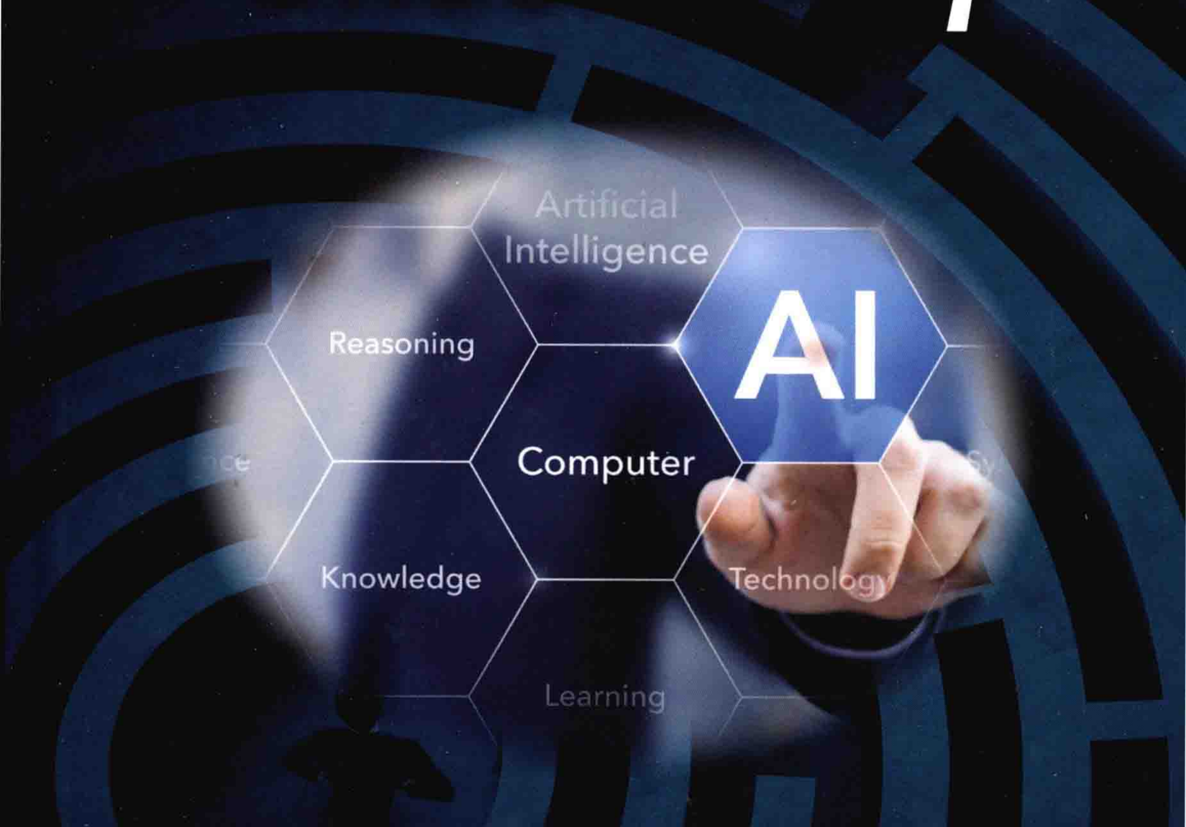
XIZHAO WANG ▪ JUNHAI ZHAI

LEARNING WITH Uncertainty



CRC Press
Taylor & Francis Group

LEARNING WITH Uncertainty

XIZHAO WANG
JUNHAI ZHAI

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper
Printed at CPI on sustainably sourced paper
Version Date: 20161021

International Standard Book Number-13: 978-1-4987-2412-8 (Hardback)

Library of Congress Cataloging-in-Publication Data

Names: Wang, Xizhao, author. | Zhai, Junhai, author.
Title: Learning with uncertainty / Xizhao Wang, Junhai Zhai.
Description: Boca Raton : CRC Press, [2016] | Includes bibliographical references and index.
Identifiers: LCCN 2016030316| ISBN 9781498724128 (acid-free paper) | ISBN 9781498724135 (e-book)
Subjects: LCSH: Machine learning. | Fuzzy decision making. | Decision trees.
Classification: LCC Q325.5 .W36 2016 | DDC 006.3/1--dc23
LC record available at https://lccn.loc.gov/2016030316

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

LEARNING WITH
Uncertainty

Preface

Learning is an essential way for humans to gain wisdom and furthermore for machines to acquire intelligence. It is well acknowledged that learning algorithms will be more flexible and more effective if the uncertainty can be modeled and processed during the process of designing and implementing the learning algorithms.

