Semi-supervised Gaussian Process Classifiers

Vikas Sindhwani
Department of Computer Science
University of Chicago
Chicago, IL 60615, USA
vikass@cs.uchicago.edu

Wei Chu
CCLS
Columbia University
New York, NY 10115, USA
chuwei@gatsby.ucl.ac.uk

S. Sathiya Keerthi
Yahoo! Research
3333 Media Studios North
Burbank, CA 91504, USA
selvarak@yahoo-inc.com

Abstract

In this paper, we propose a graph-based construction of semi-supervised Gaussian process classifiers. Our method is based on recently proposed techniques for incorporating the geometric properties of unlabeled data within globally defined kernel functions. The full machinery for standard supervised Gaussian process inference is brought to bear on the problem of learning from labeled and unlabeled data. This approach provides a natural probabilistic extension to unseen test examples. We employ Expectation Propagation procedures for evidence-based model selection. In the presence of few labeled examples, this approach is found to significantly outperform cross-validation techniques. We present empirical results demonstrating the strengths of our approach.

1 Introduction

Practitioners of machine learning methods frequently encounter the following situation: large amounts of data can be cheaply and automatically collected, but subsequent labeling requires expensive and fallible human participation. This has recently motivated a number of efforts to design semi-supervised inference algorithms that aim to match the performance of well-trained supervised methods, using a small pool of labeled examples and a large collection of unlabeled data. The scarcity of labeled examples greatly exaggerates the need to incorporate additional sources of prior knowledge, to perform careful model selection, and to deal with noise in labels. A Bayesian framework is ideally suited to handle such issues. In this paper, we construct semi-supervised Gaussian processes, and demonstrate and discuss its practical utility on a number of classification problems.

Our methods are based on the geometric intuition that for many real world problems, unlabeled examples often identify structures, such as data clusters or low dimensional manifolds, whose knowledge may potentially improve inference. For example, one might expect high correlation between class labels of data points within the same cluster or nearby on a manifold. These, respectively, are the *cluster* and *manifold* assumptions for semi-supervised learning.

We utilize the graph-based construction of semi-supervised kernels in [Sindhwani *et al.*, 2005]. The data geometry is modeled as a graph whose vertices are the labeled and unlabeled examples, and whose edges encode appropriate neighborhood relationships. In intuitive terms, in the limit of infinite unlabeled data we imagine a convergence of this graph to the underlying geometric structure of the probability distribution generating the data. The smoothness enforced by regularization operators (such as the graph Laplacian) over functions on the vertices of this graph is transferred onto a Reproducing Kernel Hilbert space (RKHS) of functions defined over the entire data space. This gives rise to a new RKHS in which standard supervised kernel methods perform semi-supervised inference. In this paper, we apply similar ideas for Gaussian processes, with an aim to draw the benefits of a Bayesian approach when only a small labeled set is available. The semi-supervised kernel of [Sindhwani *et al.*, 2005] is motivated through Bayesian considerations, and used for performing Gaussian process inference. We apply expectation propagation (EP) (see [Rasmussen and Williams, 2006] and references therein) for approximating posterior processes and calculating evidence for model selection.

We point out some aspects of the proposed method:

a) Our method, hereafter abbreviated as SSGP, can be seen as providing a Bayesian analogue of Laplacian Support Vector Machines (LapSVM) and Laplacian Regularized Least Squares (LapRLS) proposed in [Sindhwani *et al.*, 2005; Belkin *et al.*, 2006] and semi-supervised logistic regression proposed in [Krishnapuram *et al.*, 2004]. In this paper, we empirically demonstrate that when labeled examples are scarce, model selection by evidence in SSGP is a significant improvement over cross-validation in LapSVM and LapRLS.

b) Several graph-based Bayesian approaches incorporating unlabeled data have recently been proposed. Methods have been designed for transductive Bayesian learning using a graph-based prior in [Zhu *et al.*, 2003; Kapoor *et al.*, 2005]. Since the core probabilistic model in these methods is defined only over the finite collection of labeled and unlabeled inputs, out-of-sample extension to unseen test data requires additional follow-up procedures. By contrast, our approach possesses a Gaussian process model over the entire input space that provides natural out-of-sample prediction.

