Graphical Models: Modeling, Optimization, and Hilbert Space Embedding

Xinhua Zhang

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

March 2010

Declaration

Except where otherwise indicated, this thesis is my own original work.

Xinhua Zhang
March 2010

The following table gives the collaborators of each chapter, and the publications.

Chapter	Collaborators	Publication
1	Douglas Aberdeen and SVN Vishwanthan	ICML 2007, (Zhang et al., 2007)
2	Thore Graepel and Ralf Herbrich	AISTATS 2010, (Zhang et al., 2010)
3	Alex Smola, Le Song, and Arthur Gretton	NIPS 2008, (Zhang et al., 2009)
4	SVN Vishwanthan and Ankan Saha	Submitted to COLT 2010

Two other papers published but not included in this thesis are (Song, Zhang, Smola, Gretton, & Schölkopf, 2008b) and (Cheng, Vishwanathan, & Zhang, 2008).

To my parents for their love and support.

Acknowledgements

It is finally the time to summarize my four years' PhD research in a single coherent document, and I would like to take this opportunity to thank all the people who helped me along the way.

First of all, I wish to express my utmost gratitude to S. V. N. Vishwanthan, known to all as Vishy, for his long time guidance in my research and life. Being energetic and enthusiastic with research, Vishy has given me most valuable hand-on guidance in both theoretical and practical problems. I still remember those long discussions into late night, both in Vishy's office and at his home. After becoming a faculty at Purdue University, Vishy kindly offered me a visiting scholar position and I considerably benefited from this visit at Purdue.

My gratitude goes to Alex Smola, who showed me how to perform creative research and impressed me with his novel ideas. Despite the pain of following and digesting his thoughts, it is almost always fruitful to explore in the directions Alex suggested and this high level supervision greatly helped improve my ability to conduct independent research.

Thank you to Douglas Aberdeen for introducing reinforcement learning to me, and guiding me through the project of conditional random fields for reinforcement learning. It was really a pity that Doug left NICTA after the project grew into a good shape.

Many thanks to Microsoft Research Cambridge (MSRC), who offered me a 12-week internship from September 2008. My mentors Thore Graepel and Ralf Herbrich are especially creative, knowledgeable and friendly to help me in both theory and practice.

Thank you to Wray Buntine for teaching me nonparametric Bayesian, and chairing my supervisory panel after Alex and Vishy left NICTA.

My gratitude goes to the CSL/RSISE/CECS administration staff, especially Michelle Moravec, Di Kossatz, and Deb Pioch for creating a friendly and efficient environment for learning and research, keeping me on track, and helping me focus on research. Also thank you to Norma Lucas at Purdue University for all the administrative help.

I have also greatly benefited from discussions with other researchers, students and visitors of NICTA, of the Computer Sciences Lab of ANU, and of the Department of Statistics at Purdue University. The following list is absolutely not complete: Marconi Barbosa, Justin Bedo, Olivier Buffet, Tiberio Caetano, Dmitry Kamenetsky, Li Cheng, Nan Ding, Markus Hegland, Jiayuan Huang, Knut Hüper, Marcus Hutter, Quoc Viet

Le, Novi Quadrianto, Mark Reed, Scott Sanner, Nic Schraudolph, Hao Shen, Javen (Qinfeng) Shi, Le Song, Owen Thomas, Choon Hui Teo, Tao Wang, Chris Webers, and Jin Yu.

I gratefully acknowledge the scholarship from NICTA, fellowship from Microsoft Research Asia, Vice-Chancellor's grant from ANU for attending conferences and visiting Purdue University, and travel grants from Pascal, Pascal 2, ICML, and NIPS.

Xinhua Zhang
March 2010

Abstract

Over the past two decades graphical models have been widely used as powerful tools for compactly representing distributions. On the other hand, kernel methods have been used extensively to come up with rich representations. This thesis aims to combine graphical models with kernels to produce compact models with rich representational abilities.

Graphical models are a powerful underlying formalism in machine learning. Their graph theoretic properties provide both an intuitive modular interface to model the interacting factors, and a data structure facilitating efficient learning and inference. The probabilistic nature ensures the global consistency of the whole framework, and allows convenient interface of models to data.

Kernel methods, on the other hand, provide an effective means of representing rich classes of features for general objects, and at the same time allow efficient search for the optimal model. Recently, kernels have been used to characterize distributions by embedding them into high dimensional feature space. Interestingly, graphical models again decompose this characterization and lead to novel and direct ways of comparing distributions based on samples.

Among the many uses of graphical models and kernels, this thesis is devoted to the following four areas:

Conditional random fields for multi-agent reinforcement learning Conditional random fields (CRFs) are graphical models for modeling the probability of labels given the observations. They have traditionally been trained with using a set of observation and label pairs. Underlying all CRFs is the assumption that, conditioned on the training data, the label sequences of different training examples are independent and identically distributed (*iid*). We extended the use of CRFs to a class of temporal learning algorithms, namely policy gradient reinforcement learning (RL). Now the labels are no longer *iid*. They are actions that update the environment and affect the next observation. From an RL point of view, CRFs provide a natural way to model joint actions in a decentralized Markov decision process. They define how agents can communicate with each other to choose the optimal joint action. We tested our framework on a synthetic network alignment problem, a distributed sensor network, and a road traffic control system. Using tree sampling by Hamze & de Freitas (2004) for inference, the RL methods employing CRFs clearly outperform those which do not

model the proper joint policy.

