
Semi-Supervised Learning

Olivier Chapelle
Bernhard Schölkopf
Alexander Zien

The MIT Press
Cambridge, Massachusetts
London, England

©2006 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

Typeset by the authors using L^AT_EX 2_ε

Printed and bound in the United States of America

Library of Congress Cataloging-in-Publication Data

Contents

Preface	1
1 Introduction to Semi-Supervised Learning	3
1.1 Supervised, Unsupervised, and Semi-Supervised Learning	3
1.2 When Can Semi-Supervised Learning Work?	6
1.3 Classes of Algorithms and Organization of This Book	9
I Generative Models	15
2 A Taxonomy for Semi-Supervised Learning Methods	17
<i>Matthias Seeger</i>	
2.1 The Semi-Supervised Learning Problem	17
2.2 Paradigms for Semi-Supervised Learning	19
2.3 Examples	23
2.4 Conclusions	31
3 Semi-Supervised Text Classification Using EM	33
<i>Kamal Nigam, Andrew McCallum, Tom Mitchell</i>	
3.1 Introduction	33
3.2 A Generative Model for Text	35
3.3 Experimental Results with Basic EM	40
3.4 Using a More Expressive Generative Model	43
3.5 Overcoming the Challenges of Local Maxima	48
3.6 Conclusions and Summary	53
4 Risks of Semi-Supervised Learning	55
<i>Fabio Cozman, Ira Cohen</i>	
4.1 Do Unlabeled Data Improve or Degrade Classification Performance? . . .	55
4.2 Understanding Unlabeled Data: Asymptotic Bias	57
4.3 The Asymptotic Analysis of Generative Semi-Supervised Learning	61
4.4 The Value of Labeled and Unlabeled Data	64
4.5 Finite Sample Effects	67
4.6 Model Search and Robustness	67
4.7 Conclusion	68

5	Probabilistic Semi-Supervised Clustering with Constraints	71
	<i>Sugato Basu, Mikhail Bilenko, Arindam Banerjee, Raymond Mooney</i>	
5.1	Introduction	72
5.2	HMMRF Model for Semi-Supervised Clustering	73
5.3	HMMRF-KMEANS Algorithm	78
5.4	Active Learning for Constraint Acquisition	90
5.5	Experimental Results	92
5.6	Related Work	96
5.7	Conclusions	98
II	Low-Density Separation	99
6	Transductive Support Vector Machines	101
	<i>Thorsten Joachims</i>	
6.1	Introduction	101
6.2	Transductive Support Vector Machines	104
6.3	Why Use Margin on the Test Set?	106
6.4	Experiments and Applications of TSVMs	107
6.5	Solving the TSVM Optimization Problem	110
6.6	Connection to Related Approaches	111
6.7	Summary and Conclusions	112
7	Semi-Supervised Learning Using Semi-Definite Programming	113
	<i>Tijl De Bie, Nello Cristianini</i>	
7.1	Relaxing SVM Transduction	113
7.2	An Approximation for Speedup	120
7.3	General Semi-Supervised Learning Settings	122
7.4	Empirical Results	123
7.5	Summary and Outlook	127
8	Gaussian Processes and the Null-Category Noise Model	131
	<i>Neil D. Lawrence, Michael I. Jordan</i>	
8.1	Introduction	131
8.2	The Noise Model	135
8.3	Process Model and Effect of the Null-Category	137
8.4	Posterior Inference and Prediction	139
8.5	Results	141
8.6	Discussion	143
9	Entropy Regularization	145
	<i>Yves Grandvalet, Yoshua Bengio</i>	
9.1	Introduction	145
9.2	Derivation of the Criterion	146
9.3	Optimization Algorithms	149

9.4	Related Methods	152
9.5	Experiments	154
9.6	Conclusion	159
9.7	Proof of Theorem 9.1	159
10	Data-Dependent Regularization	163
	<i>Adrian Corduneanu, Tommi Jaakkola</i>	
10.1	Introduction	163
10.2	Information Regularization on Metric Spaces	168
10.3	Information Regularization and Relational Data	176
10.4	Discussion	182
III	Graph-Based Methods	183
11	Label Propagation and Quadratic Criterion	185
	<i>Yoshua Bengio, Olivier Delalleau, Nicolas Le Roux</i>	
11.1	Introduction	185
11.2	Label Propagation on a Similarity Graph	186
11.3	Quadratic Cost Criterion	190
11.4	From Transduction to Induction	196
11.5	Incorporating Class Prior Knowledge	197
11.6	Curse of Dimensionality for Semi-Supervised Learning	198
11.7	Discussion	207
12	The Geometric Basis of Semi-Supervised Learning	209
	<i>Vikas Sindhwani, Misha Belkin, Partha Niyogi</i>	
12.1	Introduction	209
12.2	Incorporating Geometry in Regularization	212
12.3	Algorithms	215
12.4	Data-Dependent Kernels for Semi-Supervised Learning	221
12.5	Linear Methods for Large-Scale Semi-Supervised Learning	223
12.6	Connections to Other Algorithms and Related Work	224
12.7	Future Directions	226
13	Discrete Regularization	227
	<i>Dengyong Zhou, Bernhard Schölkopf</i>	
13.1	Introduction	227
13.2	Discrete Analysis	229
13.3	Discrete Regularization	234
13.4	Conclusion	238
14	Semi-Supervised Learning with Conditional Harmonic Mixing	239
	<i>Christopher J. C. Burges, John C. Platt</i>	
14.1	Introduction	239

14.2	Conditional Harmonic Mixing	243
14.3	Learning in CHM Models	244
14.4	Incorporating Prior Knowledge	248
14.5	Learning the Conditionals	249
14.6	Model Averaging	250
14.7	Experiments	250
14.8	Conclusions	259
IV	Change of Representation	263
15	Graph Kernels by Spectral Transforms	265
	<i>Xiaojin Zhu, Jaz Kandola, John Lafferty, Zoubin Ghahramani</i>	
15.1	The Graph Laplacian	266
15.2	Kernels by Spectral Transforms	268
15.3	Kernel Alignment	269
15.4	Optimizing Alignment Using QCQP for Semi-Supervised Learning . . .	270
15.5	Semi-Supervised Kernels with Order Constraints	271
15.6	Experimental Results	273
15.7	Conclusion	277
16	Spectral Methods for Dimensionality Reduction	279
	<i>Lawrence K. Saul, Kilian Q. Weinberger, Fei Sha, Jihun Ham, Daniel D. Lee</i>	
16.1	Introduction	279
16.2	Linear Methods	280
16.3	Graph-Based Methods	282
16.4	Kernel Methods	288
16.5	Discussion	291
17	Modifying Distances	295
	<i>Sajama, Alon Orlitsky</i>	
17.1	Introduction	295
17.2	Estimating DBD Metrics	298
17.3	Computing DBD Metrics	306
17.4	Semi-Supervised Learning Using Density-Based Metrics	313
17.5	Conclusions and Future Work	315
V	Semi-Supervised Learning in Practice	317
18	Large-Scale Algorithms	319
	<i>Olivier Delalleau, Yoshua Bengio, Nicolas Le Roux</i>	
18.1	Introduction	319
18.2	Cost Approximations	320
18.3	Subset Selection	323

18.4 Discussion	326
19 Semi-Supervised Protein Classification Using Cluster Kernels	329
<i>Jason Weston, Christina Leslie, Eugene Ie, William Stafford Noble</i>	
19.1 Introduction	329
19.2 Representations and Kernels for Protein Sequences	331
19.3 Semi-Supervised Kernels for Protein Sequences	334
19.4 Experiments	338
19.5 Discussion	345
20 Prediction of Protein Function from Networks	347
<i>Hyunjung Shin, Koji Tsuda</i>	
20.1 Introduction	347
20.2 Graph-Based Semi-Supervised Learning	350
20.3 Combining Multiple Graphs	351
20.4 Experiments on Function Prediction of Proteins	354
20.5 Conclusion and Outlook	359
21 Analysis of Benchmarks	363
21.1 The Benchmark	363
21.2 Application of SSL Methods	368
21.3 Results and Discussion	376
VI Perspectives	381
22 An Augmented PAC Model for Semi-Supervised Learning	383
<i>Maria-Florina Balcan, Avrim Blum</i>	
22.1 Introduction	384
22.2 A Formal Framework	386
22.3 Sample Complexity Results	389
22.4 Algorithmic Results	397
22.5 Related Models and Discussion	401
23 Metric-Based Approaches for Semi-Supervised Regression and Classification	405
<i>Dale Schuurmans, Finnegan Southey, Dana Wilkinson, Yuhong Guo</i>	
23.1 Introduction	405
23.2 Metric Structure of Supervised Learning	407
23.3 Model Selection	409
23.4 Regularization	419
23.5 Classification	428
23.6 Conclusion	431
24 Transductive Inference and Semi-Supervised Learning	437
<i>Vladimir Vapnik</i>	

24.1 Problem Settings	437
24.2 Problem of Generalization in Inductive and Transductive Inference	438
24.3 Structure of the VC Bounds and Transductive Inference	441
24.4 The Symmetrization Lemma and Transductive Inference	442
24.5 Bounds for Transductive Inference	443
24.6 The Structural Risk Minimization Principle for Induction and Transduction	444
24.7 Combinatorics in Transductive Inference	446
24.8 Measures of the Size of Equivalence Classes	446
24.9 Algorithms for Inductive and Transductive SVMs	448
24.10 Semi-Supervised Learning	454
24.11 Conclusion: Transductive Inference and the New Problems of Inference .	454
24.12 Beyond Transduction: Selective Inference	455
25 A Discussion of Semi-Supervised Learning and Transduction	457
References	463
Index	480
Notation	480
Notation and Symbols	481

Preface

During the last years, semi-supervised learning has emerged as an exciting new direction in machine learning research. It is closely related to profound issues of how to do inference from data, as witnessed by its overlap with *transductive inference* (the distinctions are yet to be made precise).

At the same time, dealing with the situation where relatively few labeled training points are available, but a large number of unlabeled points are given, it is directly relevant to a multitude of practical problems where it is relatively expensive to produce labeled data, e.g., the automatic classification of web pages. As a field, semi-supervised learning uses a diverse set of tools and illustrates, on a small scale, the sophisticated machinery developed in various branches of machine learning such as **kernel methods** or **Bayesian techniques**.

As we work on semi-supervised learning, we have been aware of the lack of an authoritative overview of the existing approaches. In a perfect world, such an overview should help both the practitioner and the researcher who wants to enter this area. A well researched monograph could ideally fill such a gap; however, the field of semi-supervised learning is arguably not yet sufficiently mature for this. Rather than writing a book which would come out in three years, we thus decided instead to provide an up-to-date edited volume, where we invited contributions by many of the leading proponents of the field. To make it more than a mere collection of articles, we have attempted to ensure that the chapters form a coherent whole and use consistent notation. Moreover, we have written a short introduction, a dialogue illustrating some of the ongoing debates in the underlying philosophy of the field, and we have organized and summarized a comprehensive *benchmark* of semi-supervised learning.