c) While the focus of this paper is to derive SSGP and verify its usefulness on binary classification tasks, we wish to comment on its potential in the following respects: (1) SSGP also provides a general framework for other learning tasks, such as regression, multiple-class classification and ranking. (2) SSGP also produces class probabilities and confidence estimates. These can be used to design active learning schemes. (3) Efficient gradient-based strategies can be designed for choosing multiple model parameters such as regularization and kernel parameters.

We begin by briefly reviewing Gaussian process classification in Section 2 and then propose and evaluate SSGP in the following sections.

2 Background and Notation

In the standard formulation for learning from examples, patterns $\mathbf{x}$ are drawn from some space $\mathcal{X}$, typically a subset of $\mathcal{R}^d$ and associated labels y come from some space $\mathcal{Y}$. For ease of presentation, in this paper we will focus on binary classification, so that $\mathcal{Y}$ can be taken as $\{+1, -1\}$. We assume there is an unknown joint probability distribution $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ over the space of patterns and labels. A set of patterns $\{\mathbf{x}_i\}_{i=1}^{l+u}$ is drawn i.i.d from the marginal $\mathcal{P}_{\mathcal{X}}(\mathbf{x}) = \sum_{\mathcal{Y}} \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}(\mathbf{x}, y)$; the label y_i associated with $\mathbf{x}_i$ is drawn from the conditional distribution $\mathcal{P}_{\mathcal{Y}}(y|\mathbf{x}_i)$ for the first l patterns. We are interested in leveraging the unlabeled patterns $\{\mathbf{x}_j\}_{j=l+1}^{l+u}$ for improved inference.

To set our notation for the discussion ahead, we will denote $X_L = \{\mathbf{x}_i\}_{i=1}^{l}$ as the set of labeled examples with associated labels $Y_L = \{y_i\}_{i=1}^{l}$, $X_U = \{\mathbf{x}_i\}_{i=l+1}^{l+u}$ as the set of unlabeled patterns. We further denote all the given patterns as X_D, i.e., $X_D = X_L \cup X_U$. Other patterns that are held out for test purposes (or have not been collected yet) are denoted as X_T. The set of all patterns (labeled, unlabeled and test) is denoted as X, i.e., $X = X_D \cup X_T$, but we will often also use X to denote a generic dataset.

In the standard Gaussian process setting for supervised learning, one proceeds by choosing a covariance function $K : \mathcal{X} \times \mathcal{X} \mapsto \Re$. With any point $\mathbf{x} \in \mathcal{X}$, there is an associated latent variable $f_{\mathbf{x}}$. Given any dataset X, the latent variables $\mathbf{f}_X = \{f_{\mathbf{x}}\}_{\mathbf{x}\in X}$ are treated as random variables in a zero-mean Gaussian process indexed by the data points. The covariance between $f_{\mathbf{x}}$ and $f_{\mathbf{z}}$ is fully determined by the coordinates of the data points $\mathbf{x}$ and $\mathbf{z}$, and is given by $K(\mathbf{x}, \mathbf{z})$. Thus, the prior distribution over these variables is a multivariate Gaussian, written as: $\mathcal{P}(\mathbf{f}_X) = \mathcal{N}(0, \Sigma_{XX})$ where Σ_{XX} is the covariance matrix with elements $K(\mathbf{x}, \mathbf{z})$ with $\mathbf{x}, \mathbf{z} \in X$.

The Gaussian process classification model relates the variable $f_{\mathbf{x}}$ at $\mathbf{x}$ to the label y through a *probit* noise model such as $y = \text{sign}(f_{\mathbf{x}} + \xi)$ where $\xi \sim \mathcal{N}(0, \sigma_n^2)$. Given the latent variables $\mathbf{f}_L$ associated with labeled data points, the class labels Y_L are independent Bernoulli variables and their joint likelihood is given by: $\mathcal{P}(Y_L|\mathbf{f}_L) = \prod_{i=1}^{l} \mathcal{P}(y_i|f_{\mathbf{x}_i}) = \prod_{i=1}^{l} \Phi\left(\frac{y_i f_{\mathbf{x}_i}}{\sigma_n}\right)$ where Φ is the cumulative density function of the normal distribution.

