Preface

This book is an outgrowth of ten years of research at the University of Florida Computational NeuroEngineering Laboratory (CNEL) in the general area of statistical signal processing and machine learning. One of the goals of writing the book is exactly to bridge the two fields that share so many common problems and techniques but are not yet effectively collaborating.

Unlike other books that cover the state of the art in a given field, this book cuts across engineering (signal processing) and statistics (machine learning) with a common theme: learning seen from the point of view of information theory with an emphasis on Renyi's definition of information. The basic approach is to utilize the information theory descriptors of entropy and divergence as nonparametric cost functions for the design of adaptive systems in unsupervised or supervised training modes. Hence the title: *Information-Theoretic Learning* (ITL). In the course of these studies, we discovered that the main idea enabling a synergistic view as well as algorithmic implementations, does not involve the conventional central moments of the data (mean and covariance). Rather, the core concept is the α-norm of the PDF, in particular its expected value ($\alpha = 2$), which we call the information potential. This operator and related nonparametric estimators link information theory, optimization of adaptive systems, and reproducing kernel Hilbert spaces in a simple and unconventional way.

Due to the pervasive nature of learning, the reading of the material requires prior basic knowledge on a broad set of subjects such as information theory, density estimation, adaptive filtering, pattern recognition, reproducing kernel Hilbert spaces (RKHS), and kernel machines. Because there are few researchers with such broad interests, the first chapter provides, in simple terms, the minimal foundations of information theory, adaptive filtering, and RKHS, while the appendix reviews density estimation. Once the reader is able to grasp these fundamentals, the book develops a nonparametric framework that is rich in understanding, setting the stage for the evolution of a new generation of algorithms of varying complexity. This book is therefore

useful for professionals who are interested in improving the performance of traditional algorithms as well as researchers who are interested in exploring new approaches to machine learning.

This thematic view of a broad research area is a double-sided sword. By using the same approach to treat many different problems it provides a unique and unifying perspective. On the other hand, it leaves out many competing alternatives and it complicates the evaluation of solutions. For this reason, we present many examples to illustrate and compare performance with conventional alternatives in the context of practical problems. To be more specific, the reader will find:

- Information-theoretic cost functions for linear and nonlinear adaptive filtering that have low complexity but are robust to impulsive noise, and extract valuable structure from the error signal
- Information-theoretic cost functions for classification and unsupervised learning and a new principle of self-organization
- A RKHS for ITL defined on a space of probability density functions that simplify statistical inference
- A new similarity function called correntropy that extends the conventional correlation

The book is organized as follows.

Chapter 1 covers the foundations of information theory, an overview of adaptive systems, and also the basic definitions of RKHS.

Chapter 2 presents the foundations of Renyi's entropy, divergence, mutual information, and their estimators based on the information potential. This is a foundational chapter, and readers should spend time understanding the concepts, and practicing with the algorithms for estimating the ITL descriptors directly from data. The chapter concludes with fast computational algorithms.

Chapter 3 develops the idea of error entropy criterion (EEC) minimization and its minimum error entropy (MEE) algorithm to adapt learning systems. An analysis of the cost function is undertaken and key properties of the error entropy criterion are presented. One of the main reasons why the EEC is useful in practical applications is its robustness to outliers. We establish the link between the EEC and Huber's robust statistics through a weighted least squares point of view. In so doing we define a new function called correntropy that can also be used to train adaptive filters and is easier to compute than EEC. Correntropy defines a metric in the data space and it is directly related to entropy. The chapter ends with a method to adapt the kernel size parameter in adaptive systems training.

Chapter 4 develops a set of algorithms to adapt linear filters using MEE. Basically all the practical gradient-based algorithms are covered: the MEE batch algorithm, the MEE recursive information potential that saves computation, the MEE stochastic information gradient (SIG) that mimics Widrow's LMS algorithm, the MEE self adjusting stepsize, and the normalized MEE. We also present a fixed-point algorithm (no stepsize) with higher complexity

but that is much faster because it explores second-order information content of the cost function. The chapter ends with a comparison with the error correntropy criterion, which has practical computational advantages.

Chapter 5 addresses filtering (regression) problems extending the training algorithms for nonlinear systems. We show how to integrate backpropagation with the error entropy costs, so the reader is able by the end of this chapter to train nonlinear systems with entropic costs. Incidentally, this is really the type of systems that benefit from the error entropy cost because most of the time the errors created are non-Gaussian. Comparisons with traditional mean square error cost are provided. A brief overview of advanced search methods with ITL algorithms is also presented.

Chapter 6 changes the focus to classification problems. The techniques necessary to train classifiers with MEE have already been established in Chapter 5, so this chapter addresses the usefulness of error entropy costs for classification, which is a harder problem than regression. Alternatively, nonparametric classifiers using a MAP approach can be easily implemented and work reasonably well in small-dimensional spaces. For classification, the idea of utilizing the dissimilarity between class labels and system output separately (instead of creating the error) is appealing because of Fano's bound. We extend the cost function to include the ITL divergence measures and quadratic mutual information, and show that this alternative cost function is beneficial not only to train classifiers but also for feature selection. The chapter ends with a proof that the classification error can be lower and upper bounded (i.e., can be bracketed) by Renyi's entropy for alpha greater and smaller than one, respectively.

