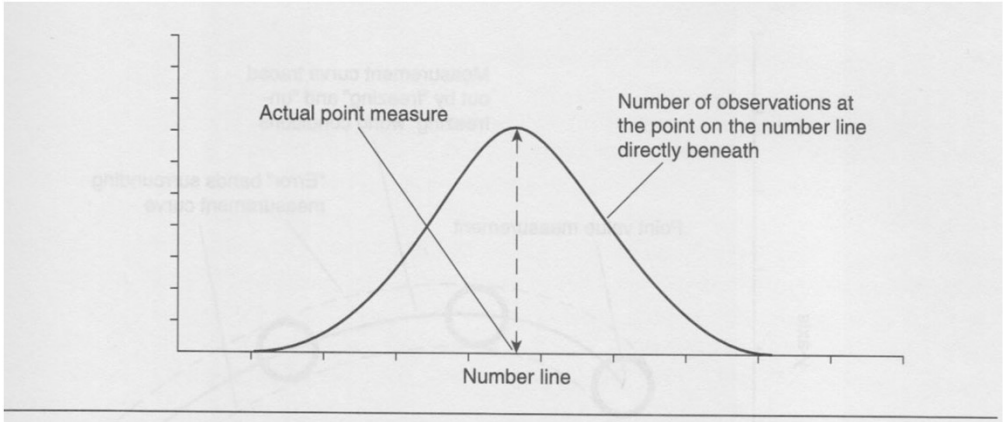



**Figure 2.1** Unbiased noise spreads the measurements evenly around the measurement point. Most cluster near the actual value.

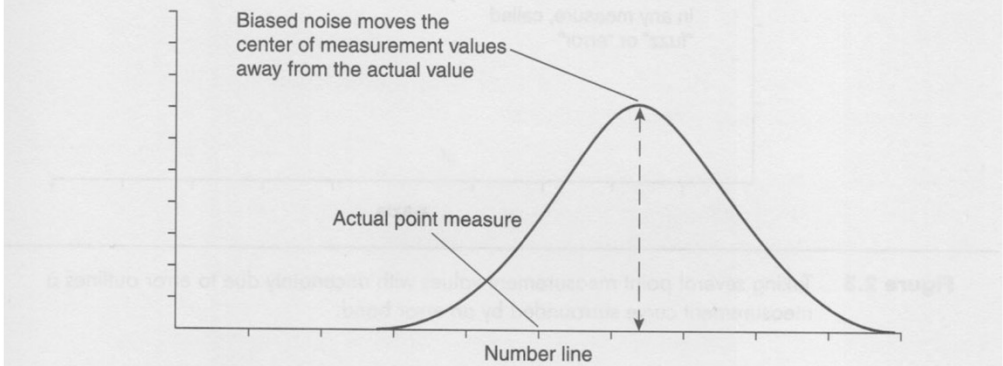



**Figure 2.2** Biased noise makes most of the measurements cluster around a point that is not the true measurement.

Fig. 2.3 shows an example, where each fuzzy circle represents such a single measurement. The central point represents the idealized point value and its surrounding the unavoidable accompanying error.

Suppose now that the world is ”unfrozen”, conditions allowed to change minutely, and then ”refrozen”. If the driving factors are linearly related to the measurement, this under slightly changed circumstances is slightly changed in direction and distance from the first measurement.