Combining this likelihood term with the Gaussian prior for the latent variables associated with the labeled dataset X_L, we obtain the posterior distribution: $\mathcal{P}(\mathbf{f}_L|Y_L) = \prod_{i=1}^{l} \mathcal{P}(y_i|f_{\mathbf{x}_i})\mathcal{N}(0, \Sigma_{LL})/P(Y_L)$ where we use Σ_{LL} as a shorthand for $\Sigma_{X_L X_L}$. The normalization factor $\mathcal{P}(Y_L) = \int \mathcal{P}(Y_L|\mathbf{f}_L)\mathcal{P}(\mathbf{f}_L)d\mathbf{f}_L$ is known as the evidence for the model parameters (such as parameters of the covariance function and the noise level σ_n^2).

The posterior distribution above is non-Gaussian. To preserve computational tractability, a family of inference techniques can be applied to approximate the posterior as a Gaussian. Some popular methods include Laplace approximation, mean-field methods, and expectation propagation (EP). In this paper we will use the EP procedure, based on empirical findings that it outperforms Laplace approximation. In EP, the key idea is to approximate the non-Gaussian part in the posterior in the form of an un-normalized Gaussian, i.e. $\prod_{i=1}^{l} \mathcal{P}(y_i|f_{\mathbf{x}_i}) \approx c\, \mathcal{N}(\mu, A)$. The parameters c, μ, A are obtained by locally minimizing the Kullback-Liebler divergence between the posterior and its approximation.

With a Gaussian approximation to the posterior in hand, the distribution of the latent variable $f_{\mathbf{x}_t}$ at a test point $\mathbf{x}_t$, as given by the following integral, becomes a tractable quantity: $\mathcal{P}(f_{\mathbf{x}_t}|Y_L) = \int \mathcal{P}(f_{\mathbf{x}_t}|\mathbf{f}_L)\mathcal{P}(\mathbf{f}_L|Y_L)d\mathbf{f}_L \approx \mathcal{N}(\mu_t, \sigma_t^2)$, where $\mu_t = \Sigma_{Lt}^T \Sigma_{LL}^{-1}\, \mu$ and $\sigma_t^2 = K(\mathbf{x}_t, \mathbf{x}_t) - \Sigma_{Lt}^T\, (\Sigma_{LL}^{-1} - \Sigma_{LL}^{-1}\ A\ \Sigma_{LL}^{-1})\ \Sigma_{Lt}$. Here, the column vector $\Sigma_{Lt} = [K(\mathbf{x}_t, \mathbf{x}_1), \ldots, K(\mathbf{x}_t, \mathbf{x}_l)]^T$ and Σ_{LL} is the gram matrix of K over the labeled inputs. The conditional distribution $\mathcal{P}(f_{\mathbf{x}_t}|\mathbf{f}_L)$ is a multi-variate Gaussian with mean $\Sigma_{Lt}^T \Sigma_{LL}^{-1}\mathbf{f}_L$ and covariance matrix $K(\mathbf{x}_t, \mathbf{x}_t) - \Sigma_{Lt}^T \Sigma_{LL}^{-1} \Sigma_{Lt}$.

Finally, one can compute the Bernoulli distribution over the test label y_t, which for the probit noise model becomes $\mathcal{P}(y_t|Y_L) = \Phi\left(\mu_t/\sqrt{\sigma_n^2 + \sigma_t^2}\right)$. For more details on Gaussian processes, we point the reader to [Rasmussen and Williams, 2006] and references therein.

Our interest now is to utilize unlabeled data for Gaussian process inference.

