Chapter 2
Fuzziness and Induction

This chapter examines the foundations of IFC by analyzing the concepts of deduction, fuzziness, and induction. The first subsection explains the classical concepts of sharp and deductive logic and classification; in this section, it is presupposed that all terms are clearly defined. The second section explains what happens when those definitions have fuzzy boundaries and provides the tools, fuzzy logic and fuzzy classification, to reason about this. However, there are many terms that do not only lack a sharp boundary of term definition but also lack a priori definitions. Therefore, the third subsection discusses how such definitions can be inferred through inductive logic and how such inferred propositional functions define inductive fuzzy classes. Finally, this chapter proposes a method to derive precise definitions of vague concepts—membership functions—from data. It develops a methodology for membership function induction using normalized likelihood comparisons, which can be applied to fuzzy classification of individuals.

2.1 Deduction

This subsection discusses deductive logic and classification, analyzes the classical as well as the mathematical (Boolean) approaches to propositional logic, and shows their application to classification. Deduction provides a set of tools for reasoning about propositions with a priori truth-values—or inferences of such values. Thus, in the first subsection, the concepts of classical two-valued logic and algebraic Boolean logic are summarized. The second subsection explains how propositional functions imply classes and, thus, provide the mechanism for classification.

2.1.1 Logic

In the words of John Stuart Mill (1843), logic is "the science of reasoning, as well as an art, founded on that science" (p. 18). He points out that the most central entity of logic is the statement, called a *proposition*:

> The answer to every question which it is possible to frame, is contained in a proposition, or assertion. Whatever can be an object of belief, or even of disbelief, must, when put into words, assume the form of a proposition. All truth and all error lie in propositions. What, by a convenient misapplication of an abstract term, we call a truth, is simply a true proposition. (p. 27)

The central role of propositions indicates the importance of linguistics in philosophy. Propositions are evaluated for their truth, and thus, assigned a truth-value because knowledge and insight is based on true statements.

Consider the universe of discourse in logic: The set of possible statements or propositions, $\mathcal{P}$. Logicians believe that there are different levels of truth, usually two (true or false); in the general case, there is a set, $\mathcal{T}$, of possible truth-values that can be assigned to propositions. Thus, the proposition $p \in \mathcal{P}$ is a meaningful piece of information to which a truth-value, $\tau(p) \in \mathcal{T}$, can be assigned. The corresponding mapping of $\tau : \mathcal{P} \to \mathcal{T}$ from propositions $\mathcal{P}$ to truth-values $\mathcal{T}$ is called a *truth function*.

In general logic, operators can be applied to propositions. A unary operator, $O_1 : \mathcal{P}_1 \to \mathcal{T}$, maps a single proposition into a set of transformed truth-values. Accordingly, a binary operator, $O_2 : \mathcal{P}_1 \times \mathcal{P}_2 \to \mathcal{T}$, assigns a truth-value to a combination of two propositions, and an *n*-ary operator, $O_n : \mathcal{P}_1 \times \cdots \times \mathcal{P}_n \to \mathcal{T}$, is a mapping of a combination of *n* propositions to a new truth-value.

The logic of two-valued propositions is the science and art of reasoning about statements that can be either true or false. In the case of two-valued logic, or *classical logic* (CL), the set of possible truth values, $\mathcal{T}^{CL} := \{true, false\}$, contains only two elements, which partitions the class of imaginable propositions $\mathcal{P}$ into exactly two subclasses: the class of false propositions and the class of true ones.

With two truth-values, there are four (2^2) possible unary logical operators; however, there is only one possible non-trivial unary operator other than identity, truth, and falsehood: A proposition, $p \in \mathcal{P}$, can be *negated* (not p), which inverts the truth-value of the original proposition. Accordingly, for a combination of two propositions, p and q, each with two truth-values, there are 16 (2^{2^2}) possible binary operators. The most common binary logical operators are disjunction, conjunction, implication, and equivalence: A *conjunction* of two propositions, p and q, is true if both propositions are true. A *disjunction* of two propositions, p or q, is true if one of the propositions is true. An *implication* of q by p is true if, whenever p is true, q is true as well. An *equivalence* of two propositions is true if p implies q and q implies p.

Classical logic is often formalized in the form of a propositional calculus. The syntax of classical propositional calculus is described by the concept of variables, unary and binary operators, formulae, and truth functions. Every proposition is represented by a variable (e.g., p); every proposition and every negation of a

proposition is a term; every combination of terms by logical operators is a formula; terms and formulae are themselves propositions; negation of the proposition p is represented by $\neg p$; conjunction of the two propositions p and q is represented by $p \wedge q$; disjunction of the two propositions p and q is represented by $p \vee q$; implication of the proposition q by the proposition p is represented by $p \Rightarrow q$; equivalence between the two propositions p and q is represented by $p \equiv q$; and there is a truth function, $\tau^{CL} : \mathcal{P} \to \mathcal{T}^{CL}$, mapping from the set of propositions p into the set of truth values $\mathcal{T}$. The semantics of propositional calculus are defined by the values of the truth function, as formalized in *Formula* (2.1) through *Formula* (2.5).

$$\tau^{CL}(\neg p) := \begin{cases} \text{if}(\tau(p) = \text{true}) & \text{false} \\ \text{else} & \text{true.} \end{cases} \tag{2.1}$$

$$\tau^{CL}(p \wedge q) := \begin{cases} \text{if}\big(\tau(p) = \tau(q) = \text{true}\big) & \text{true} \\ \text{else} & \text{false.} \end{cases} \tag{2.2}$$

$$\tau^{CL}(p \vee q) := \begin{cases} \text{if}\big(\tau(p) = \tau(q) = \text{false}\big) & \text{false} \\ \text{else} & \text{true.} \end{cases} \tag{2.3}$$

$$\tau^{CL}(p \Rightarrow q) := \tau^{CL}(\neg p \vee q) \tag{2.4}$$

$$\tau^{CL}(p \equiv q) := \tau^{CL}(p \Rightarrow q \wedge q \Rightarrow p) \tag{2.5}$$

George Boole (1847) realized that logic can be calculated using the numbers 0 and 1 as truth values. His conclusion was that logic is mathematical in nature:

> I am then compelled to assert, that according to this view of the nature of Philosophy, Logic forms no part of it. On the principle of a true classification, we ought no longer to associate Logic and Metaphysics, but Logic and Mathematics. (p. 13)

In Boole's mathematical definition of logic, the numbers 1 and 0 represents the truth-values and logical connectives are derived from arithmetic operations: subtraction from1 as negation and multiplication as conjunction. All other operators can be derived from these two operators through application of the laws of logical equivalence. Thus, in Boolean logic (*BL*), the corresponding propositional calculus is called *Boolean algebra*, stressing the conceptual switch from metaphysics to mathematics. Its syntax is defined in the same way as that of *CL*, except that the Boolean truth function, $\tau^{BL} : \mathcal{P} \to \mathcal{T}^{BL}$, maps from the set of propositions into the set of Boolean truth values, $\mathcal{T}^{BL} := \{0, 1\}$, that is, the set of the two numbers 0 and 1.

The Boolean truth function τ^{BL} defines the semantics of Boolean algebra. It is calculated using multiplication as conjunction and subtraction from 1 as negation, as formalized in *Formula* (2.6) through *Formula* (2.8). Implication and equivalence can be derived from negation and disjunction in the same way as in classical propositional calculus.

$$\tau^{BL}(\neg p) := 1 - \tau^{BL}(p) \tag{2.6}$$

$$\tau^{BL}(p \wedge q) := \tau^{BL}(p) \cdot \tau^{BL}(q) \tag{2.7}$$

$$\begin{aligned}
\tau^{BL}(p \vee q) &:= \neg\big(\neg p \wedge \neg q\big) \\
&= 1 - \big(1 - \tau^{BL}(p)\big) \cdot \big(1 - \tau^{BL}(q)\big)
\end{aligned} \tag{2.8}$$

2.1.2   Classification

Class logic, as defined by Glubrecht, Oberschelp, and Todt (1983), is a logical system that supports statements applying a *classification operator*. Classes of objects can be defined according to logical propositional functions. According to Oberschelp (1994), a class, $C = \{i \in U \mid \Pi(i)\}$, is defined as a collection of individuals, i, from a universe of discourse, U, satisfying a propositional function, Π, called the *classification predicate*. The domain of the classification operator, $\{.\,|.\} : \mathbb{P} \to U^{*}$, is the class of propositional functions $\mathbb{P}$ and its range is the powerclass of the universe of discourse U^{*}, which is the class of possible subclasses of U. In other words, the class operator assigns subsets of the universe of discourse to propositional functions. A universe of discourse is the set of all possible individuals considered, and an individual is a real object of reference. In the words of Bertrand Russell (1919), a propositional function is "an expression containing one or more undetermined constituents, such that, when values are assigned to these constituents, the expression becomes a proposition" (p. 155).