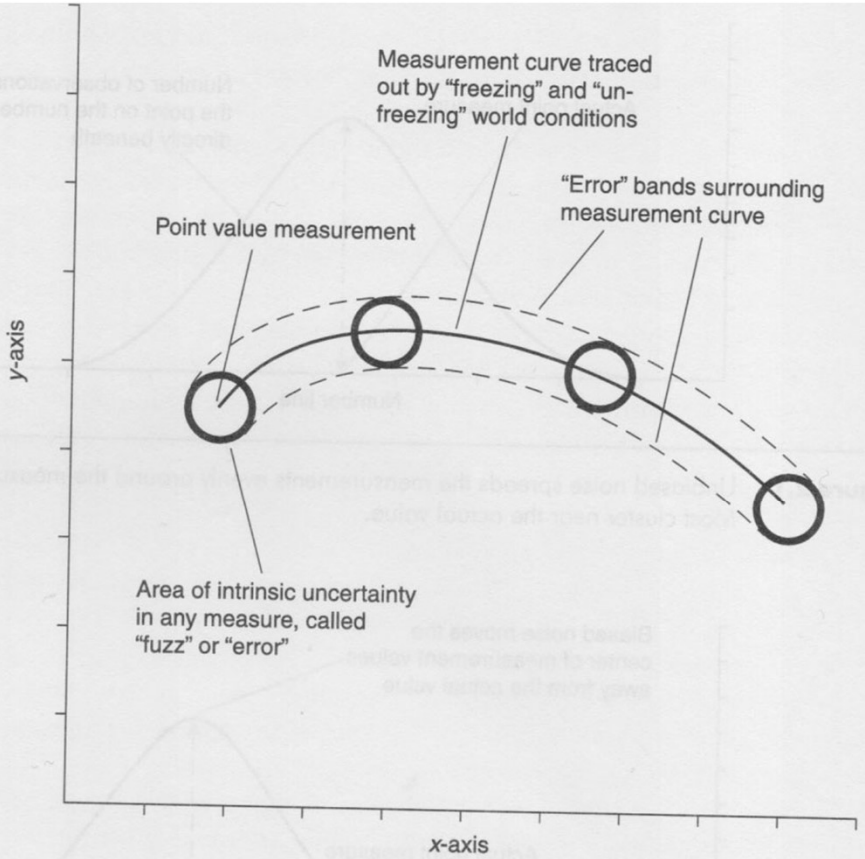



**Figure 2.3** Taking several point measurement values with uncertainty due to error outlines a measurement curve surrounded by an error band.

Let us assume that the measurements in Fig. 2.3 represented the interest of a bank account. Perhaps a small change in interest rate persuaded one to take all one's money out of that bank and deposit it in another bank. If they measure such from several banks, the situation might be like that in Fig. 2.4. The situation would be a complicated "curve bundle" and describes how the point measurements might map onto the world under slightly different circumstances.

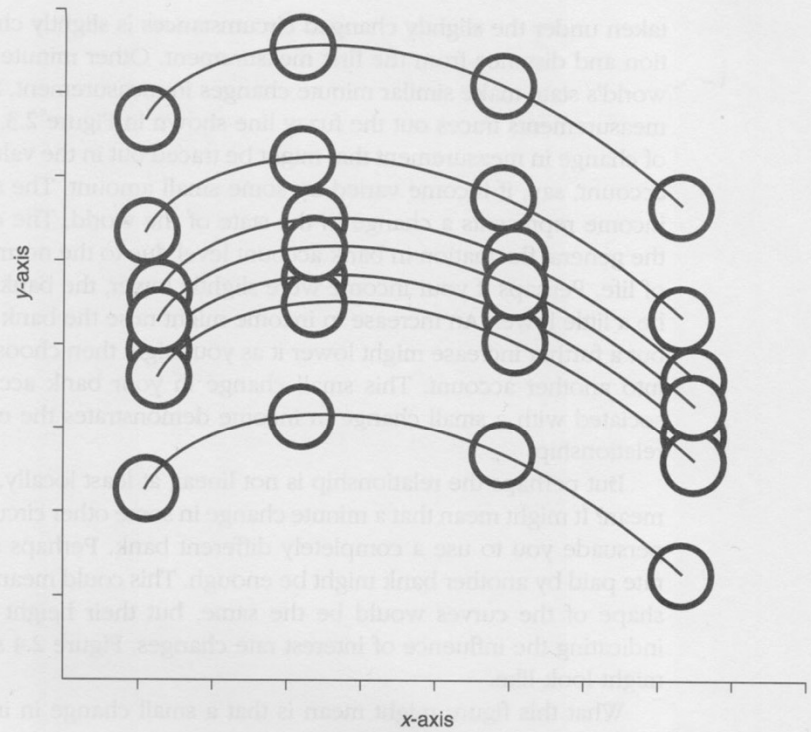



**Figure 2.4** Groups and clusters of curves that result when a small change in world conditions makes a nonlinear or "step" change in the measured values.