3 Semi-supervised Gaussian Processes

3.1 Kernels for Semi-supervised Learning

A symmetric positive semi-definite function $K(\cdot, \cdot)$ can serve as the covariance function of a Gaussian process and also as a kernel function of a deterministic Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of functions $\mathcal{X} \rightarrow \Re$. The RKHS and the GP associated with the same function $K(\cdot, \cdot)$ are closely related through a classical isometry between $\mathcal{H}$ and a Hilbert space of random variables spanned by the GP (i.e. random variables of the form $\sum_k \alpha_k f_{\mathbf{x}_k}$ and their mean square limits)[1]. In this section, we review the construction of an RKHS that is adapted for semi-supervised learning of deterministic classifiers [Sindhwani *et al.*, 2005]. The kernel of this RKHS is then used as the covariance function of SSGP.

In the context of learning deterministic classifiers, for many choices of kernels, the norm $\|f\|_{\mathcal{H}}$ can be interpreted as a smoothness measure over functions f in $\mathcal{H}$. The norm can then be used to impose a complexity structure over $\mathcal{H}$ and

[1]For an RKHS function f, the notation $f(\mathbf{x})$ means the function f evaluated at $\mathbf{x}$, whereas for a GP $f_{\mathbf{x}}$ refers to the latent random variable associated with $\mathbf{x}$.

learning algorithms can be developed based on minimizing functionals of the form: $V(f, X_L, Y_L) + \gamma\|f\|^2_{\mathcal{H}}$, where V is a loss function that measures how well f fits the data. Remarkably, for loss functions that only involve f through point evaluations, a representer theorem states that the minimizer is of the form $f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$ so that only the α_i remain to be computed. This provides the algorithmic basis of algorithms like Regularized Least Squares (RLS) and Support Vector Machines (SVM), for squared loss and hinge loss, respectively. In a semi-supervised setting, unlabeled data may suggest alternate measures of complexity, such as smoothness with respect to data manifolds or clusters. Thus, if the space $\mathcal{H}$ contains functions that are smooth in these respects, it is only required to re-structure $\mathcal{H}$ by refining the norm using labeled and labeled data X_D. A general procedure to perform this operation is as follows: we define $\tilde{\mathcal{H}}$ to be the space of functions from $\mathcal{H}$ with the modified data-dependent inner product: $\langle f, g\rangle_{\tilde{\mathcal{H}}} = \langle f, g\rangle_{\mathcal{H}} + \mathbf{f}^T M \mathbf{g}$, where $\mathbf{f}$ and $\mathbf{g}$ are the vectors $\{f(\mathbf{x})\}_{\mathbf{x}\in X_D}$ and $\{g(\mathbf{x})\}_{\mathbf{x}\in X_D}$ respectively, and M is a symmetric positive semi-definite matrix. The norm induced by this modified inner product combines the original *ambient* smoothness with an *intrinsic* smoothness measure defined in terms of the matrix M. The definition of M is based on the construction of a data adjacency graph that acts as an empirical substitute of the intrinsic geometry of the marginal $\mathcal{P}_{\mathcal{X}}$. M can be derived from the graph Laplacian L; for example, $M = L$ or $M = \sum_p \beta_p L^p$ are popular choices for families of graph regularizers. The Laplacian matrix of the graph implements an empirical version of the Laplace-Beltrami operator when the underlying space is a Riemannian manifold. This operator measures smoothness with respect to the manifold. The space $\tilde{\mathcal{H}}$ can be shown to be an RKHS. With the new data-dependent norm, $\tilde{\mathcal{H}}$ becomes better suited, as compared to the original function space $\mathcal{H}$, for semi-supervised learning tasks where the cluster/manifold assumptions hold. The form of the new kernel $\tilde{K}$ associated with $\tilde{\mathcal{H}}$ can be derived (see [Sindhwani *et al.*, 2005]) in terms of the kernel function K using reproducing properties of an RKHS and orthogonality arguments, and is given by:

$$\tilde{K}(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{z}) - \Sigma^T_{D\mathbf{x}}(I + M\Sigma_{DD})^{-1} M \Sigma_{D\mathbf{z}} \quad (1)$$

where $\Sigma_{D\mathbf{x}}$ (and similarly $\Sigma_{D\mathbf{z}}$) denotes the column vector: $[K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_{l+u}, \mathbf{x})]^T$. Laplacian SVM and Laplacian RLS solve regularization problems in the RKHS $\tilde{\mathcal{H}}$ with hinge and squared loss respectively.