In contrast, *classification* is the process of grouping individuals who satisfy the same predicate into a class. A (Boolean) classification corresponds to a membership function, $\mu_C : U \to \{0, 1\}$, which indicates with a Boolean truth-value whether an individual is a member of a class, given the individual's classification predicate. As shown by *Formula* (2.9), the membership μ of individual i in class $C = \{i \in U \mid \Pi(i)\}$ is defined by the truth-value τ of the classification predicate $\Pi(i)$. In Boolean logic, the truth-values are assumed to be certain. Therefore, classification is *sharp* because the truth values are either exactly 0 or exactly 1.

$$\mu_C(i) := \tau(\Pi(i)) \in \{0, 1\} \tag{2.9}$$

Usually, the classification predicate that defines classes refers to attributes of individuals. For example, the class "tall people" is defined by the predicate "tall," which refers to the attribute "height." An attribute, X, is a function that characterizes individuals by mapping from the universe of discourse U to the set of possible characteristics χ (*Formula* 2.10).

$$X : U \to \chi \tag{2.10}$$

There are different types of values encoding characteristics. *Categorical* attributes have a discrete range of symbolic values. *Numerical* attributes have a range of

numbers, which can be natural or real. *Boolean* attributes have Boolean truth-values $\{0, 1\}$ as a range. *Ordinal* attributes have a range of categories that can be ordered.

On one hand, the distinction between univariate and multivariate classification, the *variety*, depends on the number of attributes considered for the classification predicate. The *dimensionality* of the classification, on the other hand, depends on the number of dimensions, or linearly independent attributes, of the classification predicate domain.

In a *univariate classification* (*UC*), the classification predicate Π refers to one attribute, X, which is true for an individual, i, if the feature $X(i)$ equals a certain characteristic, $c \in \chi$.

$$\mu_{UC}(i) := \tau^{BL}(X(i) = c) \tag{2.11}$$

In a *multivariate classification* (*MVC*), the classification predicate refers to multiple element attributes. The classification predicate is true for an individual, i, if an aggregation, a, of several characteristic constraints has a given value, $c \in \chi$.

$$\mu_{MVC}(i) := \tau^{BL}(a(X_1(i), \cdots, X_n(i)) = c) \tag{2.12}$$

A *multidimensional classification* (*MDC*) is a special case of a multivariate classification that refers to n-tuples of attributes, such that the resulting class is functionally dependent on the combination of all n attributes.

$$\mu_{MDC}(i) := \tau^{BL}\left(\begin{bmatrix} X_1(i) \\ \vdots \\ X_n(i) \end{bmatrix} = \begin{bmatrix} C_1 \\ \vdots \\ C_m \end{bmatrix}\right) \tag{2.13}$$

This distinction between multivariate and multidimensional classification is necessary for the construction of classification functions. Multivariate classifications can be derived as functional aggregates of one-dimensional membership functions, in which the influence of one attribute to the resulting aggregate does not depend on the other attributes. In contrast, in multidimensional classification, the combination of all attributes determines the membership value, and thus, one attribute has different influences on the membership degree for different combinations with other attribute values. Therefore, multidimensional classifications need multidimensional membership functions that are defined on n-tuples of possible characteristics.

2.2 Fuzziness

> There are many misconceptions about fuzzy logic. Fuzzy logic is not fuzzy. Basically,
> fuzzy logic is a precise logic of imprecision and approximate reasoning. (Zadeh, 2008,
> p. 2051)

Fuzziness, or vagueness (Sorensen, 2008), is an uncertainty regarding concept boundaries. In contrast to ambiguous terms, which have several meanings, vague terms have one meaning, but the extent of it is not sharply distinguishable. For example, the word *tall* can be ambiguous, because a tall cat is usually smaller than a small horse. Nevertheless, the disambiguated predicate "tall for a cat" is vague, because its linguistic concept does not imply a sharp border between tall and small cats.

Our brains seem to love boundaries. Perhaps, making sharp distinctions quickly was a key cognitive ability in evolution. Our brains are so good at recognizing limits, that they construct limits where there are none. This is what many optical illusions are based on: for example, Kaniza's (1976) Illusory Square (Fig. 2.1).

An ancient symbol of sharp distinction between classes is the yin and yang symbol (Fig. 2.2). It symbolizes a dualistic worldview—the cosmos divided into light and dark, day and night, and so on.

Nevertheless, in reality, the transition between light and dark is gradual during the 24 h of a day. This idea of gradation of our perceptions can be visualized by a fuzzy yin and yang symbol (Fig. 2.3). Sorensen (2008) explains that many-valued logics have been proposed to solve the philosophical implications of vagueness. One many-valued approach to logic is fuzzy logic, which allows infinite truth-values in the interval between 0 and 1.

In the next section, introducing membership functions, fuzzy sets, and fuzzy propositions are discussed; these are the bases for fuzzy logic, which in fact, is a precise logic for fuzziness. Additionally, it is shown how fuzzy classifications are derived from fuzzy propositional functions.

2.2.1 Fuzzy Logic

Lotfi Zadeh (2008) said, "Fuzzy logic is not fuzzy" (p. 2751). Indeed, it is a precise mathematical concept for reasoning about fuzzy (vague) concepts. If the domain of those concepts is ordinal, membership can be distinguished by its *degree*. In classical set theory, an individual, i, of a universe of discourse, U, is either completely a member of a set or not at all. As previously explained, according to Boolean logic, the membership function $\mu_S : U \rightarrow \{0, 1\}$, for a crisp set S, maps from individuals to sharp truth-values. As illustrated in Fig. 2.4, a sharp set (the big dark circle) has a clear boundary, and individuals (the small bright circles) are either a member of it or not. However, one individual is not entirely covered by the big dark circle, but is also not outside of it.

Fig. 2.1 There is no square. Adapted from "Subjective Contours" by G. Kaniza, 1976, Copyright 1976 by Scientific American, Inc
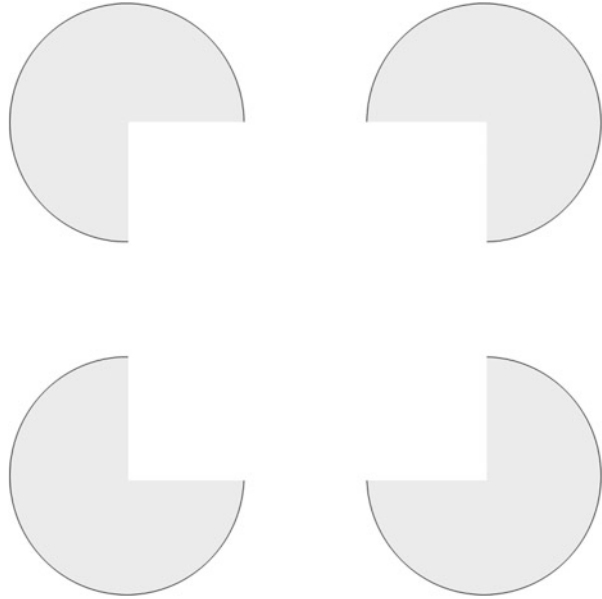
In contrast, a set is called *fuzzy* by Zadeh (1965) if individuals can have a gradual degree of membership to it. In a fuzzy set, as shown by Fig. 2.5, the limits of the set are blurred. The degree of membership of the elements in the set is gradual, illustrated by the fuzzy gray edge of the dark circle. The membership function, $\mu_F : U \rightarrow [0, 1]$, for a fuzzy set, F, indicates the degree to which individual i is a member of F in the interval between 0 and 1. In Fig. 2.4, the degree of membership of the small circles i is defined by a normalization n of their distance d from the center c of the big dark circle b, $\mu_b(i) = n(d(i, c))$. In the same way as in classical set theory, set operators can construct complements of sets and combine two sets by union and intersection. Those operators are defined by the fuzzy membership function. In the original proposal of Zadeh (1965), the set operators are defined by subtraction from 1, minimum and maximum. The complement, $\overline{F}$, of a fuzzy set, F, is derived by subtracting its membership function from 1; the union of two sets, $F \cup G$, is derived from the maximum of the membership degrees; and the intersection of two sets, $F \cap G$, is derived from the minimum of the membership degrees.

Accordingly, fuzzy subsets and fuzzy power sets can be constructed. Consider the two fuzzy sets A and B on the universe of discourse U. In general, A is a fuzzy subset of B if the membership degrees of all its elements are smaller or equal to the membership degrees of elements in B (Formula 2.14). Thus, a fuzzy power set, $\widetilde{B^*}$, of a (potentially fuzzy) set B is the class of all its fuzzy subsets (Formula 2.15).