# Bias vs. variance

To evaluate goodness of prediction of a model ***in theory***, we calculate error. There are two components of error: *bias* and *variance*. Bias can be defined as the error that cannot be reduced by increasing the size of a data set. It is a systematic error that would be present even if the size were approaching infinity. In reality bias cannot be precisely defined. Thus, the following is just a theoretical approach.

Bias *B* is calculated as the difference between the estimated expected value *E*(·) (estimate of mean) and the (hypothetical) true value of some variable *p* ($\hat{p}$ its estimate). Its squared value and variance are two components of mean squared error *MSE*.

$$B = E(\hat{p}) - p \qquad MSE(\hat{p}) = E(\hat{p} - p)^2 = S^2(\hat{p}) + B^2(\hat{p}) = \text{variance} + \text{bias}^2$$

In the preceding, an unbiased estimation of the sample variance is:

$$S^2 = \frac{\sum_i \left(\hat{p}_i - p_i\right)^2}{n-1}$$

(It is as biased if divided by *n.* In practice this is meaningful for small *n* only).

Variance can be defined as an additional error (to bias) that is incurred given a finite sample, because of sensitivity of random fluctuations. The simplest example is the sample mean that estimates the population mean, i.e., a small subset of the (quite hypothetical) whole. The sample mean is more or less biased. The bias-variance is ***theoretically*** according to Fig. 2.5.

***Remark! Bias and variance can be reduced, e.g., with better measuring equipment and more accurate measurements, but they cannot be fully deleted.***

Fig. 2.5 Illustration of the bias-variance dilemma.

If the model complexity increases, this leads to overfitting (overtraining or overlearning) of a model. We should be able to find something between the extremes of Fig. 2.6.

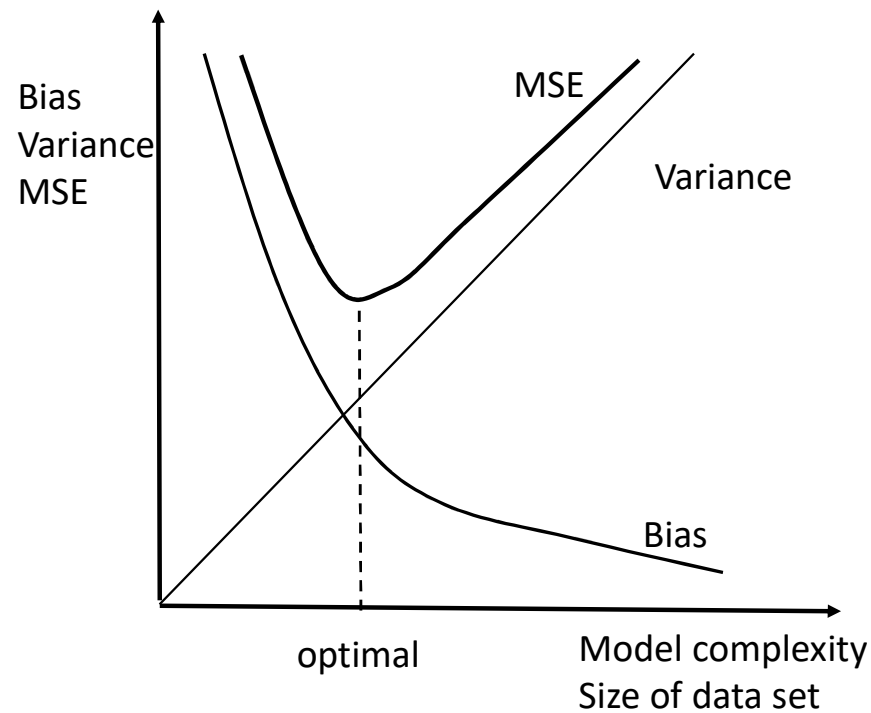
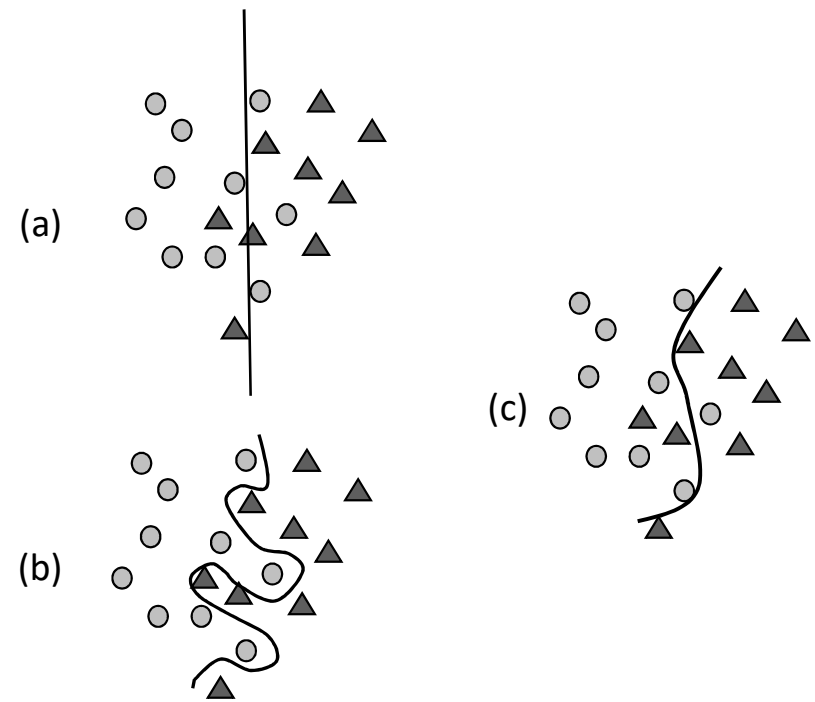