Benchmarks are helpful for the practitioner to decide which algorithm should be chosen for a given application. At the same time, they are useful for researchers to choose issues to study and further develop. By evaluating and comparing the performance of many of the presented methods on a set of eight benchmark problems, this book aims at providing guidance in this respect. The problems are designed to reflect and probe the different assumptions that the algorithms build on. All data sets can be downloaded from the book web page, which can be found at <http://www.kyb.tuebingen.mpg.de/ssl-book/>.

Finally, we would like to give thanks to everybody who contributed towards the success of this book project, in particular to Karin Bierig, Sabrina Nielebock, Bob Prior, to all chapter authors, and to the chapter reviewers.

1 Introduction to Semi-Supervised Learning

1.1 Supervised, Unsupervised, and Semi-Supervised Learning

In order to understand the nature of semi-supervised learning, it will be useful first to take a look at supervised and unsupervised learning.

1.1.1 Supervised and Unsupervised Learning

Traditionally, there have been two fundamentally different types of tasks in machine learning.

unsupervised
learning

The first one is *unsupervised learning*. Let $X = (x_1, \dots, x_n)$ be a set of n examples (or points), where $x_i \in \mathcal{X}$ for all $i \in [n] := \{1, \dots, n\}$. Typically it is assumed that the points are drawn i.i.d. (independently and identically distributed) from a common distribution on \mathcal{X} . It is often convenient to define the $(n \times d)$ -matrix $\mathbf{X} = (x_i^\top)_{i \in [n]}^\top$ that contains the data points as its rows. The goal of unsupervised learning is to find interesting structure in the data X . It has been argued that the problem of unsupervised learning is fundamentally that of estimating a density which is likely to have generated X . However, there are also weaker forms of unsupervised learning, such as quantile estimation, clustering, outlier detection, and dimensionality reduction.

supervised learning

The second task is *supervised learning*. The goal is to learn a mapping from x to y , given a training set made of pairs (x_i, y_i) . Here, the $y_i \in \mathcal{Y}$ are called the labels or targets of the examples x_i . If the labels are numbers, $\mathbf{y} = (y_i)_{i \in [n]}^\top$ denotes the column vector of labels. Again, a standard requirement is that the pairs (x_i, y_i) are sampled i.i.d. from some distribution which here ranges over $\mathcal{X} \times \mathcal{Y}$. The task is well defined, since a mapping can be evaluated through its predictive performance on test examples. When $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = \mathbb{R}^d$ (or more generally, when the labels are continuous), the task is called *regression*. Most of this book will focus on *classification* (there is some work on regression in chapter 23), i.e., the case where y takes values in a finite set (discrete labels).

generative methods

There are two families of algorithms for supervised learning. *Generative* algorithms try to model the

class-conditional density $p(x|y)$ by some unsupervised learning procedure.¹ A predictive density can then be inferred by applying Bayes theorem:

$$p(y|x) = \frac{p(x|y)p(y)}{\int_{\mathbf{y}} p(x|y)p(y)dy}. \quad (1.1)$$

discriminative
methods

In fact, $p(x|y)p(y) = p(x, y)$ is the joint density of the data, from which pairs (x_i, y_i) could be generated. *Discriminative* algorithms do not try to estimate how the x_i have been generated, but instead concentrate on estimating $p(y|x)$. Some discriminative methods even limit themselves to modeling whether $p(y|x)$ is greater than or less than 0.5; an example of this is the support vector machine (SVM). It has been argued that discriminative models are more directly aligned with the goal of supervised learning and therefore tend to be more efficient in practice. These two frameworks are discussed in more detail in sections 2.2.1 and 2.2.2.

1.1.2 Semi-Supervised Learning

standard setting of
SSL

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. **In addition to unlabeled data, the algorithm is provided with some supervision information – but not necessarily for all examples.** Often, this information will be the targets associated with some of the examples. In this case, the data set $X = (x_i)_{i \in [n]}$ can be divided into two parts: the points $X_l := (x_1, \dots, x_l)$, for which labels $Y_l := (y_1, \dots, y_l)$ are provided, and the points $X_u := (x_{l+1}, \dots, x_{l+u})$, the labels of which are not known. This is “standard” semi-supervised learning as investigated in this book; most chapters will refer to this setting.

SSL with
constraints

Other forms of partial supervision are possible. For example, there may be constraints such as “these points have (or do not have) the same target” [cf. Abu-Mostafa, 1995]. This more general setting is considered in chapter 5. The different setting corresponds to a different view of semi-supervised learning: In chapter 5, SSL is seen as unsupervised learning guided by constraints. In contrast, most other approaches see SSL as supervised learning with additional information on the distribution of the examples x . The latter interpretation seems to be more in line with most applications, where the goal is the same as in supervised learning: to predict a target value for a given x_i . However, this view does not readily apply if the number and nature of the classes are not known in advance but have to be inferred from the data. In contrast, SSL as unsupervised learning with constraints may still remain applicable in such situations.

transductive
learning
inductive learning

A problem related to SSL was introduced by Vapnik already several decades ago: so-called *transductive learning*. In this setting, one is given a (labeled) training set and an (unlabeled) test set. The idea of transduction is to perform predictions only for the test points. This is in contrast to *inductive learning*, where the goal is to output a prediction function which is defined on the entire space \mathcal{X} . Many methods described in this book

1. For simplicity, we are assuming that all distributions have densities, and thus we restrict ourselves to dealing with densities.

will be transductive; in particular, this is rather natural for inference based on graph representations of the data. This issue will be addressed again in section 1.2.4.

1.1.3 A Brief History of Semi-Supervised Learning

self-learning

Probably the earliest idea about using unlabeled data in classification is self-learning, which is also known as self-training, self-labeling, or decision-directed learning. This is a wrapper-algorithm that repeatedly uses a supervised learning method. It starts by training on the labeled data only. In each step a part of the unlabeled points is labeled according to the current decision function; then the supervised method is retrained using its own predictions as additional labeled points. This idea has appeared in the literature already for some time (e.g., Scudder [1965], Fralick [1967], Agrawala [1970]).

An unsatisfactory aspect of self-learning is that the effect of the wrapper depends on the supervised method used inside it. If self-learning is used with empirical risk minimization and 1-0-loss, the unlabeled data will have no effect on the solution at all. If instead a margin maximizing method is used, as a result the decision boundary is pushed away from the unlabeled points (cf. chapter 6). In other cases it seems to be unclear what the self-learning is really doing, and which assumption it corresponds to.

transductive inference

Closely related to semi-supervised learning is the concept of transductive inference, or transduction, pioneered by Vapnik [Vapnik and Chervonenkis, 1974, Vapnik and Sterin, 1977]. In contrast to inductive inference, no general decision rule is inferred, but only the labels of the unlabeled (or test) points are predicted. An early instance of transduction (albeit without explicitly considering it as a concept) was already proposed by Hartley and Rao [1968]. They suggested a combinatorial optimization on the labels of the test points in order to maximize the likelihood of their model.

mixture of Gaussians

It seems that semi-supervised learning really took off in the 1970s when the problem of estimating the Fisher linear discriminant rule with unlabeled data was considered [Hosmer Jr., 1973, McLachlan, 1977, O'Neill, 1978, McLachlan and Ganesalingam, 1982]. More precisely, the setting was in the case where each class-conditional density is Gaussian with equal covariance matrix. The likelihood of the model is then maximized using the labeled and unlabeled data with the help of an iterative algorithm such as the expectation-maximization (EM) algorithm [Dempster et al., 1977]. Instead of a mixture of Gaussians, the use of a mixture of multinomial distributions estimated with labeled and unlabeled data has been investigated in [Cooper and Freeman, 1970].

Later, this one component per class setting has been extended to several components per class [Shahshahani and Landgrebe, 1994] and further generalized by Miller and Uyar [1997].

theoretical analysis

Learning rates in a probably approximately correct (PAC) framework [Valiant, 1984] have been derived for the semi-supervised learning of a mixture of two Gaussians by Ratsaby and Venkatesh [1995]. In the case of an *identifiable* mixture, Castelli and Cover [1995] showed that with an infinite number of unlabeled points, the probability of error has an exponential convergence (w.r.t. the number of labeled examples) to the Bayes risk. Identifiable means that given $P(\mathbf{x})$, the decomposition in $\sum_y P(y)P(\mathbf{x}|y)$ is unique. This seems a relatively strong assumption, but it is satisfied, for instance, by mixtures

of Gaussians. Related is the analysis in [Castelli and Cover, 1996] in which the class-conditional densities are known but the class priors are not.

text applications Finally, the interest in semi-supervised learning increased in the 1990s, mostly due to applications in natural language problems and text classification [Yarowsky, 1995, Nigam et al., 1998, Blum and Mitchell, 1998, Collins and Singer, 1999, Joachims, 1999].

Note that, to our knowledge, Merz et al. [1992] were the first to use the term “semi-supervised” for classification with both labeled and unlabeled data. It has in fact been used before, but in a different context than what is developed in this book; see, for instance, [Board and Pitt, 1989].

1.2 When Can Semi-Supervised Learning Work?

A natural question arises: is semi-supervised learning meaningful? More precisely: in comparison with a supervised algorithm that uses only labeled data, can one hope to have a more accurate prediction by taking into account the unlabeled points? As you may have guessed from the size of the book in your hands, in principle the answer is “yes.” However, there is an important prerequisite: that the distribution of examples, which the unlabeled data will help elucidate, be relevant for the classification problem.

In a more mathematical formulation, one could say that the knowledge on $p(x)$ that one gains through the unlabeled data has to carry information that is useful in the inference of $p(y|x)$. If this is not the case, semi-supervised learning will not yield an improvement over supervised learning. It might even happen that using the unlabeled data degrades the prediction accuracy by misleading the inference; this effect is investigated in detail in chapter 4.

smoothness assumption One should thus not be too surprised that for semi-supervised learning to work, certain *assumptions* will have to hold. In this context, note that plain supervised learning also has to rely on assumptions. In fact, chapter 22 discusses a way of formalizing assumptions of the kind given below within a PAC-style framework. One of the most popular such assumptions can be formulated as follows.

Smoothness assumption of supervised learning: ² If two points x_1, x_2 are close, then so should be the corresponding outputs y_1, y_2 .

Clearly, without such assumptions, it would never be possible to generalize from a finite training set to a set of possibly infinitely many unseen test cases.