Fig. 2.2 Black or white:
conventional yin and yang
symbol with a sharp
distinction between
opposites, representing
metaphysical dualism.
Adapted from http://www.
texample.net/tikz/
examples/yin-and-yang/
(accessed 02.2012) with
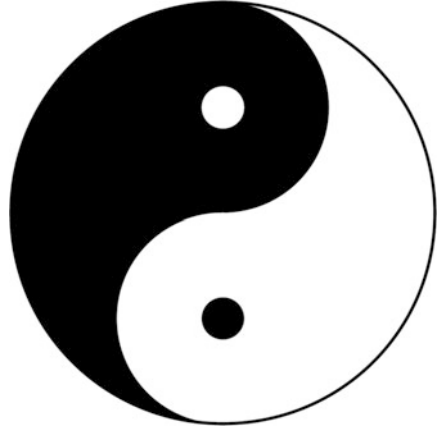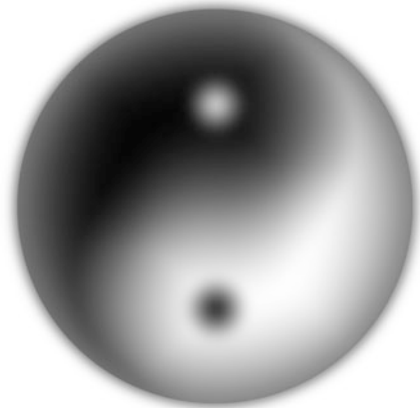permission (creative
commons license CC BY
2.5)

Fig. 2.3 Shades of *grey*:
fuzzy yin and yang symbol
with a gradation between
opposites, representing
metaphysical monism

$$A \widetilde{\subseteq} B := \forall x \in U : \mu_A(x) \le \mu_B(x) \tag{2.14}$$

$$B^{\widetilde{*}} := \left\{ A \widetilde{\subseteq} U \mid A \widetilde{\subseteq} B \right\} \tag{2.15}$$

With the tool of fuzzy set theory in hand, the sorites paradox cited in the introduction (Chap. 1) can be tackled in a much more satisfying manner. A heap of wheat grains can be defined as a fuzzy subset, *Heap* $\widetilde{\subseteq}$ $\mathbb{N}$, of natural numbers $\mathbb{N}$ of wheat grains. A heap is defined in the English language as "a great number or large quantity" (merriam-webster.com, 2012b). For instance, one could agree that 1,000 grains of wheat is a large quantity, and between 1 and 1,000, the "heapness" of a grain collection grows logarithmically. Thus, the membership function of the number of grains $n \in \mathbb{N}$ in the fuzzy set *Heap* can be defined according to Formula (2.16). The resulting membership function is plotted in Fig. 2.6.

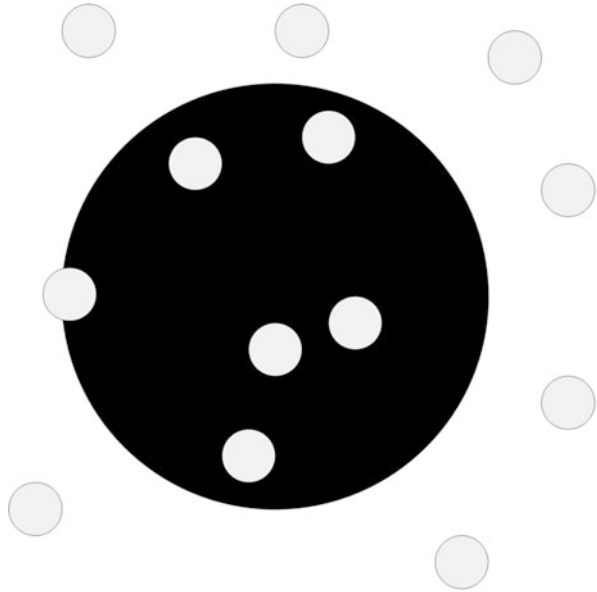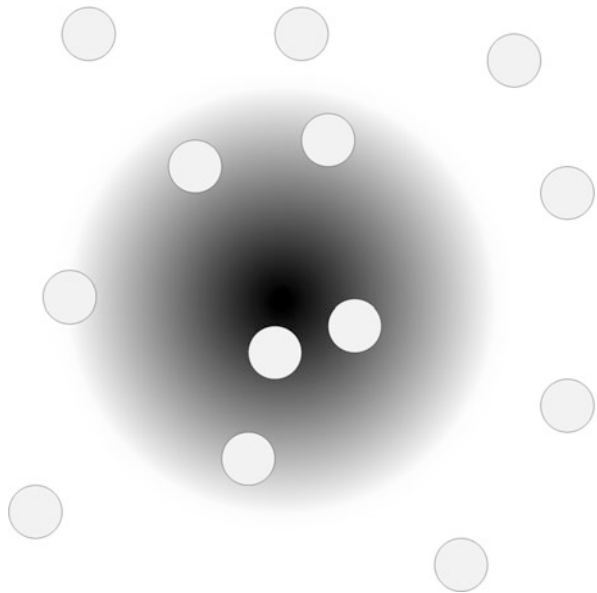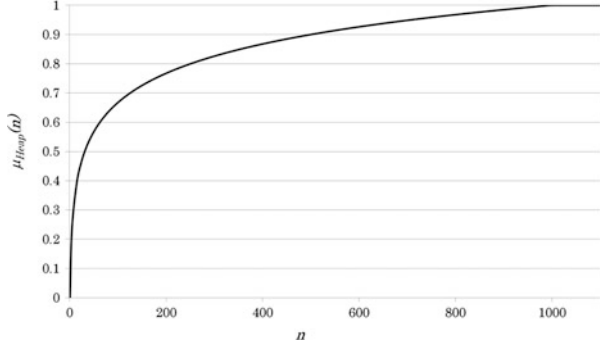**Fig. 2.4** A visualization of a classical set with sharp boundaries

Fig. 2.5 A visualization of a fuzzy set

$$\mu_{Heap}(n) := \begin{cases} 0 \text{ if } n = 0 \\ 1 \text{ if } n > 1000 \\ 0.1448\ln(n) \text{ else.} \end{cases} \qquad (2.16)$$

Based on the concept of fuzzy sets, Zadeh (1975a) derived *fuzzy propositions* (*FP*) for approximate reasoning: A fuzzy proposition has the form "*x is L*," where

Fig. 2.6 Fuzzy set theory applied to the sorites paradox

x is an individual of a universe of discourse U and L is a *linguistic term*, defined as a fuzzy set on U. As stated by Formula (2.17), the truth-value τ^{FL} of a fuzzy proposition is defined by the degree of membership μ_L of x in the linguistic term L.

$$\tau^{FP}(x \text{ is } L) := \mu_L(x) \tag{2.17}$$

If A is an attribute of x, a fuzzy proposition can also refer to the corresponding attribute value, such as $x \text{ is } L := A(x) \text{ is } L$. The fuzzy set L on U is equivalent to the fuzzy set L on the domain of the attribute, or dom(A). In fact, the set can be defined on arbitrarily deep-nested attribute hierarchies concerning the individual. As an example, let us look at the fuzzy proposition, "Mary is blond." In this sentence, the linguistic term "blond" is a fuzzy set on the set of people, which is equivalent to a fuzzy set *blond* on the color of people's hair (Formula 2.18).

$$\tau^{FP}(\text{"Mary is blond"}) = \mu_{blond}(\text{Mary}) \equiv \mu_{blond}(\text{color(hair(Mary)))} \tag{2.18}$$

Fuzzy propositions (*FP*) can be combined to construct fuzzy formulae using the usual logic operators *not* ($\neg$), *and* ($\wedge$), and *or* ($\vee$), for which the semantics are defined by the fuzzy truth function $\tau^{FP} : \mathcal{F} \rightarrow [0, 1]$, mapping from the class of fuzzy propositions $\mathcal{F}$ into the set of Zadehan truth values in the interval between 1 and 0. Let "$x \text{ is } P$" and "$x \text{ is } Q$" be two fuzzy propositions on the same individual. Then their combination to fuzzy formulae is defined as follows (Formula 2.19 through Formula 2.21): negation by the inverse of the corresponding fuzzy set, conjunction by intersection of the corresponding fuzzy sets, and disjunction by union of the corresponding fuzzy sets.

$$\tau^{FP}(\neg(x \text{ is } P)) := \mu_{\overline{P}}(x); \tag{2.19}$$

$$\tau^{FP}(x \text{ is } P \wedge x \text{ is } Q) := \mu_{P \cap Q}(x); \tag{2.20}$$

$$\tau^{FP}(x \text{ is } P \vee x \text{ is } Q) := \mu_{P \cup Q}(x) \tag{2.21}$$

Zadeh's fuzzy propositions are derived from statements of the form "X is Y." They are based on the representation operator is : $U \times \widetilde{U^*} \rightarrow \mathcal{F}$ mapping from the

universe U of discourse and its fuzzy powerset $\widetilde{U^*}$ to the class of fuzzy propositions $\mathcal{F}$. Consequently, fuzzy propositions in the sense of Zadeh are limited to statements about degrees of membership in a fuzzy set.

Generally, logic with fuzzy propositions—or more precisely, a propositional logic with Zadehan truth values in the interval between 0 and 1, a *"Zadehan Logic"* (*ZL*)—can be viewed as a generalization of Boole's mathematical analysis of logic to a gradual concept of truth. In that sense, *ZL* is a simple generalization of Boolean logic (*BL*), in which the truth value of any proposition is not only represented by numbers, but also can be anywhere in the interval between 0 and 1.