Fig. 2.6. (a) Too simple and (b) too complicated ”models” for separating two classes. (c) Another that is a better ”generalizing” model if the data set changes slightly, e.g., along with a new data set from the same data source.



# Tying measurements to the real world

Sometimes measurements are described as consisting of two components: the actual absolute perfect value and distortion. The latter is often referred to as error. Use of ”error” has unfortunate connotations as if there is somehow something wrong with the measurement. The distortion is actually an integral part of the measurement. While some part of the distortion may indeed result from a mistake on the part of measurement, much of the distortion is not only unavoidable, but is actually a crucial part of what is being measured.

In data mining, the error is included in a more or less random phenomenon called ’noise’.

# 2.2 Types of variables or measurements

Variable types are categorized in many ways. Some of the distinctions arise from the nature of the properties of measured phenomena or matters, while others arise from the use to which they are put.

A variable represents a measurement that can take on a number of particular values, with a different value possible for each case or instance.

The group of variables that can be indicated by the position of a single point (value) on some particular scale are called *scalar variables.* There are also such that require more values, vector variables, but scalar variables are much more usual and important in data mining.

# Scalar measurements

**Nominal scalar measurements (nominal variables)**

Values that are nominally scaled carry the least amount of information of the types of measurements to be considered. *Nominal* values are just to name things. They are like labels used for purposes of identification. There is no inherent order in nominal measurements. They could be such as the eye colors of subjects like blue, brown and green. The only computational operation usable for them is equivalence relation operator or ’is equal to, =’ (or ≠). We can compute neither maximum nor mean for a set of nominal values, but use only ’*mode*’ (the most frequent value) as statistical central value.

**Categorical scale measurements (categorical variables)**

*Categorical* measurements name groups of things, not individual entities. They are much like nominal values, but also allow values to be grouped in meaningful ways. As with nominal measurements, they are no more than labels for different groups. They are, for instance, marital status or mother tongue. Even if they were encoded numerically, no order could be given them. Such variable values can also be presented with characters or symbols. Numerically labeled or not, all that can be said about the categorical values is whether they are different. There is no information included in categorization to indicate how different they are from each other. We cannot express that a plumber is twice a carpenter.

## Ordinal scale measurements (ordinal variables)

When we include more information to variables than the nominal or categorical have we use ordinal measurements. For instance, we can order concepts or values ’short’, ’medium’ and ’tall’.

The ranking of ordinal values must be done subject to a very particular condition, called *transitivity*. This means that if, say, grade A is ranked higher than B, and B higher than C, then A must be higher than C, i.e., if A>B and B>C, then A >C. Mode and *median* can be calculated for ordinal values.