3.2 Data-dependent Conditional Prior

SSGP uses the semi-supervised kernel function $\tilde{K}$ defined above as a covariance function for Gaussian process learning. Thus, in SSGP the covariance between $f_{\mathbf{x}_i}$ and $f_{\mathbf{x}_j}$ not only depends on the ambient coordinates of $\mathbf{x}_i$ and $\mathbf{x}_j$, but also on geometric properties of the set X_D. In this section, we discuss the derivation of $\tilde{K}$ from Bayesian considerations.

Our general approach is summarized as follows. We begin with a standard Gaussian process. Given unlabeled and labeled data X_D, we define a joint probability distribution over the associated latent process variables $\mathbf{f}_D$ and an abstract collection of random variables, denoted as $\mathcal{G}$, whose instantiation is interpreted as realization of a certain geometry. Then, conditioning on the geometry of unlabeled data through these variables, a Bayesian update gives a posterior over the latent variables. SSGP is the resulting posterior process, which incorporates the localized spatial knowledge of the data via Bayesian learning.

There are many ways to define an appropriate likelihood evaluation for the geometry variables $\mathcal{G}$. One simple formulation for $\mathcal{P}(\mathcal{G}|\mathbf{f}_D)$ is given by:

$$\mathcal{P}(\mathcal{G}|\mathbf{f}_D) = \frac{1}{\mathcal{Z}} \exp\left(-\frac{1}{2}\mathbf{p}_D^T \,\mathbf{M}\, \mathbf{p}_D\right) \quad (2)$$

where $\mathbf{p}_D = [\mathcal{P}(y = 1|f_{\mathbf{x}_1}), \ldots, \mathcal{P}(y = 1|f_{\mathbf{x}_{l+u}})]^T = [\Phi(f_{\mathbf{x}_1}), \ldots, \Phi(f_{\mathbf{x}_{l+u}})]^T$ is a column vector of conditional label probabilities for latent variables associated with X_D, $\mathbf{M}$ is a graph-based matrix such as the graph Laplacian in Section 3.1, and $\mathcal{Z}$ is a normalization factor. $\mathcal{P}(\mathcal{G}|\mathbf{f}_D)$ may be interpreted as a measure of how much $\mathbf{f}_D$ corroborates with a given geometry, computed in terms of the smoothness of the associated conditional distributions with respect to unlabeled data. This implements a specific assumption about the connection between the true underlying unknown conditional and marginal distributions – that if two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ are *close* in the *intrinsic* geometry of $\mathcal{P}_{\mathcal{X}}$, then the conditional distributions $\mathcal{P}(y|\mathbf{x}_1)$ and $\mathcal{P}(y|\mathbf{x}_2)$ are similar, i.e., as a function of x, $\mathcal{P}(y|\mathbf{x})$ is smooth with respect to $\mathcal{P}_{\mathcal{X}}$.

Since the likelihood form (Eqn 2) leads to a non-Gaussian posterior, a natural substitute is to use an un-normalized Gaussian form, as commonly used with EP approximations. The posterior is then approximated as:

$$P(\mathbf{f}_D|\mathcal{G}) \approx c\, \exp\left(-\frac{1}{2}\mathbf{f}_D^T\, M\, \mathbf{f}_D\right) \mathcal{P}(\mathbf{f}_D)/P(\mathcal{G}) \quad (3)$$

Such a form captures the functional dependency between the latent and the geometry variables while rendering subsequent computations tractable. The value of c is inconsequential in defining this approximate posterior distribution (since c cancels due to the normalizing term $P(\mathcal{G})$), although it is important for evidence computations. In Section 3.3, we further comment on the role of the c and propose the use of partial evidence for model selection. The matrix M is approximated by the Laplacian matrix of the data-adjacency graph, to correspond to the deterministic algorithms introduced in [Sindhwani *et al.*, 2005; Belkin *et al.*, 2006]. We note that c, M can alternatively be computed from EP calculations, leading to a novel graph regularizer and tractable computations for the full evidence. We outline more details in this direction in [Chu *et al.*, 2006].