According to the *Stanford Encyclopedia of Philosophy* (Hajek, 2006), fuzzy logic, in the narrow sense, is a "symbolic logic with a comparative notion of truth developed fully in the spirit of classical logic" ("Fuzzy Logic," paragraph3). If *ZL* is viewed as a generalization of *BL*, fuzzy propositions of the form "*X is Y*" are a special case, and propositions and propositional functions of any form can have gradual values of truth. Accordingly, *ZL* is defined by the truth function $\tau^{ZL} : \mathcal{P} \to \mathcal{T}^{ZL}$ mapping from the class of propositions $\mathcal{P}$ to the set of Zadehan truth-values $\mathcal{T}^{ZL} = [0, 1]$. Consequently, fuzzy set membership is a special case of fuzzy proposition, and the degree of membership of individual x in another individual y can be defined as the value of truth of the fuzzy proposition $x \in y$ (Formula 2.22).

$$\mu_y(x) := \tau^{ZL}(x \in y) \tag{2.22}$$

The Zadehan truth function τ^{ZL} defines the semantics of *ZL*. As in Boolean algebra, its operators can be defined by subtraction from 1 as negation, and multiplication as conjunction, as formalized in Formula (2.23) and Formula (2.24). Disjunction, implication, and equivalence can be derived from negation and conjunction in the same way as in Boolean logic.

$$\tau^{ZL}(\neg p) := 1 - \tau^{ZL}(p) \tag{2.23}$$
$$\tau^{ZL}(p \wedge q) := \tau^{ZL}(p) \cdot \tau^{ZL}(q) \tag{2.24}$$

In that light, any proposition with an uncertain truth-value smaller than 1 or greater than 0 is a fuzzy proposition. Additionally, every function with the range [0,1] can be thought of as a truth function for a propositional function. For example, statistical likelihood $L(y|x)$ can be seen as a truth function for the propositional function, "*y* is likely if *x*," as a function of x. This idea is the basis for IFC proposed in the next section. The usefulness of this generalization is shown in the chapter on applications, in which fuzzy propositions such as "customers with characteristic *X* are likely to buy product *Y*" are assigned truth-values that are computed using quantitative prediction modeling.

2.2.2 Fuzzy Classification

A *fuzzy class,* $\widetilde{C} := \sim \{ i \in U \mid \widetilde{\Pi}(i) \}$, is defined as a fuzzy set $\widetilde{C}$ of individuals i, whose membership degree is defined by the Zadehan truth-value of the proposition $\widetilde{\Pi}(i)$. The classification predicate, $\widetilde{\Pi}$, is a propositional function interpreted in *ZL*. The domain of the *fuzzy class operator,* $\sim \{.|.\} : \mathbb{P} \to U^{\widetilde{*}}$, is the class of propositional functions, $\mathbb{P}$, and the range is the fuzzy power set, $U^{\widetilde{*}}$ (the set of fuzzy subsets) of the universe of discourse, U. In other words, the fuzzy class operator assigns fuzzy subsets of the universe of discourse to propositional functions.

Fuzzy classification is the process of assigning individuals a membership degree to a fuzzy set, based on their degrees of truth of the classification predicate. It has been discussed, for example, by Zimmermann (1997), Del Amo et al. (1999), and Meier et al. (2008). A fuzzy classification is achieved by a membership function, $\mu_{\widetilde{C}} : U \to [0, 1]$, that indicates the degree to which an individual is a member of a fuzzy class, $\widetilde{C}$, given the corresponding fuzzy propositional function, $\widetilde{\Pi}$. This membership degree is defined by the Zadehan truth-value of the corresponding proposition, $\widetilde{\Pi}(i)$, as formalized in Formula (2.25).

$$\mu_{\widetilde{C}}(i) := \tau^{ZL}\big(\widetilde{\Pi}(i)\big) \tag{2.25}$$

In the same way as in crisp classification, the fuzzy classification predicate refers to attributes of individuals. Additionally, Zadehan logic introduces two new types of characteristics. *Zadehan attributes* have a range of truth values represented by $\mathcal{T}^{ZL} := [0, 1]$. Linguistic attributes have a range of linguistic terms (fuzzy sets) together with the Zadehan truth-value of membership in those terms (Zadeh, 1975b).

In a *univariate fuzzy classification* (*UF*), the fuzzy classification predicate $\widetilde{\Pi}$ refers to one attribute, X, and it corresponds to the membership degree of the attribute characteristic $X(i)$ in a given fuzzy restriction (Zadeh, 1975a), $R \in \chi^{\widetilde{*}}$, which is a fuzzy subset of possible characteristics χ (Formula 2.26).

$$\mu_{UF}(i) := \tau^{ZL}(X(i) \text{ is } R) \tag{2.26}$$

In a *multivariate fuzzy classification* (*MVF*), $\widetilde{\Pi}$ refers to multiple attributes. The truth function of the classification predicate for an individual, i, equals to an aggregation, a, of several fuzzy restrictions of multiple attribute characteristics, $X_j(i)$, $j = 1 \ldots n$ (Formula 2.27).

$$\mu_{MVF}(i) := a\big(\tau^{ZL}(X_1(\text{i}) \text{ is } R_1, \ldots, X_n(\text{i}) \text{ is } R_n)\big) \tag{2.27}$$

In a *multidimensional fuzzy classification* (*MDF*), $\widetilde{\Pi}$ refers to n-tuples of functionally independent attributes. The membership degree of individuals in a

multidimensional class is based on an n-dimensional fuzzy restriction, R^n (Formula 2.28), which is a multidimensional fuzzy set on the Cartesian product of the attribute ranges with a multidimensional membership function of $\mu_{MDF} : range\ (X_1) \times \ldots \times range\ (X_n) \rightarrow [0, 1]$.

$$\mu_{MDF}(i) := \tau^{ZL} \left(\begin{bmatrix} X_1(i) \\ \vdots \\ X_n(i) \end{bmatrix} \text{ is } R^n \right) \tag{2.28}$$

2.3   Induction

Given a set of certainly true statements, deduction works fine. The problem is that the only certainty philosophy can offer is Descartes' "I think therefore I am" proposition; however, postmodern philosophers are not so sure about the *I* anymore (Precht, 2007, p. 62 ff). Therefore, one should be given a tool to reason under uncertainty, and this tool is induction. In this chapter, inductive logic is analyzed, the application of induction to fuzzy classification is discussed, and a methodology for membership function induction using normalized ratios and differences of empirical conditional probabilities and likelihoods is proposed.

2.3.1   Inductive Logic

Traditionally, induction is defined as drawing general conclusions from particular observations. Contemporary philosophy has shifted to a different view because, not only are there inductions that lead to *particular* conclusions, but also there are deductions that lead to *general* conclusions. According to Vickers (2009) in the *Stanford Encyclopedia of Philosophy* (SEP), it is agreed that induction is a form of inference that is contingent and ampliative ("The contemporary notion of induction", paragraph 3), in contrast to deductive inference, which is necessary and explicative. Induction is contingent, because inductively inferred propositions are not necessarily true in all cases. And it is ampliative because, in Vickers words, "induction can amplify and generalize our experience, broaden and deepen our empirical knowledge"("The contemporary notion of induction", paragraph 3). In another essay in the SEP, inductive logic is defined as "a system of evidential support that extends deductive logic to less-than-certain inferences" (Hawthorne, 2008, "Inductive Logic," paragraph 1). Hawthorne admits that there is a degree of fuzziness in induction: In an inductive inference, "the premises should *provide some degree of support* for the conclusion" ("Inductive Logic," para. 1). The degree of support for an inductive inference can thus be viewed as a fuzzy restriction of possible inferences, in the sense of Zadeh (1975a). Vickers (2009) explains that the problem of induction is two-fold: The epistemic problem is to define a method to distinguish appropriate from inappropriate inductive inference. The metaphysical

problem is to explain in what substance the difference between reliable and unreliable induction actually exists.

Epistemologically, the question of induction is to find a suitable method to infer propositions under uncertainty. State of the art methods rely on empirical probabilities or likelihoods. There are many interpretations of probability (Hájek, 2009). For the context of this thesis, one may agree that a mathematical probability, $P(A)$ numerically represents how probable it is that a specific proposition A is true: $P(A) \equiv \tau^{ZL}("\ A\ is\ probable\ ")$; and that the disjunction of all possible propositions, the probability space Ω, is certain, i.e., $P(\Omega) = 1$.

