How Accurate Is the Input Data?

Different models can be used to describe real-life phenomena: deterministic, probabilistic, fuzzy, models in which we have interval-valued or fuzzy-valued probabilities, etc. Models are usually not absolutely accurate. It is therefore important to know how accurate is a given model. In other words, it is important to be able to measure a mismatch between the model and the empirical data. In this chapter, we describe an approach of measuring this mismatch which is based on the notion of utility, the central notion of utility theory.

The main results of this chapter first appeared in [206]. In one of the following application chapters (Chapter 35), we show that a similar approach can be used to measure the loss of privacy.

Formulation and Analysis of the Problem, and the Corresponding Results

Models are usually approximate. In most areas of science and engineering, we only have approximate models for the real-world phenomena, i.e., models which are not 100% accurate. Since the models are approximate, their predictions are also only approximate.

It is desirable to gauge the accuracy of a model. In order to understand how accurate are the models’ predictions, we need to know how accurate are the models themselves.

An ideal way to gauge the quality of a model is to compare it with the empirical data, i.e., to validate this model.

Simplest case: deterministic phenomena. Let us start with the simplest situation, when we have a deterministic phenomenon and we have a deterministic model which describes this phenomenon. In this situation, we can simply compare the measured value of the desired quantity with the values predicted by the model.
In such a situation, the difference between the actual and predicted values is a reasonable measure of a mismatch between the real-life phenomenon and the model.

**Real-life situation: non-deterministic phenomena.** In real life, many phenomena are non-deterministic. For such phenomena, we cannot predict the exact values of the corresponding quantities; at best, we can predict the **probabilities** of different values of these quantities.

To validate such models, we must therefore compare the predicted probability distribution with the empirical probability distribution. In such situations, it is not completely clear how we can measure the mismatch between the corresponding probability distributions, i.e., how we can gauge the validity of the probabilistic models.

**Additional complexity and relation to fuzzy techniques.** In practice, the situation is even more complex. Based on a finite sample of real-life events, we cannot uniquely determine the corresponding empirical distribution: we can only provide, with different degrees of confidence, bounds on the corresponding probabilities.

In other words, for each event, instead of a single value of its probability, we get a nested family of confidence intervals corresponding to different levels of uncertainty. Nested families are, in effect, equivalent to fuzzy numbers; see, e.g., [90, 156, 246, 252], so hopefully, techniques for processing fuzzy numbers will be helpful in this comparison.

**What we do in this chapter.** In this chapter, we mainly consider the case of probability distributions. The last section discusses the possibility of extending these results to a more general case of interval-valued probability distributions (p-boxes) and nested (= fuzzy) families of such interval-valued objects.

**Utility approach: a reminder.** In decision making theory, it is proven that under certain reasonable assumptions, a person’s preferences are defined by his or her utility function \( U(x) \) which assigns to each possible outcome \( x \) a real number \( U(x) \) called utility; see, e.g., [150] and [285].

In many real-life situations, a person’s choice \( s \) does not determine the outcome uniquely, we may have different outcomes \( x_1, \ldots, x_n \) with probabilities, correspondingly, \( p_1, \ldots, p_n \). For such a choice, we can describe the utility \( U(s) \) associated with this choice as the expected value of the utility of outcomes:

\[
U(s) = E[U(x)] = p_1 \cdot U(x_1) + \ldots + p_n \cdot U(x_n).
\]

Among several possible choices, a user selects the one for which the utility is the largest: a possible choice \( s \) is preferred to a possible choice \( s' \) (denoted \( s > s' \)) if and only if

\[
U(s) > U(s').
\]

In the general case, when we have a (1-dimensional) probability distribution with the cumulative distribution function (cdf) \( F(x) \), the utility is described by a similar formula