Uncertainty is a common phenomenon in machine learning, which can be found in every stage of learning, such as data preprocessing, algorithm design, and model selection. Furthermore, one can find the impact of uncertainty processing on various machine learning techniques; for instance, uncertainty can be used as a heuristic to generate decision tree in inductive learning, it can be employed to measure the significance degree of samples in active learning, and it can also be applied in ensemble learning as a heuristic to select the basic classifier for integration. This book makes an initial attempt to systematically discuss the modeling and significance of uncertainty in some processes of learning and tries to bring some new advancements in learning with uncertainty.

The book contains five chapters. Chapter 1 is an introduction to uncertainty. Four kinds of uncertainty, that is, randomness, fuzziness, roughness, and nonspecificity, are briefly introduced in this chapter. Furthermore, the relationships among these uncertainties are also discussed in this chapter. Chapter 2 introduces the induction of decision tree with uncertainty. The contents include how to use uncertainty to induce crisp decision trees and fuzzy decision trees. Chapter 2 also discusses how to use uncertainty to improve the generalization ability of fuzzy decision trees. Clustering under uncertainty environment is discussed in Chapter 3. Specifically, the basic concepts of clustering are briefly reviewed in Section 3.1. Section 3.2 introduces two types of clustering, that is, partition-based clustering and hierarchy-based clustering. Validation functions of clustering are discussed in Section 3.3 and feature-weighted fuzzy clustering is addressed in next sections. In Chapter 4, we first present an introduction to active learning in Section 4.1. Two kinds of active learning techniques, that is, uncertainty sampling and query by committee, are discussed in Section 4.2. Maximum ambiguity–based active learning is presented in Section 4.3. A learning approach for support vector machine is presented in Section 4.4. Chapter 5 includes five sections. An introduction to ensemble learning is reviewed in Section 5.1.

Bagging and boosting, multiple fuzzy decision trees, and fusion of classifiers based on upper integral are discussed in the next sections, respectively. The relationship between fuzziness and generalization in ensemble learning is addressed in Section 5.5.

Taking this opportunity, we deliver our sincere thanks to those who offered us great help during the writing of this book. We appreciate the discussions and revision suggestions given by those friends and colleagues, including Professor Shuxia Lu, Professor Hongjie Xing, Associate Professor Huimin Feng, Dr. Chunru Dong, and our graduate students, Shixi Zhao, Sheng Xing, Shaoxing Hou, Tianlun Zhang, Peizhou Zhang, Bo Liu, Liguang Zang, etc. Our thanks also go to the editors who help us plan and organize this book.

The book can serve as a reference book for researchers or a textbook for senior undergraduates and postgraduates majored in computer science and technology, applied mathematics, automation, etc. This book also provides some useful guidelines of research for scholars who are studying the impact of uncertainty on machine learning and data mining.