In practice, probabilities can be estimated by relative frequencies, or sampled empirical probabilities p in a sample of n observations, defined by the ratio between the number of observations, i, in which the proposition A_i is true, and the total number of observations (Formula 2.29).

$$P(A) \approx p(A) := \frac{\sum_{i=1}^{n} \tau(A_i)}{n} \tag{2.29}$$

A conditional probability (Weisstein, 2010a) is the probability for an outcome x, given that y is the case, as formalized in Formula (2.30).

$$P(x \mid y) = \frac{P(x \wedge y)}{P(y)} \tag{2.30}$$

Empirical sampled conditional probabilities can be applied to compute likelihoods. According to James Joyce, "in an unfortunate, but now unavoidable, choice of terminology, statisticians refer to the inverse probability $P_H(E)$ as the 'likelihood' of H on E" (Joyce, 2003, "Conditional Probabilities and Bayes' Theorem," paragraph 5). The likelihood of the hypothesis H is an estimate of how probable the evidence or known data E is, given that the hypothesis is true. Such a probability is called a "posterior probability" (Hawthorne, 2008, "inductive Logic," paragraph 5), that is, a probability after measurement, shown by Formula (2.31).

$$L(H \mid E) := p(E \mid H) \tag{2.31}$$

In the sense of Hawthorne (2008), the general law of likelihood states that, for a pair of incompatible hypotheses H_1 and H_2, the evidence E supports H_1 over H_2, if and only if $p(E|H_1) > p(E|H_2)$ The likelihood ratio (LR) measures the strength of evidence for H_1 over H_2 (Formula 2.32). Thus, the "likelihoodist" (sic; Hawthorne, 2008, "Likelihood Ratios, Likelihoodism, and the Law of Likelihood," paragraph 5) solution to the epistemological problem of induction is the likelihood ratio as measure of support for inductive inference.

$$LR(H_1 > H_2 \mid E) := \frac{L(H_2 \mid E)}{L(H_1 \mid E)} \tag{2.32}$$

According to Hawthorne, the prior probability of a hypothesis, $p_0(H)$, that is, an estimated probability prior to measurement of evidence E, plays an important role for inductive reasoning. Accordingly, Bayes' theorem can be interpreted and rewritten using measured posterior likelihood and prior probability in order to apply it to the evaluation of scientific hypotheses. According to Hawthorne (2008), the posterior probability of hypothesis H conditional to evidence E is equal to the product of the posterior likelihood of H given E and the prior probability of H, divided by the (measured) probability of E (Formula 2.33).

$$p(H \mid E) = \frac{L(H \mid E) \cdot p_0(H)}{p(E)} \tag{2.33}$$

What if there is fuzziness in the data, in the features of observations, or in the theories? How is likelihood measured when the hypothesis or the evidence is fuzzy? If this fuzziness is ordinal, that is, if the extent of membership in the fuzzy terms can be ordered, a membership function can be defined, and an empirical probability of fuzzy events can be calculated. Analogous to Dubois and Prade (1980), a fuzzy event $\widetilde{A}$ in a universe of discourse U is a fuzzy set on U with a membership function $\mu_{\widetilde{A}} : U \to [0, 1]$. For categorical elements of U, the estimated probability after n observations is defined as the average degree of membership of observations i in $\widetilde{A}$, as formalized in Formula (2.34).

$$P\left(\widetilde{A}\right) \approx p\left(\widetilde{A}\right) = \frac{\sum_{i=1}^{n} \mu_{\widetilde{A}}(i)}{n} \tag{2.34}$$

By application of Formula (2.34) to Formula (2.31), the likelihood of ordinal fuzzy hypothesis $\widetilde{H}$, given ordinal fuzzy evidence $\widetilde{E}$, can be defined as a conditional probability of fuzzy events, as shown in Formula (2.35).

$$L\left(\widetilde{H} \mid \widetilde{E}\right) = p\left(\widetilde{E} \mid \widetilde{H}\right) = \frac{\sum_{i=1}^{n} \mu_{\widetilde{H} \cap \widetilde{E}}(i)}{\sum_{i=1}^{n} \mu_{\widetilde{H}}(i)} \tag{2.35}$$

The question of the metaphysical problem of induction is: what is the substance of induction? In what kind of material does the difference between reliable and unreliable inductive inference exist? The importance of this question cannot be underestimated, since reliable induction enables prediction. A possible answer could be that the substance of an induction is the amount of information contained in the inference. This answer presupposes that information is a realist category, as suggested by Chmielecki (1998). According to Shannon's information theory

(Shannon, 1948), the information contained in evidence x about hypothesis y is equal to the difference between the uncertainty (entropy), $H(y)$, about the hypothesis y and the resulting uncertainty, $H_x(y)$, after observation of the evidence x, $I(x, y) = H(y) - H_x(y) = \Sigma_x\Sigma_y p(x \wedge y)\log_2 p(x \wedge y)/(p(x)p(y))$. Shannon's quantity of information is defined in terms of joint probabilities. However, by application of Shannon's theory, the metaphysical problem of induction is transferred to a metaphysical problem of probabilities because, according to Shannon, the basic substance of information is the probability of two signals occurring simultaneously compared to the probability of occurring individually. (One could link this solution to the concept of quantum physical particle probability waves [Greene, 2011], but this would go beyond the scope of this thesis and would be highly speculative; therefore, this link is not explored here. Suffice it to state that probability apparently is a fundamental construct of matter and waves as well as of information and induction.)

2.3.2 Inductive Classification

Inductive classification is the process of assigning individuals to a set based on a classification predicate derived by an inductive inference. Inductive classification can be automated as a form of supervised machine learning (Witten & Frank, 2005): a class of processes (algorithms or heuristics) that learn from examples to decide whether an individual, i, belongs to a given class, y, based on its attributes. Generally, supervised machine learning processes induce a model from a dataset, which generalizes associations in the data in order to provide support for inductive inference. This model can be used for predicting the class membership of new data elements. Induced classification models, called *classifiers*, are first trained using a training set with known class membership. Then, they are applied to a test or prediction set in order to derive class membership predictions. Examples of classification learning algorithms that result in classifications are decision trees, classification rules, and association rules. In those cases, the model consists of logical formulae of attribute values, which predict a crisp class value.

Data are signs (signals) that represent knowledge such as numbers, characters, or bits. The basis for automated data analysis is a systematic collection of data on individuals. The most frequently used data structure for analytics is the matrix, in which every individual, i (a customer, a transaction, a website, etc.), is represented by a row, and every attribute, X_k, is represented by a column. Every characteristic, $X_k(i)$, of individual i for attribute X_k is represented by one scalar value within the matrix.

A training dataset d is an $m \times (n + 1)$ matrix with m rows, n columns for $X_1, \ldots, X_n$ and a column Y indicating the actual class membership. The columns X_k, $1 \le k \le n$ are called *analytic variables*, and Y is called the *target variable*, which indicates membership in a target class y. In case of a binary classification, for

each row index i, the label $Y(i)$ is equal to 1 if and only if individual i is in class y (Formula 2.36).

$$Y(i) := \begin{cases} 1 & \text{if } i \in y \\ 0 & \text{else.} \end{cases} \tag{2.36}$$

A machine learning process for inductive sharp classification generates a model $M_y(i)$, mapping from the Cartesian product of the analytic variable ranges into the set $\{0, 1\}$, indicating inductive support for the hypothesis that $i \in y$. As discussed in the section on induction, the model should provide support for inductive inferences about an individual's class membership: Given $M_y(i) = 1$, the likelihood of $i \in y$ should be greater than the likelihood of $i \notin y$.

The inductive model M_y can be applied for prediction to a new dataset with unknown class indicator, which is either a test set for performance evaluation or a prediction set, where the model is applied to forecast class membership of new data. The test set or prediction set d' has the same structure as the training set d, except that the class membership is unknown, and thus, the target variable Y is empty. The classifier M_y, derived from the training set, can be used for predicting the class memberships of representations of individuals $i \in d$. The model output prediction $M_y(i)$ yields an inductive classification defined by $\{i \mid M_y(i) = 1\}$.

In order to evaluate the quality of prediction of a crisp classifier model, several measures can be computed. In this section, likelihood ratio and Pearson correlation are mentioned. The greater the ratio between likelihood for target class membership, given a positive prediction, and the likelihood for target class membership, given a negative prediction, the better the inductive support of the model. Thus, the predictive model can be evaluated by the likelihood ratio of target class membership given the model output (Formula 2.37).

$$LR\big(Y(i) = 1 \mid M_y(i) = 1\big) := \frac{p\big(M_y(i) = 1 \mid Y(i) = 1\big)}{p\big(M_y(i) = 1 \mid Y(i) = 0\big)} \tag{2.37}$$

Working with binary or Boolean target indicators and model indicators allows the evaluation of predictive quality by a measure of correlation of the two variables M_y and Y (Formula 2.38). The correlation between two numerical variables can be measured by the Pearson correlation coefficient as the ratio between the covariance of the two variables and the square root of the product of individual variances (Weisstein, 2010b).

$$\text{corr}\big(M_y, Y\big) = \frac{E\big((M_y - \text{avg}(M_y))(Y - \text{avg}(Y))\big)}{\text{stddev}\big(M_y\big) \cdot \text{stddev}(Y)} \tag{2.38}$$

The advantage of the correlation coefficient is its availability in database systems. Every standard SQL (structured query language) database has an implementation of correlation as an aggregate function. Thus, using the correlation

coefficient, evaluating the predictive performance of a model in a database is fast and simple. However, it is important to stress that evaluating predictions with a measure of correlation is only meaningful if the target variable as well as the predictive variable are Boolean, Zadehan, or numeric in nature. It will not work for ordinal or categorical target classes, except if they are transformed into a set of Boolean variables.