To proceed further, we make the assumption that given $\mathbf{f}_D$, $\mathcal{G}$ is independent of latent variables at other points, i.e., if the data set X contains X_D and a set of unseen test points X_T, we have:

$$\mathcal{P}(\mathcal{G}|\mathbf{f}_X) = \mathcal{P}(\mathcal{G}|\mathbf{f}_D) \quad (4)$$

This assumption allows out of sample extension without the need to recompute the graph for a new dataset.

The posterior distribution of $\mathbf{f}_X$ given $\mathcal{G}$ is: $\mathcal{P}(\mathbf{f}_X|\mathcal{G}) \propto \mathcal{P}(\mathcal{G}|\mathbf{f}_X)\mathcal{P}(\mathbf{f}_X) = \mathcal{P}(\mathcal{G}|\mathbf{f}_D)\mathcal{P}(\mathbf{f}_X)$.

The prior distribution for $\mathbf{f}_X$ is a Gaussian $\mathcal{N}(0, \Sigma_{XX})$ given by the standard Gaussian processes. In the form of block matrices, the prior $\mathcal{P}(\mathbf{f}_X)$ can be written as follows: $\mathbf{f}_X = \begin{bmatrix} \mathbf{f}_D \\ \mathbf{f}_T \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{DD} & \Sigma_{DT} \\ \Sigma_{DT}^T & \Sigma_{TT} \end{bmatrix}\right)$, where we use the shorthand D for X_D and T for X_T. The posterior distribution of $\mathbf{f}_X$ conditioned on $\mathcal{G}$ can be written as a zero-mean Gaussian distribution $\mathcal{P}(\mathbf{f}_X|\mathcal{G}) \propto \exp(-\frac{1}{2}\mathbf{f}_X^T \tilde{\Sigma}_{XX}^{-1} \mathbf{f}_X)$ where

$$\tilde{\Sigma}_{XX}^{-1} = \begin{bmatrix} \Sigma_{DD} & \Sigma_{DT} \\ \Sigma_{DT}^T & \Sigma_{TT} \end{bmatrix}^{-1} + \begin{bmatrix} M & 0 \\ 0 & 0 \end{bmatrix}$$

Proposition: Given Eqn 3, for any finite collection of data points X, the random variables $\mathbf{f}_X = \{f(\mathbf{x})\}_{\mathbf{x}\in X}$ *conditioned on* $\mathcal{G}$ have a multivariate normal distribution $\mathcal{N}(0, \tilde{\Sigma}_{XX})$, where $\tilde{\Sigma}_{XX}$ is the covariance matrix whose elements are given by evaluating the following kernel function $\tilde{K} : \mathcal{X} \times \mathcal{X} \mapsto \Re$,

$$\tilde{K}(\mathbf{x}, \mathbf{z}) = K(\mathbf{x}, \mathbf{z}) - \Sigma_{D\mathbf{x}}^T (I + M\Sigma_{DD})^{-1} M \Sigma_{D\mathbf{z}} \quad (5)$$

for $\mathbf{x}, \mathbf{z} \in \mathcal{X}$. Here $\Sigma_{D\mathbf{x}}$ denotes the column vector $[K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x})]^T$.

This proposition shows that the Gaussian process conditioned on the geometry variable $\mathcal{G}$ is a Gaussian process with a modified covariance function $\tilde{K}$. A proof of this result uses straightforward matrix algebra and is omitted for brevity.

SSGP is the posterior process obtained by conditioning the original GP with respect to $\mathcal{G}$. Note that the form of its covariance function $\tilde{K}$ is the same as in Eqn 1, which is derived from properties of RKHS.