The ordinal scale does not require that the amount of the difference between ordinal values should be specified. Notwithstanding this, they are often applied ”liberally”, encoded with integers 0, 1, 2, …, $m_{max}$ (an appropriate maximum) and used as the following variable type.

**Interval scale measurements (variables)**

When there is information available not only about the order for ranking the values but also about the differences in size between the values, then the scale is known as an *interval scale.* The scale carries with it the property to indicate the distance that separates the values measured. Interval variables are virtually always neasured using numbers, either integers or reals.

An interval scale is temperature measured in Celsius or Fahrenheit. Although, e.g., Celsius scale contains the value of zero, this is strictly thinking arbitrarily set and could be located at some other position for another scale.

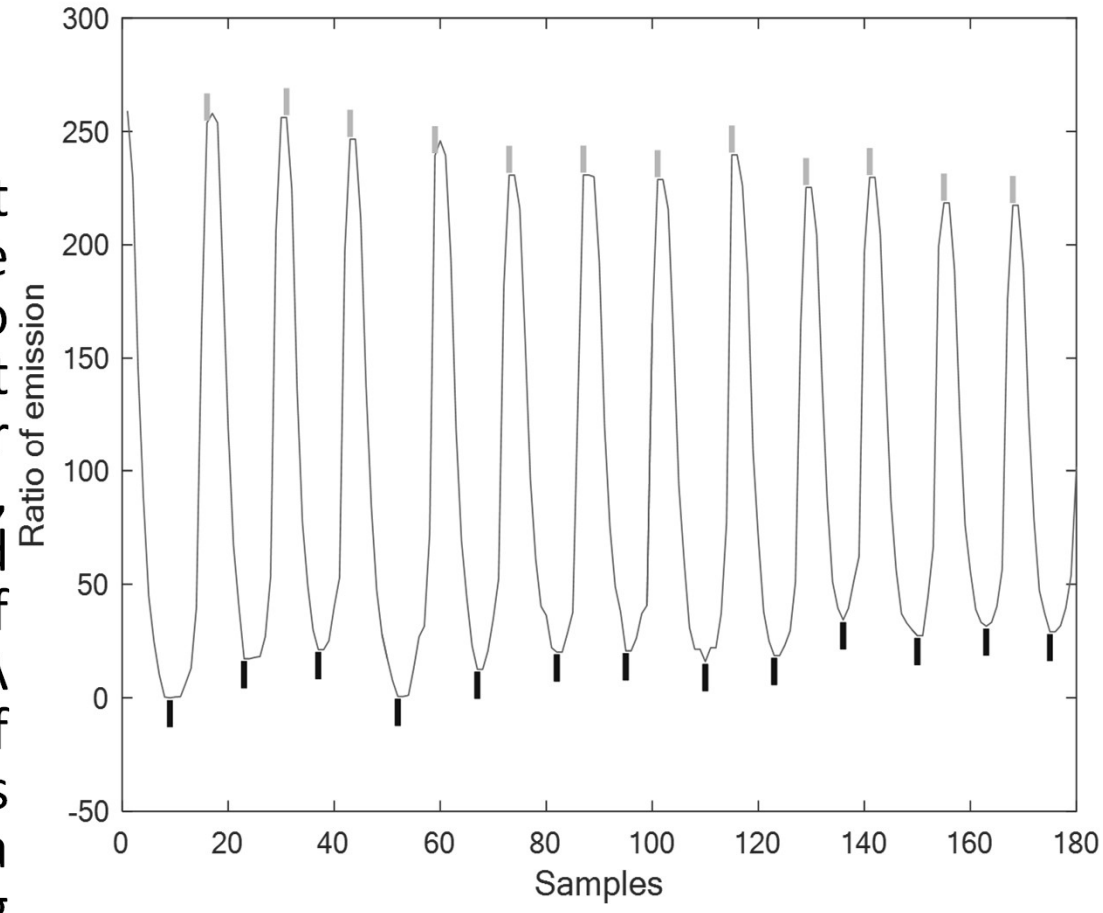
**Ratio scale measurements (variables)**

When there exists a genuine position within a scale in which it starts at a true zero point or the phenomenon disappears, say, temperature in ”absolute” Kelvin scale, weight in kg, length in m or bank balance is 0, it is called *ratio scale*.