For example, in database marketing, the process of target group selection uses classifiers to select customers who are likely to buy a certain product. In order to do this, a classifier model can be computed in the following way: Given set of customers C, we know whether they have bought product A or not. Let c be an individual customer, and C_A be the set of customers who bought product A. Then, the value $Y(c)$ of target variable Y for customer c is defined in Formula (2.39).

$$Y(c) = \begin{cases} 1 & \text{if } c \in C_A \\ 0 & \text{else.} \end{cases} \qquad (2.39)$$

The analytic variables for customers are selected from every known customer attribute, such as age, location, transaction behavior, recency, frequency, and monetary value of purchase. The aim of the classifier induction process is to learn a model, M_{C_A}, that provides a degree of support for the inductive inference that a customer is interested in the target product A. This prediction, $M_{C_A}(c) \in \{0, 1\}$, should provide a better likelihood to identify potential buyers of product A, and it should optimally correlate with the actual product usage of existing and future customers.

2.4 Inductive Fuzzy Classification

The understanding of IFC in the proposed research approach is an inductive gradation of the degree of membership of individuals in classes. In many interpretations, the induction step consists of learning *fuzzy rules* (e.g., Dianhui, Dillon, & Chang, 2001; Hu, Chen, & Tzeng, 2003; Roubos, Setnes, & Abonyi, 2003; Wang & Mendel, 1992). In this thesis, IFC is understood more generally as inducing membership functions to fuzzy classes and assigning individuals to those classes. In general, a membership function can be any function mapping into the interval between 1 and 0. Consequently, IFC is defined as the process of assigning individuals to fuzzy sets for which membership functions are generated from data so that the membership degrees are based on an inductive inference.

An inductive fuzzy class, y', is defined by a predictive scoring model, $M_y : U \to [0, 1]$, for membership in a class, y. This model represents an inductive membership function for y', which maps from the universe of discourse U into the interval between 0 and 1 (Formula 2.40).

$$\mu_{y'} : U \to [0, 1] := M_y \qquad (2.40)$$

Consider the following fuzzy classification predicate $P(i, y) :=$ "i is likely a member of y." This is a fuzzy proposition (Zadeh, 1975a) as a function of i and y, which indicates that there is inductive support for the conclusion that individual i belongs to class y. The truth function, τ^{ZL}, of this fuzzy propositional function can be defined by the membership function of an inductive fuzzy class, y'. Thus, $P(i, y)$ is a fuzzy restriction on U defined by $\mu_{y'}$ (Formula 2.41).

$$\tau^{ZL}(\text{``}i \text{ is likely a member of } y\text{''}) := M_y(i) \qquad (2.41)$$

In practice, any function that assigns values between 0 and 1 to data records can be used as a fuzzy restriction. The aim of IFC is to calculate a membership function to a fuzzy set of likely members in the target class. Hence, any type of classifier with a normalized numeric output can be viewed as an inductive membership function to the target class, or as a truth function for the fuzzy proposition $P(i, y)$. State of the art methods for IFC in that sense include linear regression, logistic regression, naïve Bayesian classification, neural networks, fuzzy classification trees, and fuzzy rules. These are classification methods yielding numerical predictions that can be normalized in order to serve as a membership function to the inductive fuzzy class y' (Formula 2.42).

$$y' := \left\{ i \in U \mid i \text{ is likely a member of } y \right\} \qquad (2.42)$$

2.4.1   Univariate Membership Function Induction

This section describes methods to derive membership functions for one variable based on inductive methods. First, unsupervised methods are described, which do not require learning from a target class indicator. Second, supervised methods for predictive membership functions are proposed.

Numerical attributes can be fuzzified in an unsupervised way, that is, without a target variable, by calculating a membership function to a fuzzy class *x is a large number*, denoted by the symbol $\uparrow$: the fuzzy set of attribute values that are large relative to the available data. This membership function, $\mu_{\uparrow} : \text{dom}(C) \to [0, 1]$, maps from the attribute domain of the target variable into the set of Zadehan truth values. This unsupervised fuzzification serves two purposes. First, it can be used to automatically derive linguistic interpretations of numerical data, such as "large" or "small." Second, it can be used to transform numerical attributes into Zadehan target variables in order to calculate likelihoods of fuzzy events. There are two approaches proposed here to compute a membership function to this class: percentile ranks and linear normalization based on minimum and maximum.

For a numeric or ordinal variable X with a value $x \in \text{dom}(X)$, the percentile rank (PR) is equal to the sampled probability that the value of the variable X is smaller than x. This sampled probability is calculated by the percentage of values in $\text{dom}(X)$
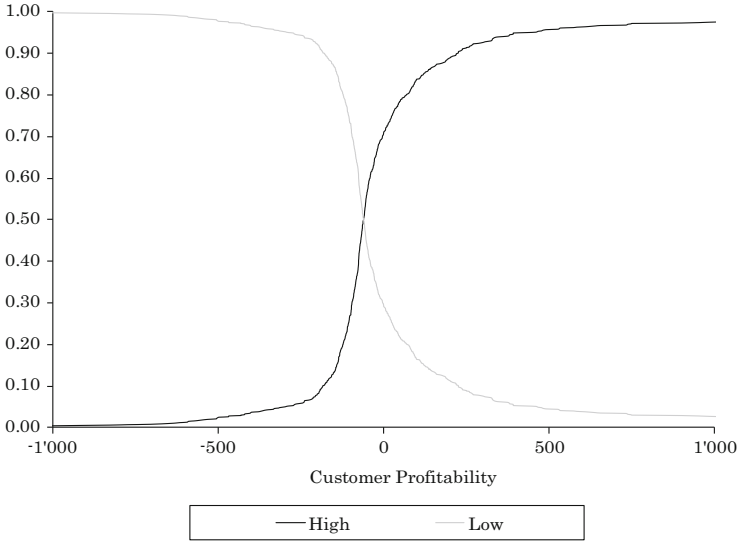
Fig. 2.7 Unsupervised IFC by percentile rank. Adapted from "An Inductive Approach to Fuzzy Marketing Analytics," by M. Kaufmann, 2009, In M. H. Hamza (Ed.), Proceedings of the 13th IASTED International Conference on Artificial Intelligence and Soft Computing: Copyright 2009 by Publisher

that are smaller than or equal to x. This sampled probability can be transformed into a degree of membership in a fuzzy set. Inductively, the sampled probability is taken as an indicator for the support of the inductive inference that a certain value, X_i, is large in comparison to the distribution of the other attribute values. The membership degree of x in the fuzzy class *"relatively large number"*, symbolized by $\uparrow$, is then defined as specified in Formula (2.43).

$$\mu_\uparrow(x) := p(X < x) \tag{2.43}$$

For example, customers can be classified by their profitability. The percentile rank of profitability can be viewed as a membership function of customers in the fuzzy set $\uparrow$ of customers with a high profitability. Figure 2.7 shows an example of an IFC-PR of customer profitability for a financial service provider.

A simpler variant of unsupervised fuzzification for generating a membership function for a *relatively large number* ($\uparrow$) is *linear normalization* (IFC-LN). For a numerical attribute C, it is defined as the relative distance to the minimal attribute value, as specified in Formula (2.44).

$$\mu_\uparrow(C(i)) := \frac{C(i) - \min(C)}{\max(C) - \min(C)} \tag{2.44}$$

For the *membership function induction* (MFI) methods in the following sections, the target variable for supervised induction must be a Zadehan variable, $Y : U \rightarrow [0, 1]$ mapping from the universe of discourse (the set of possible individuals) into the interval of Zadehan truth values between 0 and 1. Thus, Yi) indicates the degree of membership of individual i in the target class y. In the special case of a Boolean target class, Yi) is equal to 1 if $i \in y$, and it is equal to 0 if $i \notin y$. In the analytic training data, a target class indicator Y can be deduced from data attributes in the following way:

- If an attribute, A, is Zadehan with a range between 0 and 1, it can be defined directly as the target variable. In fact, if the variable is Boolean, this implies that it is also Zadehan, because it is a special case (Formula 2.45).

$$\text{Zadehan}(A) \Rightarrow \mu_y(i) := A(i) \tag{2.45}$$

- If an attribute, B, is categorical with a range of n categories, it can be transformed into n Boolean variables μ_{y^k} ($k = 1, 2, \ldots, n$), where $\mu_{y^k}(i)$ indicates whether record i belongs to class k, as specified by Formula (2.46).

$$\text{categorical}(B) \Rightarrow \mu_{y^k}(i) := \begin{cases} 1 & \text{if } B(i) = k \\ 0 & \text{else.} \end{cases} \tag{2.46}$$

- If an attribute, C, is numeric, this thesis proposes application of an unsupervised fuzzification, as previously specified, in order to derive a Zadehan target variable, as formalized in Formula (2.47). This is called an *inductive target fuzzification* (ITF).

$$\text{numerical}(C) \Rightarrow \mu_y(i) := \mu_\uparrow(C(i)) \tag{2.47}$$

The second approach for univariate membership function induction is supervised induction based on a target variable. In order to derive membership functions to inductive fuzzy classes for one variable based on the distribution of a second variable, it is proposed to normalize comparisons (ratios and differences) of likelihoods for membership function induction. For example, a normalized likelihood ratio can represent a membership degree to an inductive fuzzy class.