3.3 Model Selection

Model selection for SSGP involves choosing the kernel parameters and the noise variance σ_n (see Section 2). The definition of $\tilde{K}$ is based on the choice of a covariance function K and a graph regularization matrix M. As in [Sindhwani *et al.*, 2005], in this paper we restrict our attention to the following choices: $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right)$, $M = \frac{\gamma_I}{\gamma_A} L^p$ and use the kernel $\frac{1}{\gamma_A}\tilde{K}$. The parameters γ_A and γ_I balance ambient and intrinsic covariance. The parameters related to the computation of the graph Laplacian L are – the number of nearest neighbors NN and the graph adjacency matrix W. We set $W_{ij} = \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma_t^2}\right)$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are adjacent in a NN-nearest neighbor graph and zero otherwise.

Denote ϕ as the collection of model parameters. The optimal values of ϕ are determined by maximizing the evidence $\mathcal{P}(Y_L|\mathcal{G}, \phi)$ which is available from EP computations.

Model selection is particularly challenging in semi-supervised settings in the presence of few labeled examples. Popular model selection techniques like cross-validation (for deterministic methods like Laplacian SVM/RLS) require subsets of the already-scarce labeled data to be held out from training. Moreover, for graph-based semi-supervised learning, a proper cross-validation should also exclude the held out subset from the adjacency graph instead of simply suppressing labels, in order to ensure good out-of-sample extension. This adds a major expense since the semi-supervised kernel in Eqn 1 needs to be recomputed multiple times. Also, since cross-validation is based on counts, it may often be unable to uniquely identify a good set of parameters. By contrast, evidence maximization neither needs to suppress expensive labels nor needs to hold out labeled examples from the graph. As a continuous function of many model parameters, it is more precise in parameter selection (demonstrated in Section 4) and also amenable to gradient-based techniques. In addition to these benefits, we note the novel possibility of employing unlabeled data for model selection by maximizing the full evidence $\mathcal{P}(Y_L, \mathcal{G}|\phi) = \mathcal{P}(Y_L|\mathcal{G}, \phi)\mathcal{P}(\mathcal{G}|\phi)$. The latter term may be computed as $\mathcal{P}(\mathcal{G}|\phi) = \int \mathcal{P}(\mathcal{G}|\mathbf{f}_D)\mathcal{P}(\mathbf{f}_D)d\mathbf{f}_D$. Given Eqn 3, one can immediately derive $\log \mathcal{P}(\mathcal{G}|\phi) = \log\ c - \frac{1}{2}\log\det(M\Sigma_{DD} + I)$ where I is the identity matrix. For simplicity, in this paper we focus on comparing standard evidence maximization in SSGP with standard cross-validation in Laplacian SVM/RLS.

Table 1: Datasets with d features, n training examples (labeled+unlabeled) and t test examples.

DATASET	d	n	t	DOMAIN
MOONS	2	120	40401	SYNTHETIC
3VS5	256	1214	326	USPS DIGIT
SET1	241	1126	374	IMAGE RECOGNITION
PCMAC	7511	1460	486	20-NEWGROUPS
REUTERS	9930	458	152	REUTERS-21578

4 Experiments

We performed experiments on a synthetic two-dimensional dataset and four real world datasets for binary classification problems. The statistics of these datasets are described in Table 1.

Synthetic Data

The toy 2-D dataset MOONS, meant to visually demonstrate SSGP, is shown in Figure 1 as a collection of unlabeled examples (small black dots) and two labeled examples per class (large colored points). The classification boundary and the contours of the mean of the predictive distribution over the unit square are shown for standard GP and SSGP (Figure 1(a,c)). These plots demonstrate how SSGP utilizes unlabeled data for classification tasks. The uncertainty in prediction (Figure 1(d)) respects the geometry of the data, increasing anisotropically as one moves away from the labeled examples. Figure 1(b) also demonstrates evidence-based model selection.