Xizhao Wang
Shenzhen University, China

Junhai Zhai
Hebei University, China

Symbols and Abbreviations

Symbols

A	The set of conditional attributes
A_{ij}	The jth fuzzy linguistic term of the ith attribute
$Bel(\cdot,\cdot)$	Belief function
$bpa(\cdot,\cdot)$	Basic probability assignment function
C	The set of decision attributes
$(C)\int f\, d\mu$	Choquet fuzzy integral
$D(\cdot \parallel \cdot)$	KL-divergence
$DP(\cdot)$	Decision profile
F_α	The α cut-set of fuzzy set F
$H^\dagger$	Moore–Penrose generalized inverse of matrix H
I	Index set
$k(\cdot,\cdot)$	Kernel function
$p_{ij}^{(l)}$	The relative frequency of A_{ij} regarding to the lth class
POS_P^Q	The positive region of P with respect to Q
$P_\theta(y\|x)$	Given x, the posterior probability of y under the model θ
R^d	d dimension Euclidean space
$\underline{R}(X)$	Lower approximation of X with respect to R
$\overline{R}(X)$	Upper approximation of X with respect to R
S_B	Scatter matrix between classes
S_W	Scatter matrix within class
$(S)\int f\, d\mu$	Sugeno fuzzy integral
U	Universe
u_{ij}	The membership of the ith instance with respect to the jth class
$(U)\int f\, d\mu$	Upper fuzzy integral
$V(y_i)$	The number of votes of y_i
x_i	The ith instance

$[x]_R$	Equivalence class of x with respect to R
y_i	The class label of x_i
β	Degree of confidence
γ_P^Q	The significance of P with respect to Q
$\delta_{ij}^{(w)}$	The weighted similarity degree between the ith and jth samples
$\mu_A(x)$	Membership function

Abbreviations

AGNES	AGglomerative NESting
BIRCH	Balanced iterative reducing and clustering using hierarchies
CF	Certainty factor
CNN	Condensed nearest neighbors
CURE	Clustering Using REpresentatives
DDE	Dynamic differential evolution
DE	Differential evolution
DES	Differential evolution strategy
DIANA	DIvisive ANAlysis
DT	Decision table
ELM	Extreme learning machine
FCM	Fuzzy C-means
F-DCT	Fuzzy decision tree
FDT	Fuzzy decision table
F-ELM	Fuzzy extreme learning machine
F-KNN	Fuzzy K-nearest neighbors
FLT	Fuzzy linguistic term
FRFDT	Fuzzy rough set–based decision tree
GD	Gradient descent
GD-FWL	Gradient descent–based feature weight learning
IBL	Instance based learning
K-NN	K-nearest neighbors
LS-SVM	Least square support vector machine
MABSS	Maximum ambiguity based sample selection
MEHDE	Hybrid differential evolution with multistrategy cooperating evolution
MFDT	Multiple fuzzy decision tree
QBC	Query by committee
ROCK	RObust Clustering using linKs
SMTC-C	Similarity matrix's transitive closure
SVM	Support vector machine
WFPR	Weighted fuzzy production rules

Contents

Preface ix
Symbols and Abbreviations xi

1 Uncertainty 1
1.1 Randomness 1
1.1.1 Entropy 1
1.1.2 Joint Entropy and Conditional Entropy 3
1.1.3 Mutual Information 4
1.2 Fuzziness 5
1.2.1 Definition and Representation of Fuzzy Sets 6
1.2.2 Basic Operations and Properties of Fuzzy Sets 7
1.2.3 Fuzzy Measures 8
1.3 Roughness 9
1.4 Nonspecificity 10
1.5 Relationships among the Uncertainties 11
1.5.1 Entropy and Fuzziness 12
1.5.2 Fuzziness and Ambiguity 14
References 15

2 Decision Tree with Uncertainty 17
2.1 Crisp Decision Tree 17
2.1.1 ID3 Algorithm 18
2.1.2 Continuous-Valued Attributes Decision Trees 22
2.2 Fuzzy Decision Tree 28
2.3 Fuzzy Decision Tree Based on Fuzzy Rough Set Techniques 37
2.3.1 Fuzzy Rough Sets 37
2.3.2 Generating Fuzzy Decision Tree with Fuzzy Rough Set Technique 40
2.4 Improving Generalization of Fuzzy Decision Tree by Maximizing Fuzzy Entropy 44
2.4.1 Basic Idea of Refinement 44
2.4.2 Globally Weighted Fuzzy If-Then Rule Reasoning 44

2.4.3 Refinement Approach to Updating the Parameters............. 49
2.4.3.1 Maximum Fuzzy Entropy Principle.................. 50
References.. 55