2. Strictly speaking, this assumption only refers to continuity rather than smoothness; however, the term *smoothness* is commonly used, possibly because in regression estimation y is often modeled in practice as a smooth function of x .

1.2.1 The Semi-Supervised Smoothness Assumption

semi-supervised
smoothness
assumption

We now propose a generalization of the smoothness assumption that is useful for semi-supervised learning; we thus call it the “semi-supervised smoothness assumption”. While in the supervised case according to our prior beliefs the output varies smoothly with the distance, we now also take into account the density of the inputs. The assumption is that the label function is smoother in high-density regions than in low-density regions:

Semi-supervised smoothness assumption: If two points x_1, x_2 in a high-density region are close, then so should be the corresponding outputs y_1, y_2 .

Note that by transitivity, this assumption implies that if two points are linked by a path of high density (e.g., if they belong to the same cluster), then their outputs are likely to be close. If, on the other hand, they are separated by a low-density region, then their outputs need not be close.

Note that the semi-supervised smoothness assumption applies to both regression and classification. In the next section, we will show that in the case of classification, it reduces to assumptions commonly used in SSL. At present, it is less clear how useful the assumption is for regression problems. As an alternative, chapter 23 proposes a way to use unlabeled data for model selection that applies to both regression and classification.

1.2.2 The Cluster Assumption

cluster assumption

Suppose we knew that the points of each class tended to form a cluster. Then the unlabeled data could aid in finding the boundary of each cluster more accurately: one could run a clustering algorithm and use the labeled points to assign a class to each cluster. That is in fact one of the earliest forms of semi-supervised learning (see chapter 2). The underlying, now classical, assumption may be stated as follows:

Cluster Assumption: If points are in the same cluster, they are likely to be of the same class.

This assumption may be considered reasonable on the basis of the sheer existence of classes: if there is a densely populated continuum of objects, it may seem unlikely that they were ever distinguished into different classes.

Note that the cluster assumption does not imply that each class forms a single, compact cluster: it only means that, usually, we do not observe objects of two distinct classes in the same cluster.

The cluster assumption can easily be seen as a special case of the above-proposed semi-supervised smoothness assumption, considering that clusters are frequently defined as being sets of points that can be connected by short curves which traverse only high-density regions.

low density
separation

The cluster assumption can be formulated in an equivalent way:

Low Density Separation: The decision boundary should lie in a low-density region.

The equivalence is easy to see: A decision boundary in a high-density region would cut a cluster into two different classes; many objects of different classes in the same cluster would require the decision boundary to cut the cluster, i.e., to go through a high-density region.

Although the two formulations are conceptually equivalent, they can inspire different algorithms, as we will argue in section 1.3. The low-density version also gives additional intuition why the assumption is sensible in many real-world problems. Consider digit recognition, for instance, and suppose that one wants to learn how to distinguish a handwritten digit “0” against digit “1.” A sample point taken exactly from the decision boundary will be between a 0 and a 1, most likely a digit looking like a very elongated zero. But the probability that someone wrote this “weird” digit is very small.

1.2.3 The Manifold Assumption

manifold
assumption

A different but related assumption that forms the basis of several semi-supervised learning methods is the manifold assumption:

Manifold assumption: The (high-dimensional) data lie (roughly) on a low-dimensional manifold.

curse of
dimensionality

How can this be useful? A well-known problem of many statistical methods and learning algorithms is the so-called curse of dimensionality (cf. section 11.6.2). It is related to the fact that volume grows exponentially with the number of dimensions, and an exponentially growing number of examples is required for statistical tasks such as the reliable estimation of densities. This is a problem that directly affects generative approaches that are based on density estimates in input space. A related problem of high dimensions, which may be more severe for discriminative methods, is that pairwise distances tend to become more similar, and thus less expressive.

If the data happen to lie on a low-dimensional manifold, however, then the learning algorithm can essentially operate in a space of corresponding dimension, thus avoiding the curse of dimensionality.

As above, one can argue that algorithms working with manifolds may be seen as approximately implementing the semi-supervised smoothness assumption: such algorithms use the metric of the manifold for computing geodesic distances. If we view the manifold as an approximation of the high-density regions, then it becomes clear that in this case, the semi-supervised smoothness assumption reduces to the standard smoothness assumption of supervised learning, applied on the manifold.

Note that if the manifold is embedded into the high-dimensional input space in a curved fashion (i.e., it is not just a subspace), geodesic distances differ from those in the input space. By ensuring more accurate density estimates and more appropriate distances, the manifold assumption may be useful for classification as well as for regression.

1.2.4 Transduction

As mentioned before, some algorithms naturally operate in a transductive setting. According to the philosophy put forward by Vapnik, high-dimensional estimation problems should attempt to follow the following principle:

Vapnik's principle: When trying to solve some problem, one should not solve a more difficult problem as an intermediate step.

Consider as an example supervised learning, where predictions of labels y corresponding to some objects x are desired. Generative models estimate the density of x as an intermediate step, while discriminative methods directly estimate the labels.

In a similar way, if label predictions are only required for a given test set, transduction can be argued to be more direct than the corresponding induction: while an inductive method infers a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ on the entire space \mathcal{X} , and afterward returns the evaluations $f(x_i)$ at the test points, transduction consists of directly estimating the finite set of test labels, i.e., a function $f : X_u \rightarrow \mathcal{Y}$ only defined on the test set. Note that transduction (as defined in this book) is not the same as SSL: some semi-supervised algorithms are transductive, but others are inductive.

Now suppose we are given a transductive algorithm which produces a solution superior to an inductive algorithm trained on the same labeled data (but discarding the unlabeled data). Then the performance difference might be due to one of the following two points (or a combination thereof):

1. transduction follows Vapnik's principle more closely than induction does, or
2. the transductive algorithm takes advantage of the unlabeled data in a way similar to semi-supervised learning algorithms.

There is ample evidence for improvements being due to the second of these points. We are presently not aware of empirical results that selectively support the first point. In particular, the evaluation of the benchmark associated with this book (chapter 21) does not seem to suggest a systematic advantage of transductive methods. However, the properties of transduction are still the topic of debate, and chapter 25 tries to present different opinions.

1.3 Classes of Algorithms and Organization of This Book

Although many methods were not explicitly derived from one of the above assumptions, most algorithms can be seen to correspond to or implement one or more of them. We try to organize the semi-supervised learning methods presented in this book into four classes that roughly correspond to the underlying assumption. Although the classification is not always unique, we hope that this organization makes the book and its contents more accessible to the reader, by providing a guiding scheme.

For the same reason, this book is organized in "parts." There is one part for each class of SSL algorithms and an extra part focusing on generative approaches. Two further parts

are devoted to applications and perspectives of SSL. In the following we briefly introduce the ideas covered by each book part.

1.3.1 Generative Models

Part I presents history and state of the art of SSL with generative models. Chapter 2 starts with a thorough review of the field.

mixture models

Inference using a generative model involves the estimation of the conditional density $p(x|y)$. In this setting, any additional information on $p(x)$ is useful. As a simple example, assume that $p(x|y)$ is Gaussian. Then one can use the EM algorithm to find the parameters of the Gaussian corresponding to each class. The only difference to the standard EM algorithm as used for clustering is that the “hidden variable” associated with any labeled example is actually not hidden, but it is known and equals its class label. It implements the cluster assumption (cf. section 2.2.1), since a given cluster belongs to only one class.

This small example already highlights different interpretations of semi-supervised learning with a generative model:

- It can be seen as classification with additional information on the marginal density.
- It can be seen as clustering with additional information. In the standard setting, this information would be the labels of a subset of points, but it could also come in the more general form of constraints. This is the topic of chapter 5.

A strength of the generative approach is that knowledge of the structure of the problem or the data can naturally be incorporated by modeling it. In chapter 3, this is demonstrated for the application of the EM algorithm to text data. It is observed that, when modeling assumptions are not correct, unlabeled data can decrease prediction accuracy. This effect is investigated in depth in chapter 4.

data-dependent
priors

In statistical learning, before performing inference, one chooses a class of functions, or a prior over functions. One has to choose it according to what is known in advance about the problem. In the semi-supervised learning context, if one has some ideas about what the structure of the data tells about the target function, the choice of this prior can be made more precise after seeing the unlabeled data: one could typically put a higher prior probability on functions that satisfy the cluster assumption. From a theoretical point, this is a natural way to obtain bounds for semi-supervised learning as explained in chapter 22.

1.3.2 Low-Density Separation

Part II of this book aims at describing algorithms which try to directly implement the low-density separation assumption by pushing the decision boundary away from the unlabeled points.

transductive SVM
(TSVM)

The most common approach to achieving this goal is to use a maximum margin algorithm such as support vector machines. The method of maximizing the margin for unlabeled as well as labeled points is called the transductive SVM (TSVM). However, the corresponding problem is nonconvex and thus difficult to optimize.

One optimization algorithm for the TSVM is presented in chapter 6. Starting from the

SVM solution as trained on the labeled data only, the unlabeled points are labeled by SVM predictions, and the SVM is retrained on all points. This is iterated while the weight of the unlabeled points is slowly increased. Another possibility is the semi-definite programming SDP relaxation suggested in chapter 7.

Two alternatives to the TSVM are then presented that are formulated in a probabilistic and in an information theoretic framework, respectively. In chapter 8, **binary Gaussian process classification** is augmented by the introduction of a null class that occupies the space between the two regular classes. As an advantage over the TSVM, this allows for probabilistic outputs.

This advantage is shared by the **entropy minimization** presented in chapter 9. It encourages the class-conditional probabilities $P(y|x)$ to be close to either 1 or 0 at labeled and unlabeled points. As a consequence of the smoothness assumption, the probability will tend to be close to 0 or 1 throughout any high-density region, while class boundaries correspond to intermediate probabilities.

A different way of using entropy or information is the data-dependent regularization developed in chapter 10. As compared to the TSVM, this seems to implement the low-density separation even more directly: the standard squared-norm regularizer is multiplied by a term reflecting the density close to the decision boundary.

1.3.3 Graph-Based Methods

During the last couple of years, the most active area of research in semi-supervised learning has been in graph-based methods, which are the topic of part III of this book. The common denominator of these methods is that the data are represented by the nodes of a graph, the edges of which are labeled with the pairwise distances of the incident nodes (and a missing edge corresponds to infinite distance). If the distance of two points is computed by minimizing the aggregate path distance over all paths connecting the two points, this can be seen as an approximation of the geodesic distance of the two points with respect to the manifold of data points. Thus, graph methods can be argued to build on the manifold assumption.