For interval scale and ratio scale variables, a *mean* and *standard deviation* (and some other) as well as mode and median can be computed. Now actual differences between values can be computed as distances. Thus, several different distance measures (to be presented later) can be applied that are not possible to use for nominal or categorical values.

Not all ratio scale variables do have a unit like those above. For instance, in Fig. 2. 7 the signal amplitude values are *dimensionless* ratios of two wave lengths of emitted light.

Fig. 2.7. Cardiomyocytes (heart cells) in a laboratory culture basin were exposed to two different wavelengths of light and emissions recorded. For calcium transient analysis, regions of interest were selected from a video stream of spontaneously beating cells. A signal (mean removed) of around 12 s with all peaks recognized normal represents a normal, valid calcium cycling waveform[3].




[3]M. Juhola et al., and K. Aalto-Setälä, Signal analysis and classification methods for the calcium transient data of stem cell-derived cardiomyocytes, Computers in Biology and Medicine, 61, 1-7, 2015

## Nonscalar measurements

Nonscalar measurements need more than one component to capture additional information. Speed or, more physically, velocity requires the direction in addition to magnitude, i.e., two components. Acceleration (or deceleration) is also such. After all, these are not frequently encountered in data mining, and, thus, we do not consider them in this context. Besides, the direction is not always important. For instance, when eye movement signals such as in Fig. 1.2 are used for medical or other purposes, the angular velocity is given by using the magnitude only, since usually the direction is not physiologically an interesting indicator.

# 2.3 Continua of variables

So far we considered the way in which variable values are taken with different types of scale. Now we treat some basic and intrinsic properties of different variable types.

The preceding variable types can also be understood to be categorical or continuous so that the nominal, categorical and ordinal scale masurements are of the former type and the rest are of the latter even if sometimes integers only were used for them. Thus, even if the length were measured in m, it could be measured more accurately, e.g., in mm or µm and expressed with decimals, too.

**The qualitative-quantitative continuum**

This continuum captures the low to high information content of different variable types. Nominal (and categorical scale) variables are qualititative. Any sharp division between the two is not really present. Namely, ordinal variables could be counted in both. Nevertheless, typically it is employed as quantitative.

## The discrete-continuous continuum

Discrete variables are considered to have a limited set of values that they can take on. Continuous variables can take on any value within a range of the minimum and maximum in principle. To see this is a continuum, let us consider the bank account. Technically it is discrete, since we use the cent as the smallest unit. In theory, we could conceive it as continuous, too.

## Single-valued variables (constants)

It may seem odd to discuss a ”variable” as having only a single value. It is strictly speaking a value since it is not varying. Yet, variables that do not vary are also used. For instance, the number of week days is a constant or the number of triangle sides.

Nonetheless, a so-called *dead* variable in a data set that does not change its value does not contribute any information to the modeling process. Since *constant* carry no information in this sense, they can and should be discarded in preprocessing for data mining.

If all values of a variable in a data set are equal to 0, this variable could not distinguish data cases into different classes.

**Two-valued variables**

This is an important type of variable. It is often useful to deploy these *dichotomous* variables. A typical one is gender, ’male’ or ’female’. (In practice there may be three values, the third being ’unknown’.) Note that this type is not seen necessarily as fully equal to binary as described subsequently.

## Missing and empty values

It is frequently the case that there will be variables the values of which are missing, those that are not entered into to the data set. The concept is very important and we will later return to this topic in more detail.

Missing values may appear for the sake of various reasons. For instance, a person has not replied to some question of an inquiry or the measurement equipment has left out some value. These have to somehow be fixed (get them anew if possible or more usually to statistically estimate based on the known values) or be passed by.

An *empty* value in a variable is one for which no real-world value can be supposed (the ”classical” examplar question is whether the patient is pregnant; only she is able to be).

In a database,  missing and empty values are called *nulls*, but these are not a type of masurement. In some context (e.g. Matlab), abbreviation *NaN* (not-a-number) is used.