Chapter 7 treats clustering (the simplest of unsupervised learning methods) using ITL divergence measures. First, we discuss the Cauchy-Schwarz divergence measure as a cost function for clustering, bringing out the nice feature that optimal clusters are not necessarily spherical. Then, a gradient descent algorithm is proposed to find the data partition that minimizes this clustering cost function, and its connection to spectral clustering and optimal graph cuts is established. Gaussian mean shift is also framed as the optimization of an ITL cost function. The chapter ends with a novel information cut algorithm for graph clustering.

Chapter 8 reviews several self-organizing principles based on information-theoretic concepts to show the importance of IT descriptors as cost functions for the optimal design of unsupervised learning systems. Then, a new self-organizing principle called the principle of relevant information is presented that yields as special cases, clustering, principal curves, and vector quantization. Finally, the ITL descriptors are utilized to implement the most common forms of self-organizing principles without assumptions about the data PDFs.

Chapter 9 defines a new reproducing kernel Hilbert space on the space of PDFs with an inner product defined by the cross information potential of ITL. This RKHS provides a functional analysis perspective of ITL and

helps us understand links between statistical inference and the RKHS defined for ITL. Moreover, we show the relationship between ITL descriptors and statistical operators used in machine learning in the RKHS defined by the kernel, including an interpretation of support vector machines.

Chapter 10 defines in the space of random variables a novel generalized correlation function named correntropy. We present many properties of correntropy to make clear its statistical meaning. Based on correntropy, we propose the correntropy coefficient that is bounded by unity and zero for independent random variables, unlike the conventional correlation coefficient. By defining the concept of parametric correntropy, we propose a new correntropy dependence measure that obeys most of Renyi's postulates for dependence. We illustrate the use of correntropy in statistical inference problems, such as matched filtering, tests of nonlinear coupling and as a dependent measure between random variables.

Chapter 11 extends the concept of correntropy to random processes. The name can be properly explained in this context because correntropy (built from correlation plus entropy) looks like correlation but the sum over the lags (or dimensions) is the information potential (the argument of the log of Renyi's entropy). We show that the autocorrentropy function is a positive definite kernel and, as such, defines a novel RKHS with interesting properties. It is possible to define a correntropy spectral density that provides a spectral representation that includes, for the first time, second- and higher-order moments of the random process. We end the chapter with a case study to exemplify how to transform optimal linear algorithms to the correntropy RKHS, and a few examples in speech processing, time series analysis, a correntropy Karhunen-Loeve transform and, object recognition.

The appendix completes the book with a review of kernel density estimation and Renyi's entropy estimation.

The author is conscious that such a vast coverage of topics imposes some compromises of breadth versus depth. To help readers with different backgrounds, profiles, and goals the following flowcharts help establish a road map for the book.

Adaptive Systems (including neural networks) Theme
Ch 1 → Ch 2 → Ch 3 → Ch 4 → Ch 5 → Ch 6 → Ch 7 → Ch 8
Unsupervised Learning Theme
Ch 1 → Ch 2 → Ch 7 → Ch 8
RKHS Theme
Ch 1 → Ch 2 → Ch 9 → Ch 10 → Ch 11
Statistical Signal Processing Theme
Ch 1 → Ch 2 → Ch 3 → Ch 9 → Ch 9 → Ch 11
Pattern Recognition Theme
Ch 1 → Ch 2 → Ch 5 → Ch 6 → Ch 7 → Ch 9 → Ch 10

The book is based on a large collection of journal papers and conference proceedings produced by an extraordinary group of PhD students and

CNEL visitors who were smart enough, knowledgeable enough, and brave enough to think outside the box and ask very pertinent questions about the principles ordinarily used in statistical signal processing and machine learning. Their names are: Deniz Erdogmus, Weifeng Liu, Dongxin Xu, Robert Jenssen, Jianwu Xu, Ignacio Santamaria, Kenneth Hild II, Jeongju Han, Kyu-Hwa Jeong, Sudhir Rao, Puskal Pokharel, Rodney Morejon, Antonio Paiva, Sohan Seth, Il Park, and Abhishek Singh.

To demonstrate my appreciation for their work I consider them my coauthors and list their names in the chapters where their main contributions are centered.

The author is grateful to the National Science Foundation, in particular the Electrical, Communications and Cyber Systems Division in the Engineering Directorate which has funded the great majority of this work and the above mentioned students.

Gainesville, Florida
August, 2009.

http://www.springer.com/978-1-4419-1569-6

Information Theoretic Learning
Renyi's Entropy and Kernel Perspectives
Principe, J.C.
2010, XIV, 448 p., Hardcover
ISBN: 978-1-4419-1569-6