Most graph methods refer to the graph by utilizing the graph Laplacian. Let $g = (V, E)$ be a graph with real edge weights given by $w : E \rightarrow \mathbb{R}$. Here, the weight $w(e)$ of an edge e indicates the similarity of the incident nodes (and a missing edge corresponds to zero similarity). Now the weighted adjacency matrix (or weight matrix, for short) \mathbf{W} of the graph $g = (V, E)$ is defined by

$$\mathbf{W}_{ij} := \begin{cases} w(e) & \text{if } e = (i, j) \in E \\ 0 & \text{if } e = (i, j) \notin E \end{cases} \quad (1.2)$$

The diagonal matrix \mathbf{D} defined by $\mathbf{D}_{ii} := \sum_j \mathbf{W}_{ij}$ is called the degree matrix of g . Now there are different ways of defining the graph Laplacian, the two most prominent of which

are the normalized graph Laplacian, \mathcal{L} , and the unnormalized graph Laplacian, L :

$$\begin{aligned}\mathcal{L} &:= \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}, \\ L &:= \mathbf{D} - \mathbf{W}.\end{aligned}\tag{1.3}$$

Many graph methods that penalize nonsmoothness along the edges of a weighted graph can in retrospect be seen as different instances of a rather general family of algorithms, as is outlined in chapter 11. Chapter 13 takes a more theoretical point of view, and transfers notions of smoothness from the continuous case onto graphs as the discrete case. From that, it proposes different regularizers based on a graph representation of the data.

Usually the prediction consists of labels for the unlabeled nodes. For this reason, this kind of algorithm is intrinsically transductive, i.e., it returns only the value of the decision function on the unlabeled points and not the decision function itself. However, there has been recent work in order to extend graph-based methods to produce inductive solutions, as discussed in chapter 12.

Information propagation on graphs can also serve to improve a given (possibly strictly supervised) classification, taking unlabeled data into account. Chapter 14 presents a probabilistic method for using directed graphs in this manner.

Often the graph g is constructed by computing similarities of objects in some other representation, e.g., using a kernel function on Euclidean data points. But sometimes the original data already have the form of a graph. Examples include the linkage pattern of webpages and the interactions of proteins (see chapter 20). In such cases, the directionality of the edges may be important.

1.3.4 Change of Representation

The topic of part IV is algorithms that are not intrinsically semi-supervised, but instead perform two-step learning:

1. Perform an unsupervised step on all data, labeled and unlabeled, but ignoring the available labels. This can, for instance, be a change of representation, or the construction of a new metric or a new kernel.
2. Ignore the unlabeled data and perform plain supervised learning using the new distance, representation, or kernel.

This can be seen as direct implementation of the semi-supervised smoothness assumption, since the representation is changed in such a way that small distances in high-density regions are conserved.

Note that the graph-based methods (part III) are closely related to the ones presented in this part: the very construction of the graph from the data can be seen as an unsupervised change of representation. Consequently, the first chapter of part IV, chapter 15, discusses spectral transforms of such graphs in order to build kernels. Spectral methods can also be used for nonlinear dimensionality reduction, as extended in chapter 16. Furthermore, in chapter 17, metrics derived from graphs are investigated, for example, those derived from shortest paths.

1.3.5 Semi-Supervised Learning in Practice

Semi-supervised learning will be most useful whenever there are far more unlabeled data than labeled. This is likely to occur if obtaining data points is cheap, but obtaining the labels costs a lot of time, effort, or money. This is the case in many application areas of machine learning, for example:

- In speech recognition, it costs almost nothing to record huge amounts of speech, but labeling it requires some human to listen to it and type a transcript.
- Billions of webpages are directly available for automated processing, but to classify them reliably, humans have to read them.
- Protein sequences are nowadays acquired at industrial speed (by genome sequencing, computational gene finding, and automatic translation), but to resolve a three-dimensional (3D) structure or to determine the functions of a single protein may require years of scientific work.

Webpage classification is introduced in chapter 3 in the context of generative models.

Since unlabeled data carry less information than labeled data, they are required in large amounts in order to increase prediction accuracy significantly. This implies the need for fast and efficient SSL algorithms. Chapters 18 and 19 present two approaches to dealing with huge numbers of points. In chapter 18 methods are developed for speeding up the label propagation methods introduced in chapter 11. In chapter 19 cluster kernels are shown to be an efficient SSL method.

Chapter 19 also presents the first of two approaches to an important bioinformatics application of semi-supervised learning: the classification of protein sequences. While here the predictions are based on the protein sequences themselves, Chapter 20 moves on to a somewhat more complex setting: The information is here assumed to be present in the form of graphs that characterize the interactions of proteins. Several such graphs exist and have to be combined in an appropriate way.

This book part concludes with a very practical chapter: the presentation and evaluation of the benchmarks associated with this book (chapter 21). It is intended to give hints to the practitioner on how to choose suitable methods based on the properties of the problem.

1.3.6 Outlook

The last part of the book, part VI, is devoted to some of the most interesting directions of ongoing research in SSL.

Until now this book has mostly restricted itself to classification. Chapter 23 introduces another approach to SSL that is suited for both classification and regression, and derives algorithms from it. Interestingly it seems not to require the assumptions proposed in chapter 1.

Further, this book mostly presented *algorithms* for SSL. While the assumptions discussed above supply some intuition on when and why SSL works, and chapter 4 investigates when and why it can fail, it would clearly be more satisfactory to have a thorough

theoretical understanding of SSL in total. Chapter 22 offers a PAC-style framework that yields error bounds for SSL problems.

In chapter 24 inductive semi-supervised learning and transduction are compared in terms of Vapnik-Chervonenkis (VC) bounds and other theoretical and philosophical concepts.

The book closes with a hypothetical discussion (chapter 25) between three machine learning researchers on the relationship of (and the differences between) semi-supervised learning and transduction.

References

- B. Abboud, F. Davoine, and M. Dang. Expressive face recognition and synthesis. In *Computer Vision and Pattern Recognition Workshop*, volume 5, page 54, 2003.
- N. Abe, J. Takeuchi, and M. Warmuth. Polynomial learnability of stochastic rules with respect to the KL-divergence and quadratic distance. *IEICE, Transactions on Info and Systems*, E84-D(3):299–316, March 2001.
- Y. S. Abu-Mostafa. Machines that learn from hints. *Scientific American*, 272(4):64–69, 1995.
- A. K. Agrawala. Learning with a probabilistic teacher. *IEEE Transactions on Information Theory*, 16:373–379, 1970.
- A. Agresti. *Categorical Data Analysis*. John Wiley and Sons, 2002.
- H. Akaike. Statistical predictor information. *Annals of the Institute of Statistical Mathematics*, 22:203–271, 1970.
- H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.
- B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*. New York, Garland Science Publishing, 1998.
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- Y. Altun, D. McAllester, and M. Belkin. Maximum margin semi-supervised learning for structured variables. In *NIPS*, volume 18, 2005.
- M. R. Amini and P. Gallinari. Semi-supervised logistic regression. In *Fifteenth European Conference on Artificial Intelligence*, pages 390–394, 2002.
- J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66:17–26, 1979.
- M. Anjos. *New Convex Relaxations for the Maximum Cut and VLSI Layout Problems*. Phd thesis, Waterloo University, Waterloo, Canada, 2001.
- F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML)*. ACM Press, 2004.
- R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, 1999.
- M.-F. Balcan and A. Blum. An augmented PAC model for semi-supervised learning. *Manuscript*, 2005.
- M.-F. Balcan, A. Blum, and K. Yang. Co-training and Expansion: Towards bridging theory and practice. In *NIPS*, 2004.
- S. Baluja. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled examples. In *Advances in Neural Information Processing Systems 11*, pages 854–860, 1999.
- A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, 2005a.
- A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 2005b. To appear.
- N. Bansal, A. L. Blum, and S. Chawla. Correlation clustering. In *FOCS*, pages 238–247, 2002.
- A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proceedings of ICML*, pages 11–18, Washington, DC, 2003.
- P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities risk bounds and structural results. *Journal of Machine Learning Research*, pages 463–482, 2002.

- S. Basu, A. Banerjee, and R. J. Mooney. Semi-supervised clustering by seeding. In *Proceedings of ICML*, pages 19–26, 2002.
- S. Basu, A. Banerjee, and R. J. Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of SIAM SDM*, 2004a.
- S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of ACM SIGKDD*, pages 59–68, Seattle, WA, 2004b.
- E. B. Baum. Polynomial time algorithms for learning neural nets. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 258 – 272, 1990.
- S. Becker and G. E. Hinton. A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355:161–163, 1992.
- M. Belkin. *Problems of Learning on Manifolds*. PhD thesis, Department of Mathematics, University of Chicago, 2003.
- M. Belkin, I. Matveeva, and P. Niyogi. Regression and regularization on large graphs. In *Proceedings of the Seventeenth Annual Conference on Learning Theory*, 2004a.
- M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Proceedings of the Seventeenth Annual Conference on Computational Learning Theory (COLT 2004)*, pages 624–638, Banff, Canada, 2004b.
- M. Belkin and P. Niyogi. Semi-supervised learning on manifolds. In *Neural Information Processing Systems (NIPS)*, 2002.
- M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003a.
- M. Belkin and P. Niyogi. Using manifold structure for partially labeled classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA, 2003b. MIT Press.
- M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago, 2004c.
- M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 17–24. Society for Artificial Intelligence and Statistics, 2005. (Available electronically at <http://www.gatsby.ucl.ac.uk/aistats/>).
- R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ, 1961.
- S. Ben-David, A. Itai, and E. Kushilevitz. Learning by distances. *Information and Computation*, 117(2):240–250, 1995.
- A. Ben-Hur and D. Brutlag. Remote homology detection: A motif based approach. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 2003.
- G. M. Benedek and A. Itai. Learnability with respect to a fixed distribution. *Theoretical Computer Science*, 86: 377–389, 1991.
- Y. Bengio and N. Chapados. Extensions to metric based model selection. *Journal of Machine Learning Research*, 3:1209–1227, 2003. ISSN 1533-7928.
- Y. Bengio, O. Delalleau, and N. Le Roux. The curse of dimensionality for local kernel machines. Technical Report 1258, Département d’informatique et recherche opérationnelle, Université de Montréal, 2005.
- Y. Bengio, O. Delalleau, and N. Le Roux. The curse of highly variable functions for local kernel machines. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006a.
- Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004a.
- Y. Bengio, H. Larochelle, and P. Vincent. Non-local manifold parzen windows. In *Advances in Neural Information Processing Systems 18*. MIT Press, Cambridge, MA, 2006b.
- Y. Bengio and M. Monperrus. Non-local manifold tangent learning. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.
- Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004b.
- K. P. Bennett and A. Demiriz. Semi-supervised support vector machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *NIPS*, volume 11, pages 368–374. MIT Press, 1999.
- K. P. Bennett, A. Demiriz, and R. Maclin. Exploiting unlabeled data in ensemble methods. In *SIGKDD*