The basic idea of *inductive fuzzy classification based on normalized likelihood ratios* (IFC-NLR) is to transform inductive support of target class membership into a membership function with the following properties: The higher the likelihood of $i \in y$ in relation to $i \notin y$, the greater the degree membership of i in y'. For an attribute X, the NLR function calculates a membership degree of a value $x \, \epsilon \, \text{dom}(X)$ in the predictive class y', based on the likelihood of target class membership. The

resulting membership function is defined as a relation between all values in the domain of the attribute X and their NLRs.

As discussed in Sect. 2.3.1, following the principle of likelihood (Hawthorne, 2008), the ratio between the two likelihoods is an indicator for the degree of support for the inductive conclusion that $i \in y$, given the evidence that $X(i) = x$ In order to transform the likelihood ratio into a fuzzy set membership function, it can be normalized in the interval between 0 and 1. Luckily, for every ratio, $R = A/B$, there exists a normalization, $N = A/(A + B)$, having the following properties:

- N is close to 0 if R is close to 0.
- N is equal to 0.5 if and only if R is equal to 1.
- N is close to 1 if R is a large number.

This kind of normalization is applied to the aforementioned likelihood ratio in order to derive the NLR function. Accordingly, the membership μ of an attribute value x in the target class prediction y' is defined by the corresponding NLR, as formalized in Formula (2.48).

$$\mu_{y'}(x) := NLR(y \mid x) = \frac{L(y \mid x)}{L(y \mid x) + L(\neg y \mid x)} \tag{2.48}$$

In fact, one can demonstrate that the NLR function is equal to the posterior probability of y, conditional to x, if both hypotheses y and $\neg y$ are assumed to be of equal prior probability (Formula 2.52), by application of the second form of Bayes' theorem (Joyce, 2003), as presented in Formula (2.50). The trick is to express the probability of the evidence $p(x)$ in terms of a sum of products of prior probabilities, p_0, and measured likelihoods, L, of the hypothesis and its alternative by application of Formula (2.33).

Theorem

$$NLR(y \mid x) = p(y \mid x) \Leftrightarrow p_0(y) = p_0(\neg y) \tag{2.49}$$

Proof

$$
\begin{aligned}
p(y \mid x) &= \frac{p_0(y)L(y \mid x)}{p(x)} \quad \text{(c.f. Formula 2.33)} \\[2mm]
&= \frac{p_0(y)L(y \mid x)}{p_0(y)L(y \mid x) + p_0(\neg y)L(\neg y \mid x)} \\
&\quad [if \quad p(x) = p(y)p(x \mid y) + p(\neg y)p(x \mid \neg y) \\
&\quad and \quad p(y) := p_0(y) \ and \ p(x \mid y) := L(y \mid x)] \\[2mm]
&= \frac{L(y \mid x)}{L(y \mid x) + L(\neg y \mid x)} \quad [if \ p_0(y) :=: p_0(\neg y)] \\[2mm]
&=: NLR(y \mid x), \quad q.e.d.
\end{aligned}
\tag{2.50}
$$

Alternatively, two likelihoods can be compared by a normalized difference, as shown in Formula (2.51). In that case, the membership function is defined by a

normalized likelihood difference (NLD), and its application for classification is called *inductive fuzzy classification by normalized likelihood difference* (IFC-NLD). In general, IFC methods based on *normalized likelihood comparison* can be categorized by the abbreviation IFC-NLC.

$$\mu_{y'}(x) := NLD(y \mid x) = \frac{L(y \mid x) - L(\neg y \mid x) + 1}{2} \tag{2.51}$$

If a target attribute is continuous, it can be mapped into the Zadehan domain of numeric truth-values between 0 and 1, and membership degrees can be computed by a normalized ratio of likelihoods of fuzzy events. If the target class is fuzzy, for example because the target variable is gradual, the likelihoods are calculated by fuzzy conditional relative frequencies based on fuzzy set cardinality (Dubois & Prade, 1980). Therefore, the formula for calculating the likelihoods is generalized in order to be suitable for both sharp and fuzzy characteristics. Thus, in the general case of variables with fuzzy truth-values, the likelihoods are calculated as defined in Formula (2.52).

$$
\begin{aligned}
L(y|x) &:= \frac{\sum_{i=1}^{n} \mu_x(i)\mu_y(i)}{\sum_{i=1}^{n} \mu_y(i)} \\[2mm]
L(\neg y|x) &:= \frac{\sum_{i=1}^{n} \mu_x(i)\left(1 - \mu_y(i)\right)}{\sum_{i=1}^{n} \left(1 - \mu_y(i)\right)}
\end{aligned}
\tag{2.52}
$$

Accordingly, the calculation of membership degrees using the NLR function (Formula 2.52) works for both categorical and fuzzy target classes and for categorical and fuzzy analytic variables. For numerical attributes, the attribute values can be discretized using quantiles, and a piecewise linear function can be approximated to average values in the quantiles and the corresponding NLR. A membership function for individuals based on their attribute values can be derived by aggregation, as explained in Sect. 2.4.2.

Following the different comparison methods for conditional probabilities described by Joyce (2003), different methods for the induction of membership degrees using conditional probabilities are proposed in Table 2.1. They have been chosen in order to analytically test different Bayesian approaches listed by Joyce (2003) for their predictive capabilities. Additionally, three experimental measures were considered: logical equivalence, normalized correlation, and a measure based on minimum and maximum. In those formulae, x and y are assumed to be Zadehan with a domain of [0,1] or Boolean as a special case. These formulae are evaluated as parameters in the meta-induction experiment described in Sect. 4.2.

A method for discretization of a numerical range is the calculation of quantiles or n-tiles for the range of the analytical variable. A quantile discretization using n-tiles partitions the variable range into n intervals having the same number of individuals. The quantile $Q_n^Z(i)$ for an attribute value $Z(i)$, of a numeric attribute

Table 2.1 Proposed formulae for induction of membership degrees

Method	Formula
Likelihood of y given x (L)	$L(y|x) = p(x|y)$
Normalized likelihood ratio (NLR)	$NLR(y \mid \mathrm{x}) = \dfrac{p(x \mid y)}{p(x \mid y) + p(x \mid \neg y)}$
Normalized likelihood ratio unconditional (NLRU)	$NLRU(y \mid \mathrm{x}) \dfrac{p(x \mid y)}{p(x \mid y) + p(x)}$
Normalized likelihood difference (NLD)	$NLD(y \mid x) = \dfrac{p(x \mid y) - p(x \mid \neg y) + 1}{2}$
Normalized likelihood difference unconditional (NLDU)	$NLDU(y \mid x) = \dfrac{p(y \mid x) - p(x) + 1}{2}$
Conditional probability of y given x (CP)	$p(y|x)$
Normalized probability ratio (NPR)	$NPR(y \mid \mathrm{x}) = \dfrac{p(y \mid \mathrm{x})}{p(y \mid \mathrm{x}) + p(y \mid \neg \mathrm{x})}$
Normalized probability ratio unconditional (NPRU)	$NPRU(y \mid \mathrm{x}) = \dfrac{p(y \mid \mathrm{x})}{p(y \mid \mathrm{x}) + p(y)}$
Normalized probability difference (NPD)	$NPD(y \mid x) = \dfrac{p(y \mid x) - p(y \mid \neg x) + 1}{2}$
Normalized probability difference unconditional (NPDU)	$NPDU(y \mid x) = \dfrac{p(y \mid x) - p(y) + 1}{2}$
Equivalence—if and only if (IFF)	avg $((1 - \mathrm{x} \cdot (1 - \mathrm{y})) \cdot (1 - \mathrm{y} \cdot (1 - \mathrm{x})))$
Minimum–maximum (MM)	$\dfrac{p(y \mid x) + \min_{z \in \mathrm{dom(X)}}(p(y \mid z))}{\min_{z \in \mathrm{dom(X)}}(p(y \mid z)) + \max_{z \in \mathrm{dom(X)}}(p(y \mid z))}$
Normalized correlation (NC)	$\dfrac{corr(x,y) + 1}{2}$

Z and an individual i, is calculated using Formula (2.53), where n is the number of quantiles, m is the total number of individuals or data records, $rank_Z(i)$ is the position of the individual in the list of individuals sorted by their values of attribute Z, and $trunc(r)$ is the closest integer that is smaller than the real value r.

$$Q_n^Z(i) := trunc\left(\frac{n}{m}\left(rank_Z(i) - 1\right)\right) \qquad (2.53)$$