3 Clustering under Uncertainty Environment 59
3.1 Introduction.. 59
3.2 Clustering Algorithms Based on Hierarchy or Partition................ 60
3.2.1 Clustering Algorithms Based on Hierarchy..................... 60
3.2.2 Clustering Algorithms Based on Partition...................... 66
3.3 Validation Functions of Clustering................................... 70
3.4 Feature Weighted Fuzzy Clustering.................................. 71
3.5 Weighted Fuzzy Clustering Based on Differential Evolution........... 73
3.5.1 Differential Evolution and Dynamic Differential Evolution.... 73
3.5.1.1 Basic Differential Evolution Algorithm................ 73
3.5.1.2 Dynamic Differential Evolution Algorithm........... 77
3.5.2 Hybrid Differential Evolution Algorithm Based on Coevolution with Multi-Differential Evolution Strategy........ 78
3.6 Feature Weight Fuzzy Clustering Learning Model Based on MEHDE... 80
3.6.1 MEHDE-Based Feature Weight Learning: MEHDE-FWL.. 81
3.6.2 Experimental Analysis.. 82
3.6.2.1 Comparison between MEHDE-FWL and GD-FWL Based on FCM.............................. 84
3.6.2.2 Comparisons Based on SMTC Clustering............ 88
3.6.2.3 Efficiency Analysis of GD-, DE-, DDE-, and MEHDE-Based Searching Techniques................ 90
3.7 Summary.. 95
References.. 96

4 Active Learning with Uncertainty 99
4.1 Introduction to Active Learning...................................... 99
4.2 Uncertainty Sampling and Query-by-Committee Sampling.......... 102
4.2.1 Uncertainty Sampling... 102
4.2.1.1 Least Confident Rule................................ 102
4.2.1.2 Minimal Margin Rule................................ 103
4.2.1.3 Maximal Entropy Rule............................... 103
4.2.2 Query-by-Committee Sampling................................ 103
4.3 Maximum Ambiguity–Based Active Learning......................... 105
4.3.1 Some Concepts of Fuzzy Decision Tree........................ 106
4.3.2 Analysis on Samples with Maximal Ambiguity................ 107
4.3.3 Maximum Ambiguity–Based Sample Selection................ 109
4.3.4 Experimental Results... 111

4.4 Active Learning Approach to Support Vector Machine................ 120
4.4.1 Support Vector Machine.. 122
4.4.2 SVM Active Learning... 123
4.4.3 Semisupervised SVM Batch Mode Active Learning............ 124
4.4.4 IALPSVM: An Informative Active Learning Approach to SVM.. 125
4.4.5 Experimental Results and Discussions......................... 128
4.4.5.1 Experiments on an Artificial Data Set by Selecting a Single Query Each Time................. 128
4.4.5.2 Experiments on Three UCI Data Sets by Selecting a Single Query Each Time........................... 131
4.4.5.3 Experiments on Two Image Data Sets by Selecting a Batch of Queries Each Time........................ 136
References.. 146

5 Ensemble Learning with Uncertainty 149
5.1 Introduction to Ensemble Learning.................................. 149
5.1.1 Majority Voting and Weighted Majority Voting............... 150
5.1.2 Approach Based on Dempster–Shafer Theory of Evidence.. 151
5.1.3 Fuzzy Integral Ensemble Approach............................ 152
5.2 Bagging and Boosting... 153
5.2.1 Bagging Algorithm... 153
5.2.2 Boosting Algorithm... 154
5.3 Multiple Fuzzy Decision Tree Algorithm............................. 154
5.3.1 Induction of Multiple Fuzzy Decision Tree.................... 155
5.3.2 Experiment on Real Data Set.................................. 168
5.4 Fusion of Classifiers Based on Upper Integral........................ 170
5.4.1 Extreme Learning Machine..................................... 170
5.4.2 Multiple Classifier Fusion Based on Upper Integrals.......... 172
5.4.2.1 Upper Integral and Its Properties..................... 173
5.4.2.2 A Model of Classifier Fusion Based on Upper Integral.. 175
5.4.2.3 Experimental Results................................ 179
5.5 Relationship between Fuzziness and Generalization in Ensemble Learning... 186
5.5.1 Classification Boundary.. 186
5.5.1.1 Boundary and Its Estimation Given by a Learned Classifier.............................. 186
5.5.1.2 Two Types of Methods for Training a Classifier...... 188
5.5.1.3 Side Effect of Boundary and Experimental Verification.. 189