- International Conference on Knowledge Discovery and Data Mining*, 2002.
- J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1985.
- R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, pages 51–58, 1966.
- M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Manuscript*, 2000.
- J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B (Methodological)*, 48(3):259–302, 1986.
- A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor, 2004. Submitted for publication.
- T. De Bie and N. Cristianini. Convex transduction with the normalized cut. *Manuscript*, 2004.
- M. Bilenko and S. Basu. A comparison of inference techniques for semi-supervised clustering with hidden markov random fields. In *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields (SRL-2004)*, Banff, Canada, 2004.
- M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of ICML*, pages 81–88, Banff, Canada, 2004.
- M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of ACM SIGKDD*, pages 39–48, Washington, DC, 2003.
- C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.
- R. E. Blahut. Computation of channel capacity and rate distortion functions. In *IEEE Trans. Inform. Theory*, volume 18, pages 460–473, July 1972.
- A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 19–26, 2001.
- A. Blum, A. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22:35–52, 1998.
- A. Blum and R. Kannan. Learning an intersection of k halfspaces over a uniform distribution. *Journal of Computer and Systems Sciences*, 54(2):371–380, 1997.
- A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *ICML*, 2004.
- A. Blum and J. C. Langford. PAC-MDL bounds. In *Proceedings of COLT*, 2003.
- A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.
- A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.
- R. Board and L. Pitt. Semi-supervised learning. *Machine Learning*, 4(1):41–65, 1989.
- B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.
- L. Bottou and V. Vapnik. Local learning algorithms. *Neural Computation*, 4(6):888–900, 1992.
- S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of recent advances. *Manuscript*, 2004.
- S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.
- O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.
- O. Bousquet and D. Herrmann. On the complexity of learning the kernel matrix. In *Advances in NIPS 14*, 2002.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge UK, 2004.
- M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, 1999.
- M. Brand. Nonlinear dimensionality reduction by kernel eigenmaps. In *International Joint Conference on Artificial Intelligence*, 2003.
- M. Brand. From subspaces to submanifolds. In *Proceedings of the British Machine Vision Conference*, London, England, 2004.
- L. Breiman. Bagging predictors. *Machine Learning*, 24:123–40, 1996.

- S. E. Brenner, P. Koehl, and M. Levitt. The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Research*, 28:254–256, 2000.
- R. Bruce. Semi-supervised learning using prior probabilities and EM. In *IJCAI-01 Workshop on Text Learning: Beyond Supervision*, August 2001.
- W. L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2: 159–225, 1994.
- C. J. C. Burges. Geometric methods for feature extraction and dimensional reduction. In L. Rokach and O. Maimon, editors, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005.
- V. Castelli. *The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition*. PhD thesis, Stanford University, December 1994.
- V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16: 105–111, 1995.
- V. Castelli and T. M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, November 1996.
- G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3):315–332, 1992.
- O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.
- O. Chapelle and V. Vapnik. Model selection for support vector machines. In *Advances in NIPS 12*, 2000.
- O. Chapelle, V. Vapnik, and Y. Bengio. Model selection for small sample regression. *Machine Learning*, 48(1-3): 9–23, 2002.
- O. Chapelle, V. Vapnik, and J. Weston. Transductive inference for estimating values of functions. In *Neural Information Processing Systems (NIPS)*, 1999.
- O. Chapelle, J. Weston, L. Bottou, and V. Vapnik. Vicinal risk minimization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 416–422, Cambridge, MA, 2001. MIT Press.
- O. Chapelle, J. Weston, and B. Schölkopf. Cluster kernels for semi-supervised learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems (NIPS) 15*, pages 585–592. MIT Press, 2003.
- O. Chapelle and A. Zien. Semi-supervised classification by low density separation. In *Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.
- J. Cheng, D. Bell, and W. Liu. Learning belief networks from data: An information theory based approach. In *International Conference on Information and Knowledge Management*, pages 325–331, 1997.
- V. Cherkassky and F. Mulier. *Learning from Data: Concepts, Theory, and Methods*. Wiley, New York, 1998.
- V. Cherkassky, F. Mulier, and V. Vapnik. Comparison of VC-method with classical methods for model selection. In *Proceedings World Congress on Neural Networks*, pages 957–962, 1997.
- F. R. K. Chung. *Spectral Graph Theory*. Number 92 in Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI, 1997.
- I. Cohen, F. Cozman, N. Sebe, M. C. Cirelo, and T. Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1568, 2004.
- I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang. Learning bayesian network classifiers for facial expression recognition using both labeled and unlabeled data. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
- D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University, 2003.
- R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucke. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102:7426–7431, 2005.
- M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 189–196, 1999.
- R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large Scale Transductive SVMs. November 2005. In preparation. <http://www.kyb.tuebingen.mpg.de/bs/people/fabee/transduction.html>.
- D. B. Cooper and J. H. Freeman. On the asymptotic improvement in the outcome of supervised learning provided

- by additional nonsupervised learning. *IEEE Transactions on Computers*, C-19(11):1055–1063, November 1970.
- A. Corduneanu and T. Jaakkola. Continuation methods for mixing heterogeneous sources. In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2002.
- A. Corduneanu and T. Jaakkola. On information regularization. In *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.
- A. Corduneanu and T. Jaakkola. Distributed information regularization on graphs. In *Advances in Neural Information Processing Systems 17*, 2004.
- C. Cortes, P. Haffner, and M. Mohri. Rational kernels. *Neural Information Processing Systems 15*, 2002.
- C. Cortes and V. N. Vapnik. Support–vector networks. *Machine Learning Journal*, 20:273–297, 1995.
- T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.
- T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.
- F. G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, pages 327–331, Pensacola, Florida, 2002.
- F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-supervised learning and model search. In *Proceedings of the ICML-2003 workshop The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 111–112, 2003a.
- F. G. Cozman, I. Cohen, and M. C. Cirelo. Semi-supervised learning of mixture models. In *International Conference on Machine Learning*, pages 99–106, 2003b.
- M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1–2):69–113, 2000.
- P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.
- N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, U.K., 2000.
- N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems 14 (NIPS01)*, 2002a.
- N. Cristianini, J. Shawe-Taylor, and J. Kandola. Spectral kernel methods for clustering. In *Advances in Neural Information Processing Systems 14 (NIPS01)*, pages 649–655, 2002b.
- L. Csató. *Gaussian Processes — Iterative Sparse Approximations*. PhD thesis, Aston University, 2002.
- J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34(4):508–519, March 1999.
- S. Dasgupta. Performance guarantees for hierarchical clustering. In *Proceedings of COLT*, pages 351–363, 2002.
- S. Dasgupta, M. L. Littman, and D. McAllester. PAC generalization bounds for co-training. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2001. MIT Press.
- N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463–474, 1969.
- T. De Bie and N. Cristianini. Convex methods for transduction. In *Neural Information Processing Systems (NIPS)*, 2003.
- T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems 16 (NIPS03)*, pages 73–80, 2004a.
- T. De Bie and N. Cristianini. Kernel methods for exploratory data analysis: a demonstration on text data. In *Proc. of the International Workshop on Statistical Pattern Recognition (SPR04)*, pages 16–29, 2004b.
- T. De Bie, N. Cristianini, and R. Rosipal. Eigenproblems in pattern recognition. In E. Bayro-Corrochano, editor, *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*, pages 129–170. Springer-Verlag, Heidelberg, 2005.
- T. De Bie, M. Momma, and N. Cristianini. Efficiently learning the metric with side-information. In *Proc. of the 14th International Conference on Algorithmic Learning Theory (ALT03)*, pages 175–189, 2003.
- T. De Bie, J. Suykens, and B. De Moor. Learning from general label constraints. In *Proc. of IAPR International Workshop on Statistical Pattern Recognition (SPR04)*, pages 671–679, 2004a.
- T. De Bie, J. A. K. Suykens, and B. De Moor. Learning from general label constraints. In *Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition*, pages 671–679. Lisbon, Portugal, 2004b.

- V. de Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 721–728, Cambridge, MA, 2003. MIT Press.
- S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, 1990.
- M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.
- O. Dekel, C. D. Manning, and Y. Singer. Log-linear models for label-ranking. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.
- O. Delalleau, Y. Bengio, and N. Le Roux. Efficient non-parametric function induction in semi-supervised learning. In *Artificial Intelligence and Statistics*, 2005.
- A. Demiriz and K. P. Bennett. Optimization approaches to semi-supervised learning. In M. C. Ferris, O. L. Mangasarian, and J. S. Pang, editors, *Applications and Algorithms of Complementarity*, pages 121–141. Kluwer, 2000.
- A. Demiriz, K. P. Bennett, and M. J. Embrechts. Semi-supervised clustering using genetic algorithms. In *Proceedings of ANNIE*, pages 809–814, 1999.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
- M. Deng, T. Chen, and F. Sun. An integrated probabilistic model for functional prediction of proteins. In W. Miller, M. Vingron, S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Seventh Annual International Conference on Computational Biology (RECOMB)*, pages 95–103. ACM, 2003.
- P. Derbeko, R. El-Yaniv, and R. Meir. Error bounds for transductive learning via compression and clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1085–1092. MIT Press, 2003.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of mathematics*. Springer, New York, 1996.
- I. S. Dhillon and Y. Guan. Information theoretic clustering of sparse co-occurrence data. In *Proceedings of ICDM*, pages 517–521, 2003.
- I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001.
- T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.
- B. E. Dom. An information-theoretic external cluster-validity measure. Research Report RJ 10219, IBM, 2001.
- P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, 1997.
- D. L. Donoho and C. E. Grimes. When does Isomap recover the natural parameterization of families of articulated images? Technical Report 2002-27, Department of Statistics, Stanford University, August 2002.
- D. L. Donoho and C. E. Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Arts and Sciences*, 100:5591–5596, 2003.
- P. G. Doyle and J. L. Snell. Random walks and electric networks. *Mathematical Association of America*, 1984.
- H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999.
- S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. ACM International Conference on Information and Knowledge Management*, pages 148–155, 1998.
- J. Dunagan and S. Vempala. Optimal outlier removal in high-dimensional spaces. In *Proceedings of the 33rd ACM Symposium on Theory of Computing*, 2001.
- B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, 1975.
- B. Efron. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, 21:460–480, 1979.
- A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Inf. and Comput.*, 82:246–261, 1989.
- B. Fischer, V. Roth, and J. M. Buhmann. Clustering with the connectivity kernel. In *Advances in Neural Information Processing Systems 16*, 2004.
- A. Flaxman. Personal communication, 2003.
- D. Foster and E. George. The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975,