The rank of individual i relative to attribute Z, $rank_Z(i)$, in a dataset S is the number of other individuals, h, that have higher values in attribute Z, calculated using Formula (2.54).

$$rank_Z(i) := \left|\left\{h \in S \mid \forall i \in S : Z(h) > Z(i)\right\}\right| \qquad (2.54)$$

In order to approximate a linear function, the method of two-dimensional *piecewise linear function approximation* (PLFA) is proposed. For a list of points in $\mathbb{R}^2$, ordered by the first coordinate, $(\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle)$, for every point $\langle x_1, y_1 \rangle$ except the last one ($i = 1, 2, \ldots, n - 1$), a linear function, $f_i(x) = a_i x + b_i$, can be interpolated to its neighbor point, where a_i is the slope (Formula 2.55) and b_i is the intercept (Formula 2.56) of the straight line.

$$a_i := \frac{(y_{i+1} - y_i)}{(x_{i+1} - x_i)} \qquad (2.55)$$

$$b_i := y_i - a_i x_i \qquad (2.56)$$

For the calculation of membership degrees for quantiles, the input is a list of points with one point for every quantile k. The first coordinate is the average of the attribute values in k. The second coordinate is the inductive degree of membership $\mu_{y'}$ in target class y, given $Z(i)$ is in quantile k, for example derived using the NLR function.

$$\begin{aligned} y_k &:= \mu_{y'}(k) \\ x_k &:= avg\{Z(i) \mid Q_n^z(i) = k\} \end{aligned} \qquad (2.57)$$

Finally, a continuous, piecewise affine membership function can be calculated, truncated below 0 and above 1, and is composed of straight lines for every quantile $k = 1, \ldots, n - 1$; $n \geq 2$ of the numeric variable Z (Formula 2.58).

$$\mu_y(x) := \begin{cases} 0 & |a_1 x + b_1 \leq 0 \vee a_{n-1} x + b_{n-1} \leq 0 \\ a_1 x + b_1 & |x \leq x_2 \\ \vdots & \vdots \\ a_k x + b_k & x_k < x \leq x_{k+1} \\ \vdots & \vdots \\ a_{n-1} x + b_{n-1} & |x > x_{n-1} \\ 1 & a_1 x + b_1 \geq 1 \vee a_{n-1} x + b_{n-1} \geq 1 \end{cases} \qquad (2.58)$$

The number of quantiles can be optimized, so that the correlation of the membership function with the target variable is optimal, as illustrated in Fig. 2.8.

2.4.2   Multivariate Membership Function Induction

As shown in Fig. 2.9, the proposed process for inducing a multivariate inductive fuzzy class consists of preparing the data, inducing univariate membership functions for the attributes, transforming the attribute values into univariate target membership degrees, classifying individuals by aggregating the fuzzified attributes into a multivariate fuzzy classification, and evaluating the predictive performance of the resulting model.

The idea of the process is to develop a fuzzy classification that ranks the inductive membership of individuals, i, in the target class y gradually. This fuzzy classification will assign individuals an inductive membership degree to the predictive inductive fuzzy class y' using the multivariate model $\mu_{y'}$. The higher the
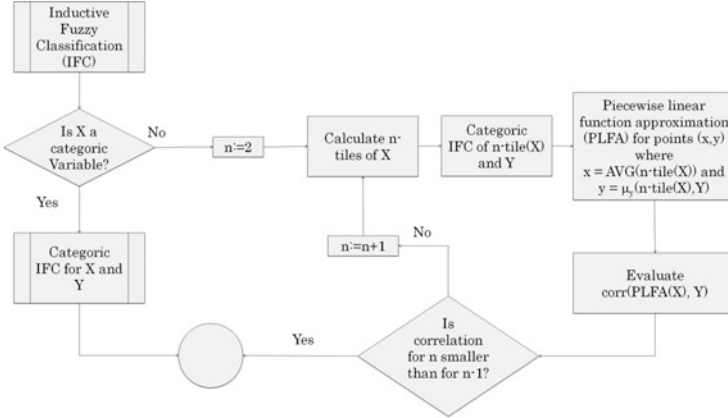
Fig. 2.8 Computation of membership functions for numerical variables

Fig. 2.9 Proposed method for multivariate IFC

inductive degree of membership $\mu_{y'}(i)$ of an individual in y', the greater the degree of inductive support for class membership in the target class y.

In order to accomplish this, a training set is prepared from source data, and the relevant attributes are selected using an interestingness measure. Then, for every attribute X_k, a membership function, $\mu_{y'}^k : \text{dom}(X_k) \to [0, 1]$, is defined. Each $\mu_{y'}^k$ is induced from the data such that the degree of membership of an attribute value $X_k(i)$ in the inductive fuzzy class y' is proportional to the degree of support for the inference that $i \in y$. After that, in the univariate classification step, each variable, X_k, is fuzzified using $\mu_{y'}^k$. The multivariate fuzzy classification step consists of aggregating the fuzzified attributes into one multivariate model, $\mu_{y'}$, of data elements that represents the membership function of individual i in y'. This inductive fuzzy class corresponds to an IFC that can be used for predictive ranking of data elements. The last step of the process is model evaluation through analyzing the prediction performance of the ranking. Comparing the forecasts with the real class memberships in a test set does this. In the following paragraphs, every step of the IFC process is described in detail.

In order to analyze the data, combining data from various sources into a single coherent matrix composes a training set and a test set. All possibly relevant attributes are merged into one table structure. The class label Y for the target variable has to be defined, calculated, and added to the dataset. The class label is

restricted to the Zadehan domain, as defined in the previous section. For multiclass predictions, the proposed process can be applied iteratively.

Intuitively, the aim is to assign to every individual a membership degree in the inductive fuzzy class y'. As explained in Sect. 2.3.1, this degree indicates support for the inference that an individual is a member of the target class y. The membership function for y' will be derived as an aggregation of inductively fuzzified attributes. In order to accomplish this, for each attribute, a univariate membership function in the target class is computed, as described in the previous section.

Once the membership functions have been induced, the attributes can be fuzzified by application of the membership function to the actual attribute values. In order to do so, each variable, X_k, is transformed into an inductive degree of membership in the target class. The process of mapping analytic variables into the interval [0, 1] is an attribute fuzzification. The resulting values can be considered a membership degree to a fuzzy set. If this membership function indicates a degree of support for an inductive inference, it is called an *inductive attribute fuzzification* (IAF), and this transformation is denoted by the symbol $\rightsquigarrow$ inFormula (2.59).

$$X_k(i) \rightsquigarrow \mu_{y'}(X_k(i)) \tag{2.59}$$

The most relevant attributes are selected before the IFC core process takes place. The proposed method for attribute selection is a ranking of the Pearson correlation coefficients (Formula 2.38) between the inductively fuzzified analytic variables and the (Zadehan) target class indicator Y. Thus, for every attribute, X_k, the relevance regarding target y is defined as the correlation of its inductive fuzzification with the target variable (see Sect. 3.1.1).

In order to obtain a multivariate membership function for individuals i derived from their fuzzified attribute values $\mu_{y'}(X_k(i))$, their attribute value membership degrees are aggregated. This corresponds to a multivariate fuzzy classification of individuals. Consequently, the individual's multivariate membership function $\mu_{y'}$: $U \rightarrow [0, 1]$ to the inductive fuzzy target class y' is defined as an aggregation, *aggr*, of the membership degrees of n attributes, X_k, $k = 1, 2, \ldots, n$ (Formula 2.60).

$$\mu_{y'}(i) := aggr\left(\mu_{y'}(X_1(i)), \ldots, \mu_{y'}(X_n(i))\right) \tag{2.60}$$

By combining the inductively fuzzified attributes into a multivariate fuzzy class of individuals, a multivariate predictive model, $\mu_{y'}$, is obtained from the training set. This corresponds to a classification of individuals by the fuzzy proposition "i is likely a member of y," for which the truth value is defined by an aggregation of the truth values of fuzzy propositions about the individual's attributes, $X_k(i)$ is y'. This model can be used for IFC of unlabeled data for predictive ranking. Applying an alpha cutoff, $\left\{i \mid \mu_{y'}(i) \geq \alpha\right\}$, an $\alpha \in [0, 1]$ leads to a binary classifier.

There are different possibilities for calculating the aggregation, *aggr*. Simpler methods use an average of the attribute membership degrees, logical conjunction

(minimum, algebraic product), or logical disjunction (maximum, algebraic sum). More sophisticated methods involve the supervised calculation of a multivariate model. In this thesis, normalized or cutoff linear regression, logistic regression, and regression trees are considered. These different aggregation methods were tested as a parameter in the meta-induction experiment described in Sect. 4.2 in order to find an optimal configuration.

Finally, in order to evaluate predictive performance, the classifier is applied to a hold-out test set, and the predictions $\mu_{y'}(i)$ are compared with the actual target variable (i). The correlation between the prediction and the target, $corr\left(\mu_{y'}, Y\right)$, can be used to compare the performance of different IFC models.