5.5.2 Fuzziness of Classifiers.. 192
5.5.2.1 Fuzziness of Classifier................................ 193
5.5.2.2 Relationship between Fuzziness and Misclassification...................................... 193
5.5.2.3 Relationship between Fuzziness and Classification Boundary.............................. 195
5.5.2.4 Divide and Conquer Strategy........................ 198
5.5.2.5 Impact of the Weighting Exponent m on the Fuzziness of Fuzzy K-NN Classifier................. 198
5.5.3 Relationship between Generalization and Fuzziness........... 199
5.5.3.1 Definition of Generalization and Its Elaboration.. 199
5.5.3.2 Classifier Selection.................................. 201
5.5.3.3 Explanation Based on Extreme (max/min) Fuzziness.. 202
5.5.3.4 Experimental Results................................ 205
References.. 216

Index **221**

Chapter 1

Uncertainty

Uncertainty is a common phenomenon in machine learning, which can be found in every phase of learning, such as data preprocessing, algorithm design, and model selection. The representation, measurement, and handling of uncertainty have a significant impact on the performance of a learning system. There are four common uncertainties in machine learning, that is, randomness [1], fuzziness [2], roughness [3], and nonspecificity [4]. In this chapter, we mainly introduce the first three kinds of uncertainty, briefly list the fourth uncertainty, and give a short discussion about the relationships among the four uncertainties.

1.1 Randomness

Randomness is a kind of objective uncertainty regarding random variables, while entropy is a measure of the uncertainty of random variables [5].

1.1.1 Entropy

Let X be a discrete random variable that takes values randomly from set $\mathcal{X}$, its probability mass function is $p(x) = Pr(X = x)$, $x \in \mathcal{X}$, denoted by $X \sim p(x)$. The definition of entropy of X is given as follows [5].

Definition 1.1 The entropy of X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x). \tag{1.1}$$

We can find from (1.1) that the entropy of X is actually a function of p; the following example can explicitly illustrate this point.

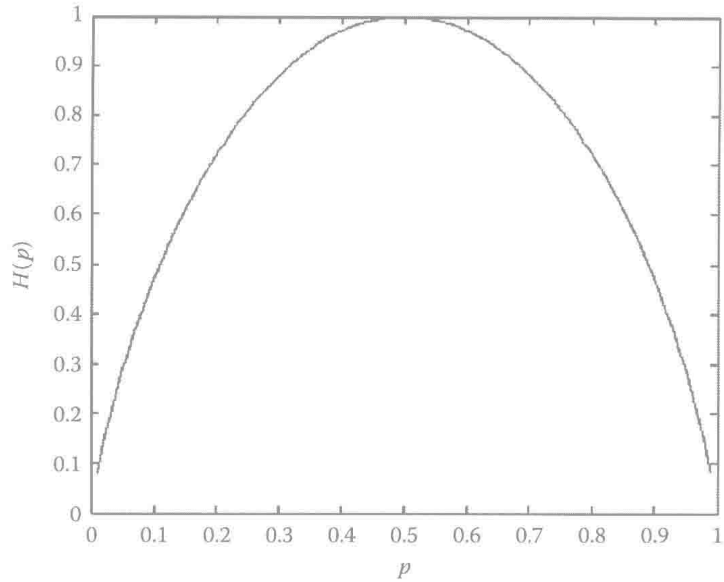


Figure 1.1 The relationship between entropy $H(p)$ and p.

Example 1.1 Let $\mathcal{X} = \{0, 1\}$, and $Pr(X = 1) = p$, $Pr(X = 0) = 1 - p$. According to Equation (1.1), the entropy of X is

$$H(X) = -p \times \log_2 p - (1 - p) \times \log_2(1 - p). \tag{1.2}$$

Obviously, $H(X)$ is a function of p. For convenience, we denote $H(X)$ as $H(p)$. The graph of the function $H(p)$ is shown in Figure 1.1. We can see from Figure 1.1 that $H(p)$ takes its maximum at $p = \frac{1}{2}$.