- 1994.
- S. C. Fralick. Learning to recognize patterns without a teacher. *IEEE Transactions on Information Theory*, 13: 57–64, 1967.
- Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.
- J. H. Friedman, J. L. Bentley, and R. A. Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226, 1977.
- J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337–407, 2000.
- N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 129–138, 1998.
- N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
- G. Fung and O. Mangasarian. Semi-supervised support vector machines for unlabeled data classification. *Optimization Methods and Software*, 15:29–44, 2001.
- C. Galarza, E. Rietman, and V. Vapnik. Applications of model selection techniques to polynomial approximation. Preprint, 1996.
- A. Gammerman, V. Vapnik, and V. Vowk. Learning by transduction. In *Conference on Uncertainty in Artificial Intelligence*, pages 148–156, 1998.
- S. Ganesalingam. Classification and mixture approaches to clustering via maximum likelihood. *Applied Statistics*, 38(3):455–466, 1989.
- S. Ganesalingam and G. McLachlan. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 65:658–662, 1978.
- S. Ganesalingam and G. McLachlan. Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation*, 9:151–158, 1979.
- A. Garg and D. Roth. Understanding probabilistic classifiers. In *ECML*, pages 179–191, 2001.
- S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–742, 1984.
- R. Ghani. Combining labeled and unlabeled data for text classification with a large number of categories. In *Proceedings of the IEEE International Conference on Data Mining*, 2001.
- R. Ghani. Combining labeled and unlabeled data for multiclass text categorization. In *ICML*, 2002.
- E. Giné and A. Guillaou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 38(6):907–921, November 2002.
- L. Goldstein and K. Messer. Optimal plug-in estimators for nonparametric functional estimation. *The Annals of statistics*, 20(3):1306–1328, 1992.
- G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins, 3rd edition, 1996.
- C. Goutte, H. Déjean, E. Gaussier, J.-M. Renders, and N. Cancedda. Combining labelled and unlabelled data: a case study on fisher kernels and transductive inference for biological entity recognition. In *Conference on Natural Language Learning (CoNLL)*, 2002.
- T. Graepel, R. Herbrich, and K. Obermayer. Bayesian transduction. In *Advances in Neural Information System Processing (NIPS99)*, volume 12, 2000.
- Y. Grandvalet. Logistic regression for partial labels. In *9th Information Processing and Management of Uncertainty in Knowledge-based Systems – IPMU'02*, pages 1935–1941, 2002.
- Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, volume 17, 2004.
- A. G. Gray and A. W. Moore. N-Body problems in statistical learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 521–527, Cambridge, MA, 2001. MIT Press.
- M. Gribskov and N. L. Robinson. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Computers and Chemistry*, 20(1):25–33, 1996.
- L. Hagen and A. B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE. Trans. on*

- Computed Aided Design*, 11:1074–1085, 1992.
- J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty First International Conference on Machine Learning (ICML-04)*, pages 369–376, Banff, Canada, 2004.
- J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. Unpublished manuscript, 1971.
- J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.
- W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer, <http://www.xplora-stat.de/ebooks/ebooks.html>, 2004.
- R. Hardt and F. H. Lin. Mappings minimizing the L^p norm of the gradient. *Communications on Pure and Applied Mathematics*, 40:556–588, 1987.
- H. O. Hartley and J. N. K. Rao. Classification and estimation in analysis of variance problems. *Review of International Statistical Institute*, 36:141–147, 1968.
- T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, New York, 1990.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.
- D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California, Santa Cruz, Santa Cruz, CA, July 1999.
- D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2001.
- M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in r^d . 2005.
- M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph laplacians. In *Proc. 18th Conf. on Learning Theory*, pages 470–485, 2005.
- J. Heinonen, T. Kilpeläinen, and O. Martio. *Nonlinear Potential Theory of Degenerate Elliptic Equations*. Oxford University Press, Oxford, 1993.
- C. Helmberg. Semidefinite programming for combinatorial optimization. Habilitationsschrift ZIB-Report ZR-00-34, TU Berlin, Konrad-Zuse-Zentrum Berlin, 2000.
- H. Hishigaki, K. Nakai, T. Ono, A. Tanigaki, and T. Takagi. Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, 18:523–531, 2001.
- D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k -center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985.
- T. Hofmann and J. Puzicha. Statistical models for co-occurrence data. Technical Report AI Memo 1625, Artificial Intelligence Laboratory, MIT, February 1998.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.
- D. W. Hosmer Jr. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics*, 29:761–770, December 1973.
- P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, pages 221–233. University of California Press, 1967.
- E. Ie, J. Weston, W. S. Noble, and C. Leslie. Multi-class protein fold recognition using adaptive codes. In *Proceedings of the International Conference on Machine Learning*, 2005.
- J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31:370–377, 2002.
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 2000.
- T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493, Cambridge, MA, 1999. MIT Press.
- T. Jebara, R. Kondor, and A. Howard. Probability product kernels. *Journal of Machine Learning*, 5:819–844, 2004.
- R. Jin and Z. Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2003.
- T. Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 143–151, 1997. URL <ftp://ftp.cs.cmu.edu/afs/cs/user/thorsten/www/icml97.ps.Z>.

- T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, pages 137–142, 1998.
- T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, pages 200–209, Bled, Slovenia, 1999. Morgan Kaufmann. URL http://www-ai.cs.uni-dortmund.de/DOKUMENTE/joachims_99c.ps.gz.
- T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer, 2002.
- T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning (ICML 2003)*, 2003.
- I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.
- M. I. Jordan, editor. *Learning in Graphical Models*. MIT Press, 1999.
- M. Kärstén. Generalization error bounds using unlabeled data. In *Proceedings Annual Conference on Computational Learning Theory (COLT-05)*, 2005.
- M. Kärstén and J. Langford. A comparison of tight generalization bounds. In *Proceedings International Conference on Machine Learning (ICML-05)*, 2005.
- S. D. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In *Proc. of the International Joint Conferences on Artificial Intelligence (IJCAI03)*, pages 561–566, 2003.
- T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.
- M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resources for deciphering genome. *Nucleic Acids Res.*, 32:D277–D280, 2004.
- N. Kasabov and S. Pang. Transductive support vector machines and applications in bioinformatics for promoter recognition. *Neural Information Processing - Letters and Reviews*, 3(2):31–38, 2004.
- M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Journal of the ACM (JACM)*, pages 983–1006, 1998.
- M. Kearns, Y. Mansour, and A. Y. Ng. An information-theoretic analysis of hard and soft assignment methods for clustering. In *Proceedings of UAI*, pages 282–293, 1997.
- M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.
- M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. In S. J. Hanson, G. A. Drastal, and R. L. Rivest, editors, *Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect*, volume 1. MIT Press, Bradford, 1994.
- S. S. Keerthi and D. DeCoste. A modified finite newton method for fast solution of large scale linear svms. *Journal of Machine Learning Research*, 2005.
- B. Kegl and L. Wang. Boosting on manifolds: adaptive regularization of base classifiers. In *NIPS*, volume 17, 2004.
- D. Klein, S. D. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of ICML*, pages 307–314, Sydney, Australia, 2002.
- J. Kleinberg. Detecting a network failure. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, pages 231–239, 2000.
- J. Kleinberg, M. Sandler, and A. Slivkins. Network failure detection and graph connectivity. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, pages 231–239, 2004.
- J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. In *Proceedings of FOCS*, pages 14–23, 1999.
- A. R. Klivans, R. O’Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, pages 177–186, 2002.
- M. Kockelkorn, A. Lüneburg, and T. Scheffer. Using transduction and multi-view learning to answer emails. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 266–277, 2003.
- R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-95)*, 1995.
- V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, pages 1902–1914, 2001.
- P. Komarek and A. Moore. Fast robust logistic regression for large sparse datasets with binary outputs. In

- Artificial Intelligence and Statistics*, 2003.
- R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML-02)*, 2002.
- B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo. On semi-supervised classification. In *NIPS*, volume 17, 2004.
- A. Krogh, M. Brown, I. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.
- A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pages 231–238, 1995.
- R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *Computational Systems Biology Conference*, 2004.
- S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.
- T. N. Lal, M. Schröder, T. Hinterberger, J. Weston, M. Bogdan, N. Birbaumer, and B. Schölkopf. Support vector channel selection in bci. *IEEE Transactions on Biomedical Engineering*, 51(6):1003–1010, 2004.
- L. Lamport. How to write a proof. *American Mathematical Monthly*, 102(7):600–608, 1993.
- G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004a.
- G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004b.
- G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, 2004c.
- T. Lange, M. H. C. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabeled data. In *Computer Vision and Pattern Recognition*, pages 731–738. San Diego, CA, 2005.
- S. Lauritzen. *Graphical Models*. Oxford Statistical Sciences. Clarendon Press, 1996.
- G. Lebanon. Learning Riemannian metrics. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers, 2003.
- I. Lee, S.V. Date, A.T. Adai, and E.M. Marcotte. A probabilistic functional network of yeast genes. *Science*, 306(5701):1555–1558, 2004.
- T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. R. Harbison, C. M. Thompson, I. Simon, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298:799–804, 2002.
- B. Leskes. The value of agreement, a new boosting algorithm. In *COLT*, pages 51 – 56, 2005.
- C. Leslie, E. Eskin, J. Weston, and W. S. Noble. Mismatch string kernels for SVM protein classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *Advances in Neural Information Processing Systems*, pages 1441–1448, Cambridge, MA, 2003. MIT Press.
- C. Leslie and R. Kuang. Fast kernels for inexact string matching. *Proceedings of COLT/KW*, 2003.
- A. Levin, P. Viola, and Y. Freund. Unsupervised improvement of visual detectors using Co-Training. In *Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003)*, pages 626–633, Nice, France, 2003. IEEE.
- D. D. Lewis. The reuters-21578 data set. <http://www.daviddlewis.com/resources/testcollections/reuters21578/>, 1997.
- D. D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine Learning: ECML-98, Tenth European Conference on Machine Learning*, pages 4–15, 1998.
- D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994. URL <http://www.research.att.com/~lewis/papers/lewis94b.ps>.
- C. Liao and W. S. Noble. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. *Proceedings of RECOMB*, 2002.
- R. Liere and P. Tadepalli. Active learning with committees for text categorization. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 591–596, 1997. URL <http://www.cs.orst.edu/~lierer/aaai97.ps>.
- J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