## Binary values

A special and very important type of dichotomous variable is the *binary* variable, which takes on only values 1 and 0. These values are typically used to indicate if some condition is true or false, or if something did or did not happen, or if some property exists or is absent.

The diffrence between binary and other dischotomous variables is that for the former only it is possible to compute similar operations to those of continuous variables, e.g., mean. This may be seen through that for 0 the property is absent. Instead, for the gender the value is either ’male’ or ’female’.

Binary values are important and useful for various matters, not only as the type of some variables in a data set. They are important for such neural networks as multilayer perceptrons that can use this kind of variable to create probability predictions of the states of outputs, in other words, real values from the interval [0,1].

Further, for nominal variable values can be transformed so that these variables can be used more ”efficiently” in computation. If a variable is the color of a subject’s eyes and there are three different values {blue, brown, green}, these can be encoded with three binary variables so that ’blue’ is equal to 100, ’brown’ equal to 010 and ’green’ to 001. This enables real-valued computation as usual, e.g., for neural network processes.

**Other discrete variables**

All of the other variables, apart from the constants and dichotomous (including binary) variables, take on at least three or more distinct values, say integers from 0 to 100. For example, course examinations are evaluated with scores from 0 to 30. Here the classes are ’rejected’ for {0,..,11} and ’passed’ for {12,..,30}.

**Continuous variables**

Continuous variables, although usually limited as to a minimum and maximum value, can, at least in theory, take on any values within a range. The only practical limit is the accuracy of representation or instrumentation technology.

Most physical variables such as temperature are continuous.

# Example: Vertigo data set

The central variables of Vertigo data set are shown in Table 2.1.[4] There are 1 nominal variable, 11 binary, 10 ordinal and 16 quantitative variables. The only nominal one was ”almost fully ordinal”, because it included four values {none, sudden, progressive, both} of which ’both’ dropping out the first three would form an ordinal variable. At first, we encoded this variable as mentioned above with three binary variables: {none=000, sudden=100, progressive=010, both=001}. Then we observed that there was only one instance having ’both’ from 815 patients. Thereafter, we applied it ”freely” as a single ordinal variable for simplicity. (We could also have left that instance out, but not the whole variable that is among the most important.)

The quantitative variables included integers, e.g, for ’age’ and these were continuous in nature.

[4]M. Siermala et al., Evaluation and classification of otoneurological data with new data analysis methods based on machine learning, Information Sciences, 177, 1963-1976, 2007

Table 2.1. Variables and their Types: B = Binary, N = Nominal, O = Ordinal, and Q = Quantitative; Category Numbers After the Nominal and Ordinal

| [1] patient's age Q | [14] hearing loss type N 4 | [27] caloric asymmetry % Q |
|---|---|---|
| [2] time from symptoms O 7 | [15] severity of tinnitus O 4 | [28] nystagmus to right Q |
| [3] frequency of spells O 6 | [16] time of first tinnitus O 7 | [29] nystagmus to left Q |
| [4] duration of attack O 6 | [17] ear infection B | [30] pursuit eye movement amplitude gain % Q |
| [5] severity of attack O 5 | [18] ear operation B | [31] and its latency (ms) Q |
| [6] rotational vertigo Q | [19] head or ear trauma: noise injury B | [32] audiometry 500 Hz right ear (dB) Q |
| [7] swinging, floating vertigo or unsteady Q | [20] chronic noise exposure B | [33] audiometry 500 Hz left ear (dB) Q |
| [8] Tumarkin-type drop attacks O 4 | [21] head trauma B | [34] audiometry 2 kHz right ear (dB) Q |
| [9] positional vertigo Q | [22] ear trauma B | [35] audiometry 2 kHz left ear (dB) Q |
| [10] unsteadiness outside attacks O 4 | [23] spontaneous nystagmus B | [36] nausea or vomiting O 4 |
| [11] duration of hearing symptoms O 7 | [24] swaying velocity of posturography eyes open (cm/s) Q | [37] fluctuation of hearing B |
| [12] hearing loss of right ear between attacks B | [25] swaying velocity of posturography eyes closed (cm/s) Q | [38] lightheadedness B |
| [13] hearing loss of left ear between attacks B | [26] spontaneous nystagmus (eye movement) velocity (°/s) Q | |