Example 1.2 Table 1.1 is a small discrete-valued data set with 14 instances. For attribute $X = Outlook$, $\mathcal{X} = \{\text{Sunny}, \text{Cloudy}, \text{Rain}\}$, we can find from Table 1.1 that $Pr(X = \text{Sunny}) = p_1 = \frac{5}{14}$, $Pr(X = \text{Cloudy}) = p_2 = \frac{4}{14}$, $Pr(X = \text{Rain}) = p_3 = \frac{5}{14}$. According to (1.1), we have

$$H(Outlook) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{4}{14}\log_2\frac{4}{14} - \frac{5}{14}\log_2\frac{5}{14} = 1.58.$$

Similarly, we have

$$H(Temperature) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14} - \frac{4}{14}\log_2\frac{4}{14} = 1.56.$$

$$H(Humidity) = -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14} = 1.00.$$

Table 1.1 A Small Data with 14 Instances

x	*Outlook*	*Temperature*	*Humidity*	*Wind*	*y (PlayTennis)*
x_1	Sunny	Hot	High	Weak	No
x_2	Sunny	Hot	High	Strong	No
x_3	Cloudy	Hot	High	Weak	Yes
x_4	Rain	Mild	High	Weak	Yes
x_5	Rain	Cool	Normal	Weak	Yes
x_6	Rain	Cool	Normal	Strong	No
x_7	Cloudy	Cool	Normal	Strong	Yes
x_8	Sunny	Mild	High	Weak	No
x_9	Sunny	Cool	Normal	Weak	Yes
x_{10}	Rain	Mild	Normal	Weak	Yes
x_{11}	Sunny	Mild	Normal	Strong	Yes
x_{12}	Cloudy	Mild	High	Strong	Yes
x_{13}	Cloudy	Hot	Normal	Weak	Yes
x_{14}	Rain	Mild	High	Strong	No

$$H(\mathit{Wind}) = -\frac{8}{14}\log_2\frac{8}{14} - \frac{6}{14}\log_2\frac{6}{14} = 0.99.$$

$$H(\mathit{PlayTennis}) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14} = 0.94.$$

It is worth noting that the probability mentioned earlier is approximated by its frequency, that is, the proportion of a value in all cases.

1.1.2 Joint Entropy and Conditional Entropy

Given two random variables X and Y, suppose that $(X, Y) \sim p(x, y)$. The joint entropy of X and Y can be defined as follows.

Definition 1.2 The joint entropy of (X, Y) is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log_2 p(x, y). \tag{1.3}$$

Given a random variable X, we can define the conditional entropy $H(Y|X)$ of a random variable Y.

Definition 1.3 Suppose $(X, Y) \sim p(x, y)$, $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= -\sum_{x\in\mathcal{X}} p(x)H(Y|X = x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log_2 p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log_2 p(y|x). \end{aligned} \tag{1.4}$$

It is necessary to note that generally $H(Y|X) \neq H(X|Y)$, but $H(X) - H(X|Y) = H(Y) - H(Y|X)$.

1.1.3 Mutual Information

Given two random variables X and Y, the mutual information of X and Y, denoted by $I(X; Y)$, is a measure of relevance of X and Y. We now give the definition of mutual information.

Definition 1.4 The mutual information of X and Y is defined as

$$I(X; Y) = H(X) - H(X|Y). \tag{1.5}$$

We can find from (1.5) that $I(X; Y)$ is the reduction of the uncertainty of X due to presentation of Y. By symmetry, it also follows that

$$I(X; Y) = H(Y) - H(Y|X). \tag{1.6}$$

Theorem 1.1 The mutual information and entropy have the following relationships [5]:

$$I(X; Y) = H(X) - H(X|Y); \tag{1.7}$$

$$I(X; Y) = H(Y) - H(Y|X); \tag{1.8}$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y); \tag{1.9}$$

$$I(X; Y) = I(Y, X); \tag{1.10}$$

$$I(X; X) = H(X). \tag{1.11}$$