- N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, fourier transform, and learnability. In *Proceedings of the Thirtieth Annual Symposium on Foundations of Computer Science*, pages 574–579, Research Triangle Park, North Carolina, October 1989.
- R. J. A. Little. Discussion on the paper by Professor Dempster, Professor Laird and Dr. Rubin. *Journal of the Royal Statistical Society, Series B*, 39(1):25, 1977.
- H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.
- J. L'öfberg. YALMIP 3, 2004. <http://control.ee.ethz.ch/~joloef/yalmip.msql>.
- D. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.
- J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
- C. Mallows. Some comments on C_p . *Technometrics*, 15:661–676, 1973.
- K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley and Sons, 2nd edition, 2000.
- A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *Learning for Text Categorization: Papers from the AAAI Workshop*, pages 41–48, 1998a. Tech. rep. WS-98-05, AAAI Press.
- A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. In *Proceedings of ICML*, Madison, WI, 1998b.
- A. McCallum, R. Rosenfeld, T. Mitchell, and A. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Machine Learning: Proceedings of the Fifteenth International Conference*, pages 359–367, 1998.
- G. J. McLachlan. Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, 70(350):365–369, 1975.
- G. J. McLachlan. Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *Journal of the American Statistical Association*, 72(358):403–406, 1977.
- G. J. McLachlan. *Discriminant analysis and statistical pattern recognition*. Wiley, 1992.
- G. J. McLachlan and S. Ganesalingam. Updating a discriminant function on the basis of unclassified data. *Communications in Statistics: Simulation and Computation*, 11(6):753–767, 1982.
- G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. John Wiley and Sons, New York, 1997.
- M. Meila. *Learning with mixtures of trees*. PhD thesis, MIT, 1999.
- S. Mendelson and P. Philips. Random subclass bounds. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT)*, 2003.
- C. J. Merz, D. C. St. Clair, and W. E. Bond. Semi-supervised adaptive resonance theory (smart2). In *International Joint Conference on Neural Networks*, volume III, pages 851–856, 1992.
- D. Miller and H. Uyar. A generalized Gaussian mixture classifier with learning based on both labelled and unlabelled data. In *Proceedings of the 1996 Conference on Information Science and Systems*, 1996.
- D. Miller and H. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 571–577, Cambridge, MA, 1997. MIT Press.
- T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.
- T. Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.
- G. D. Murray and D. M. Titterton. Estimation problems with data from a mixture. *Applied Statistics*, 27(3): 325–334, 1978.
- A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.
- E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.
- E. A. Nadaraya. *Nonparametric estimation of probability densities and regression curves*. Kluwer Academic Publishers, 1989.
- R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, 1998.
- S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-20). Technical Report CUCS-005-96, Columbia Univ., USA, February 1996.
- Y. Nesterov and A. Nemirovsky. Interior-point polynomial methods in convex programming: Theory and

- applications. *SIAM*, 13, 1994.
- A. Y. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *NIPS*, volume 14, pages 841–848. MIT Press, 2001.
- A. Y. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS 14)*, Cambridge, MA, 2002. MIT Press.
- K. Nigam. Using unlabeled data to improve text classification. Technical Report Doctoral Dissertation, CMU-CS-01-126, Carnegie Mellon University, 2001.
- K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of Co-training. In *Proc. ACM CIKM Int. Conf. on Information and Knowledge Management*, pages 86–93, 2000.
- K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 792–799, 1998.
- K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2/3):103–134, 2000.
- A. O’Hagan. Some Bayesian numerical analysis. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 345–363, Valencia, 1992. Oxford University Press.
- T. O’Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364):821–826, 1978.
- D. Opitz and J. Shavlik. Generating accurate and diverse members of a neural-network ensemble. In *Advances in Neural Information Processing Systems 8*, 1996.
- M. Ouimet and Y. Bengio. Greedy spectral embedding. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, 2005.
- A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Inc., New York, 4th edition, 2001.
- J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect twice as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–1210, 1998.
- S. Park and B. Zhang. Large scale unstructured document classification using unlabeled data and syntactic information. In *PAKDD 2003*, LNCS vol. 2637, pages 88–99. Springer, 2003.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
- F. C. N. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *ACL*, pages 183–190, Columbus, Ohio, 1993.
- D. Pierce and C. Cardie. Limitations of Co-Training for natural language learning from large datasets. In *Proc. Conference on Empirical Methods in NLP*, pages 1–9, 2001.
- J. C. Platt. Fast embedding of sparse similarity graphs. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2004. MIT Press.
- T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.
- T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri. b. In *Proc. of the Conference on Uncertainty in Geometric Computations*, pages 22–28, 2001.
- M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- J. Ratsaby and S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 412–417, 1995.
- R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.
- B. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
- J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.
- J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, chapter 14, pages 313–323. Prentice Hall, 1971.
- K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*,

- 11(9):589–594, 1990.
- K. Rose, E. Gurewitz, and G. Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38(4):1249–1257, 1992.
- S. Rosenberg. *The Laplacian on a Riemannian Manifold*. Cambridge University Press, Cambridge, UK, 1997.
- B. Rost and C. Sander. Prediction of protein secondary structure at better than 70 accuracy. *J Mol Biol*, 232(2): 584–599, July 1993.
- D. Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82(1-2):273–302, 1996.
- S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290: 2323–2326, 2000.
- Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, MA, 1996.
- H. Saigo, J. Vert, N. Ueda, and T. Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 2004.
- Sajama and A. Orlitsky. Estimating and computing density based distance metrics. In *Proc. 22th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2005.
- L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.
- C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, pages 722–726, San Francisco, 1999. Morgan Kaufman.
- C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10(2):153–178, 1993.
- C. Schaffer. A conservation law for generalization performance. In *Proceedings of International Conference on Machine Learning (ICML-94)*, pages 683–690, 1994.
- B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
- B. Schölkopf, A. J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.
- D. Schuurmans and F. Southey. Metric-based methods for adaptive model selection and regularization. *Machine Learning*, 48(1-3):51–84, 2002. Special Issue on New Methods for Model Selection and Model Combination.
- D. Schuurmans, L. Ungar, and D. Foster. Characterizing the generalization performance of model selection strategies. In *Proceedings of International Conference on Machine Learning (ICML-97)*, pages 340–348, 1997.
- G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 18:1257–1261, 2000.
- H. J. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11:363–371, 1965.
- M. Seeger. Input-dependent regularization of conditional density models. Technical report, Institute for ANC, Edinburgh, UK, 2000a. See www.kyb.tuebingen.mpg.de/bs/people/seeger.
- M. Seeger. Learning with labeled and unlabeled data. Technical report, Institute for ANC, Edinburgh, UK, 2000b. See www.kyb.tuebingen.mpg.de/bs/people/seeger.
- M. Seeger. Covariance kernels from Bayesian generative models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 905–912, Cambridge, MA, 2002. MIT Press.
- E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition specific regulators from gene expression data. *Nature Biotechnology*, 34(2):166–176, 2003a.
- E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19:i264–i272, July 2003b.
- J. A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 1999.
- F. Sha and L. K. Saul. Analysis and extension of spectral methods for nonlinear dimensionality reduction. In *Proceedings of the Twenty Second International Conference on Machine Learning (ICML-05)*, Bonn, Germany, 2005.
- B. Shahshahani and D. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, Sept 1994. URL <http://dynamo.ecn.purdue.edu/~landgreb/GRS94.pdf>.

- J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
- J. Shawe-Taylor and N. Cristianini. *Kernel methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.
- N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall. Computing gaussian mixture models with EM using equivalence constraints. In *Advances in Neural Information Processing Systems 16 (NIPS03)*, pages 465–472, 2004.
- J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
- R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.
- V. Sindhwani. Kernel machines for semi-supervised learning. Technical report, Masters Thesis, University of Chicago, 2004.
- V. Sindhwani, W. Chu, and S. S. Keerthi. Semi-supervised gaussian processes. Technical report, Yahoo! Research (In preparation), 2005a.
- V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, 2005b.
- T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
- A. Smola and R. Kondor. Kernels and regularization on graphs. In *Conference on Learning Theory, COLT/KW*, 2003.
- P. Sollich. Probabilistic interpretation and Bayesian methods for support vector machines. In *Proceedings 1999 International Conference on Artificial Neural Networks, ICANN'99*, pages 91–96, London, U.K., 1999. The Institution of Electrical Engineers.
- P. Sollich. Probabilistic methods for support vector machines. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 349–355, Cambridge, MA, 2000. MIT Press.
- K. Sparck-Jones. A statistical interpretation of term specificity and its application in retrieval. *J. of Documentation*, 28(1):11–21, 1972.
- P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.
- D. A. Spielman and S. H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 26th annual ACM symposium on Theory of computing*, pages 81–90. ACM Press, 2004.
- A. Stolcke and S. M. Omohundro. Best-first model merging for hidden Markov model induction. Technical Report TR-94-003, ICSI, University of California, Berkeley, 1994. URL <http://www.icsi.berkeley.edu/techreports/1994.html>.
- C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of statistics*, 8(6):1348–1360, 1980.
- A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64, 2000.
- R. Strichartz. *The Way of Analysis*. Jones and Bartlett, 1995.
- J. F. Sturm. Using SeDuMi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12(8):625–653, 1999. Special issue on Interior Point Methods (CD supplement with software).
- J. Sun, S. Boyd, L. Xiao, and P. Diaconis. The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. *SIAM Review*, 2005. Submitted.
- M. Szummer and T. Jaakkola. Clustering and efficient use of unlabeled examples. In *Advances in Neural Information processing systems 14*, 2001.
- M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002a.
- M. Szummer and T. Jaakkola. Partially labeled classification with markov random walks. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002b. MIT Press.
- B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the*

- Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2002.
- J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
- A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems*. W. H. Winston, Washington, DC, 1977.
- D. M. Titterton. Updating a diagnostic system using unconfirmed cases. *Applied Statistics*, 25(3):238–247, 1976.
- D. M. Titterton, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1st edition, 1985.
- S. Tong and D. Koller. Restricted bayes optimal classifiers. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 658–664, 2000.
- S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2:45–66, November 2001. ISSN 1533-7928 (electronic); 1532-4435 (paper).
- V. Tresp. A Bayesian committee machine. *Neural Computation*, 12(11):2719–2741, 2000.
- K. Tsuda and W. S. Noble. Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, 20(Suppl. 1):i326–i333, 2004.
- N. Ueda and R. Nakano. Deterministic annealing variant of the EM algorithm. In *Advances in Neural Information Processing Systems 7*, pages 545–552, 1995.
- P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.
- L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.
- T. van Allen and R. Greiner. A model selection criteria for learning belief nets: An empirical comparison. In *International Conference on Machine Learning*, pages 1047–1054, 2000.
- C. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, June 1977.
- L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.
- V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer Verlag, New York, 1982.
- V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.
- V. Vapnik and A. Chervonenkis. Uniform convergence of frequencies of occurrence of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:915–918, 1968.
- V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.
- V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, 1974.
- V. Vapnik and A. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):283–305, 1991.
- V. Vapnik and A. Sterin. On structural risk minimization or overall risk in a problem of pattern recognition. *Automation and Remote Control*, 10(3):1495–1503, 1977.
- A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, 21(6):697–700, 2003.
- S. Vempala. A random sampling based algorithm for learning the intersection of half-spaces. In *Proceedings of the 38th Symposium on Foundations of Computer Science*, pages 508–513, 1997.
- K. A. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *COLT*, pages 314–326, 1990.
- J.-P. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 1425–1432. MIT Press, 2003.
- J.-P. Vert and Y. Yamanishi. Supervised graph inference. In *NIPS*, volume 17, 2004.
- P. Vincent and Y. Bengio. Density-sensitive metrics and kernels. In *Workshop on Advances in Machine Learning*, Montréal, Québec, Canada, 2003.
- S. V. N. Vishwanathan and A. Smola. Fast kernels for string and tree matching. *Neural Information Processing Systems 15*, 2002.