Bayesian online multi-label classification Gaussian density filtering (GDF) provides fast and effective inference for graphical models (Maybeck, 1982). Based on this natural online learner, we propose a Bayesian online multi-label classification (BOMC) framework which learns a probabilistic model of the linear classifier. The training labels are incorporated to update the posterior of the classifiers via a graphical model similar to TrueSkill (Herbrich et al., 2007), and inference is based on GDF with expectation propagation. Using samples from the posterior, we label the test data by maximizing the expected F-score. Our experiments on Reuters1-v2 dataset show that BOMC delivers significantly higher macro-averaged F-score than the state-of-the-art online maximum margin learners such as LaSVM (Bordes et al., 2005) and passive-aggressive online learning (Crammer et al., 2006). The online nature of BOMC also allows us to efficiently use a large amount of training data.

Hilbert space embedment of distributions Graphical models are also an essential tool in kernel measures of independence for non-*iid* data. Traditional information theory often requires density estimation, which makes it unideal for statistical estimation. Motivated by the fact that distributions often appear in machine learning via expectations, we can characterize the distance between distributions in terms of distances between means, especially means in reproducing kernel Hilbert spaces which are called kernel *embedment*. Under this framework, the undirected graphical models further allow us to factorize the kernel embedment onto cliques, which yields efficient measures of independence for non-*iid* data (Zhang et al., 2009). We show the effectiveness of this framework for ICA and sequence segmentation, and a number of further applications and research questions are identified.

Optimization in maximum margin models for structured data Maximum margin estimation for structured data, *e.g.* (Taskar et al., 2004), is an important task in machine learning where graphical models also play a key role. They are special cases of regularized risk minimization, for which bundle methods (BMRM, Teo et al., 2007) and the closely related $\texttt{SVM}^{\texttt{Struct}}$ (Tsochantaridis et al., 2005) are state-of-the-art general purpose solvers. Smola et al. (2007b) proved that BMRM requires $O(1/\epsilon)$ iterations to converge to an ϵ accurate solution, and we further show that this rate hits the lower bound. By utilizing the structure of the objective function, we devised an algorithm for the structured loss which converges to an ϵ accurate solution in $O(1/\sqrt{\epsilon})$ iterations. This algorithm originates from Nesterov's optimal first order methods (Nesterov, 2003, 2005b).

Contents

List of Figures

List of Tables

List of Symbols

$[n]$	Index set $\{1, 2, \ldots, n\}$	3
$\mathbb{R}_+$	Set of positive real numbers $(0, \infty)$	3
$\overline{\mathbb{R}}_+$	Set of non-negative real numbers $[0, \infty)$	3
$\boldsymbol{\phi}$	Vector of sufficient statistics	3
$\top$	Transpose of matrix or vector	3
$\mathcal{X}$	Sample space, or space of feature vectors	3
$\boldsymbol{\theta}$	Natural parameter	3
d	Number of features	3
ν	Base measure of exponential family of distributions	3
Θ	The set of $\boldsymbol{\theta}$ for which $g(\boldsymbol{\theta}) < +\infty$	4
$\mathcal{P}_{\boldsymbol{\phi}}$	Exponential family generated by $\boldsymbol{\phi}$	4
$g(\boldsymbol{\theta})$	Log partition function	3
C^∞	Infinitely differentiable	4
MRF_G	Markov random field corresponding to graph G	6
$X \perp\!\!\!\perp Y \mid Z$	Random variable X is independent of Y given Z	6
$\mathcal{C}$	Set of maximal cliques	7
$k(x, y)$	Kernel function	11
$\mathbb{N}$	Set of natural numbers: $\{1, 2, \ldots\}$	12
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$	12
$\delta(\cdot)$	$\delta(\cdot) = 1$ if $\cdot$ is true, and 0 otherwise	12
$\mathcal{H}(p)$	Reproducing kernel Hilbert space (RKHS)	14
$\xrightarrow{\mathcal{H}}$	Convergence in RKHS norm	14
$\rightrightarrows$	Uniform convergence of functions	14
$C(\mathcal{X})$	Space of continuous bounded functions $\mathcal{X}$	14
$\mathcal{P}_k$	Exponential family generated by kernel k	16
n	Number of data points	17
$\mathcal{Y}^m$	Cartesian product of $\mathcal{Y}$ for m times	18
$\mathrm{KL}(p, q)$	KL divergence from density p to density q: $\int p(x) \frac{p(x)}{q(x)} \mathrm{d}x$	21
$f^\star$	Fenchel dual of function f	23
$\mathcal{M}_{\boldsymbol{\phi}}$	Marginal polytope for sufficient statistics $\boldsymbol{\phi}$	24
$\Lambda_{\boldsymbol{\phi}}$	Mapping from $\boldsymbol{\theta} \in \Theta$ to the mean $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}(x)]$	24
ri	Relative interior	24