- U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.
- U. von Luxburg, O. Bousquet, and M. Belkin. Limits of spectral clustering. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA, 2005.
- C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Olivier, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.
- V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. In *Proc. 16th International Conf. on Machine Learning*, pages 444–453. Morgan Kaufmann, San Francisco, CA, 1999.
- K. Wagstaff. *Intelligent Clustering with Instance-Level Constraints*. PhD thesis, Cornell University, 2002.
- K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained K-Means clustering with background knowledge. In *Proceedings of ICML*, pages 577–584, 2001.
- G. Wahba. *Spline Models for Observational Data*. Number 59 in CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1990.
- M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.
- L. Wang, K. L. Chan, and Z. Zhang. Bootstrapping SVM active learning by incorporating unlabelled images for image retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 629–634, 2003.
- W. Wapnik and A. Tscherwonienkis. *Theorie der Zeichenerkennung*. Akademie Verlag, Berlin, 1979.
- C. Watkins. Dynamic alignment kernels. In A. J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.
- G. S. Watson. Smooth regression analysis. *Sankhya - The Indian Journal of Statistics*, 26:359–372, 1964.
- K. Q. Weinberger, B. D. Packer, and L. K. Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *Proceedings of the Tenth International Workshop on AI and Statistics (AISTATS-05)*, Barbados, WI, 2005.
- K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal on Computer Vision*, 2005. Submitted.
- Y. Weiss. Segmentation using eigenvectors: a unifying view. In *ICCV*, pages 975–982. Kerkyra, Greece, 1999.
- J. Weston, C. Leslie, D. Zhou, A. Elisseeff, and W. S. Noble. Cluster kernels for semi-supervised protein classification. *Advances in Neural Information Processing Systems 17*, 2003a.
- J. Weston, F. Pérez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schölkopf. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, 19(6):764–771, 2003b.
- H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, January 1982.
- C. K. I. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In Michael I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *Series D: Behavioural and Social Sciences*, Dordrecht, The Netherlands, 1998. Kluwer.
- C. K. I. Williams. On a connection between kernel PCA and metric multidimensional scaling. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 675–681, Cambridge, MA, 2001. MIT Press.
- C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA, 2001. MIT Press.
- E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS 15*, pages 505–512. MIT Press, 2003.
- M. Yamasaki. Ideal boundary limit of discrete Dirichlet functions. *Hiroshima Math. J.*, 16(2):353–360, 1986.
- Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proc. 21st International ACM SIGIR Conf.*, pages 42–49, 1999.
- Y. Yang and J. O. Pedersen. Feature selection in statistical learning of text categorization. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 412–420, 1997.
- D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Meeting of the Association for Computational Linguistics*, pages 189–196, 1995.
- G. Yona, N. Linial, and M. Linial. Protomap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Structure, Function, and Genetics*, 37:360–678, 1999.
- K. Yu, V. Tresp, and D. Zhou. Semi-supervised induction with basis function. Technical report, Max-Planck Institute, 2004.

- A. L. Yuille, P. Stolorz, and J. Utans. Statistical physics, mixtures of distributions, and the EM algorithm. *Neural Computation*, 6(2):334–340, 1994.
- S. Zelikovitz and H. Hirsh. Improving short-text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.
- T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In *International Joint Conference on Machine Learning*, pages 1191–1198, 2000.
- Y. Zhang, M. Brady, and S. Smith. Hidden Markov random field model and segmentation of brain MR images. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.
- Z. Zhang and H. Zha. Principal manifolds and nonlinear dimensionality reduction by local tangent space alignment. *SIAM Journal of Scientific Computing*, 26(1):313–338, 2004.
- D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 321–328. MIT Press, Cambridge, MA, 2004.
- D. Zhou, J. Huang, and B. Schölkopf. Learning from Labeled and Unlabeled Data on a Directed Graph. In Luc De Raedt and Stefan Wrobel, editors, *Proceedings of the 22nd International Conference on Machine Learning*, 2005a.
- D. Zhou, B. Schölkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 18*, pages 1633–1640, Cambridge, MA, 2005b. MIT Press.
- X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.
- X. Zhu, Z. Ghahramani, and J. Lafferty. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning*, pages 912–912, Washington, DC, USA, 2003a.
- X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *The Twentieth International Conference on Machine Learning*, pages 912–912, Washington, DC, USA, 2003b. AAAI Press.
- X. Zhu, J. Lafferty, and Z. Ghahramani. Semi-supervised learning: From Gaussian fields to Gaussian processes. Technical Report CMU-CS-03-175, Carnegie Mellon University, 2003c.

Notation and Symbols

Sets of Numbers

\mathbb{N}	the set of natural numbers, $\mathbb{N} = \{1, 2, \dots\}$
\mathbb{R}	the set of reals
$[n]$	compact notation for $\{1, \dots, n\}$
$x \in [a, b]$	interval $a \leq x \leq b$
$x \in (a, b]$	interval $a < x \leq b$
$x \in (a, b)$	interval $a < x < b$
$ C $	cardinality of a set C (for finite sets, the number of elements)

Data

\mathcal{X}	the input domain
d	(used if \mathcal{X} is a vector space) dimension of \mathcal{X}
M	number of classes (for classification)
l, u	number of labeled, unlabeled training examples
n	total number of examples, $n = l + u$.
i, j	indices, often running over $[l]$ or $[n]$
x_i	input patterns $x_i \in \mathcal{X}$
y_i	classes $y_i \in [M]$ (for regression: target values $y_i \in \mathbb{R}$)
X	a sample of input patterns, $X = (x_1, \dots, x_n)$
Y	a sample of output targets, $Y = (y_1, \dots, y_n)$
X_l	labeled part of X , $X_l = (x_1, \dots, x_l)$
Y_l	labeled part of Y , $Y_l = (y_1, \dots, y_l)$
X_u	unlabeled part of X , $X_u = (x_{l+1}, \dots, x_{l+u})$
Y_u	unlabeled part of Y , $Y_u = (y_{l+1}, \dots, y_{l+u})$

Kernels

\mathcal{H}	feature space induced by a kernel
Φ	feature map, $\Phi : \mathcal{X} \rightarrow \mathcal{H}$
k	(positive definite) kernel
K	kernel matrix or Gram matrix, $K_{ij} = k(x_i, x_j)$

Vectors, Matrices and Norms

$\mathbf{1}$	vector with all entries equal to one
\mathbf{I}	identity matrix
A^\top	transposed matrix (or vector)
A^{-1}	inverse matrix (in some cases, pseudo-inverse)
$\text{tr}(A)$	trace of a matrix
$\det(A)$	determinant of a matrix
$\langle \mathbf{x}, \mathbf{x}' \rangle$	dot product between \mathbf{x} and \mathbf{x}'
$\ \cdot\ $	2-norm, $\ \mathbf{x}\ := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$
$\ \cdot\ _p$	p -norm, $\ \mathbf{x}\ _p := \left(\sum_{i=1}^N x_i ^p \right)^{1/p}$, $N \in \mathbb{N} \cup \{\infty\}$
$\ \cdot\ _\infty$	∞ -norm, $\ \mathbf{x}\ _\infty := \sup_{i=1}^N x_i $, $N \in \mathbb{N} \cup \{\infty\}$

Functions

\ln	logarithm to base e
\log_2	logarithm to base 2
f	a function, often from \mathcal{X} or $[n]$ to \mathbb{R} , \mathbb{R}^M or $[M]$
\mathcal{F}	a family of functions
$L_p(\mathcal{X})$	function spaces, $1 \leq p \leq \infty$

Probability

$P\{\cdot\}$	probability of a logical formula
$P(C)$	probability of a set (event) C
$p(x)$	density evaluated at $x \in \mathcal{X}$
$\mathbf{E}[\cdot]$	expectation of a random variable
$\mathbf{Var}[\cdot]$	variance of a random variable
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2

Graphs

g	graph $g = (V, E)$ with nodes V and edges E
\mathcal{G}	set of graphs
\mathbf{W}	weighted adjacency matrix of a graph ($\mathbf{W}_{ij} \neq 0 \Leftrightarrow (i, j) \in E$)
\mathbf{D}	(diagonal) degree matrix of a graph, $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$
\mathcal{L}	normalized graph Laplacian, $\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$
L	un-normalized graph Laplacian, $L = \mathbf{D} - \mathbf{W}$

SVM-related

$\rho_f(x, y)$	margin of function f on the example (x, y) , i.e., $y \cdot f(x)$
ρ_f	margin of f on the training set, i.e., $\min_{i=1}^m \rho_f(x_i, y_i)$
h	VC dimension
C	regularization parameter in front of the empirical risk term
λ	regularization parameter in front of the regularizer
\mathbf{w}	weight vector
b	constant offset (or threshold)
α_i	Lagrange multiplier or expansion coefficient
β_i	Lagrange multiplier
$\boldsymbol{\alpha}, \boldsymbol{\beta}$	vectors of Lagrange multipliers
ξ_i	slack variables
$\boldsymbol{\xi}$	vector of all slack variables
Q	Hessian of a quadratic program

Miscellaneous

I_A	characteristic (or indicator) function on a set A i.e., $I_A(x) = 1$ if $x \in A$ and 0 otherwise
δ_{ij}	Kronecker δ ($\delta_{ij} = 1$ if $i = j$, 0 otherwise)
δ_x	Dirac δ , satisfying $\int \delta_x(y) f(y) dy = f(x)$
$O(g(n))$	a function $f(n)$ is said to be $O(g(n))$ if there exist constants $C > 0$ and $n_0 \in \mathbb{N}$ such that $ f(n) \leq Cg(n)$ for all $n \geq n_0$
$o(g(n))$	a function $f(n)$ is said to be $o(g(n))$ if there exist constants $c > 0$ and $n_0 \in \mathbb{N}$ such that $ f(n) \geq cg(n)$ for all $n \geq n_0$
rhs/lhs	shorthand for “right/left hand side”
■	the end of a proof