Real World Data sets

For all the real world data sets (except 3VS5), we generated a fixed training/test split by randomly drawing one-fourth of the original data set into a test set of unseen examples, and retaining the rest as a training set of labeled and unlabeled examples. Learning curves were plotted for varying amount of labeled data, randomly chosen from the training set. For 3VS5, we chose the exact data splits used in [Lawrence and Jordan, 2004] to facilitate comparison between algorithms.

To reduce the complexity of model selection, we fix $NN = 10$ and $p = 2$ for 3VS5, SET1 and $NN = 50$ and $p = 5$ for

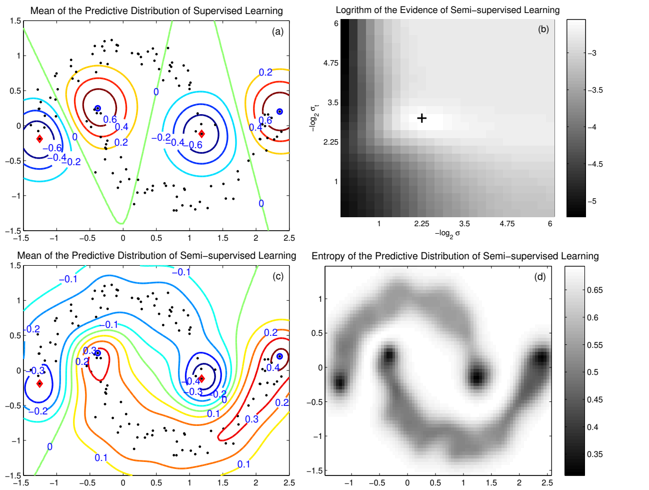

Figure 1: MOONS: The mean of the predictive distribution of supervised GP is shown in graph (a). Based on maximum evidence, the best settings are found at $\log_2 \sigma_t = -3.0$ and $\log_2 \sigma = -2.25$, shown as a cross in graph (b). The results of SSGP with the best settings are shown in graph (c) and (d).

PCMAC, REUTERS. These values are based on experimental experience from [Sindhwani *et al.*, 2005] with image and text data sets. We used weighted graphs with Euclidean distance as the graph similarity measure, and Gaussian weights with width (σ_t) set to the mean distance between adjacent vertices. The behavior of evidence-based model selection for SSGP and cross-validation (CV) based model selection for Laplacian SVM was investigated over a range of values for the parameters $\sigma, \gamma_A, \gamma_I$. For each dataset, we computed the mean length of feature vectors (σ_0) in the training set, and probed σ in the range $[\ \frac{\sigma_0}{4}\ \frac{\sigma_0}{2}\ \sigma_0\ 2\sigma_0\ 4\sigma_0\ ]$; the range for γ_A was $[10^{-6}\ 10^{-4}\ 10^{-2}\ 1\ 100]$ and the choices for ratio $\frac{\gamma_I}{\gamma_A}$ were $[0\ 1\ 10\ 100\ 1000]$ (note that a choice of 0 ignored unlabeled data and reduces the algorithm to its standard supervised mode). The noise parameter σ_n in the probit model for class labels was set to 10^{-4} in all experiments. Note that this parameter allows SSGP to potentially deal with label noise.

1. BENEFIT OF UNLABELED DATA: In Figure 2, we plot the mean error rates on the test sets for the four datasets as a function of the number of labeled examples (expressed in terms of probability of finding a labeled example in the training set) for SSGP and standard supervised GP (which ignores unlabeled data). Figure 3 shows the corresponding curves for performance on the unlabeled data. The solid curves show the mean and standard deviation of the minimum error rates for SSGP and GP over the set of parameters $\sigma, \gamma_A, \gamma_I$ ($\gamma_I = 0$ for GP) and the dashed curves show the mean and standard deviation of error rates at parameters chosen by maximizing evidence. The mean and standard deviations are computed over 10 random draws for every choice of the amount of labeled data. We make several observations: (a) By utilizing unlabeled data, SSGP makes significant performance improvements over standard GP on both the test and unlabeled sets, the gap being larger on the unlabeled set. This holds true across the entire span of varying amounts of labeled data for

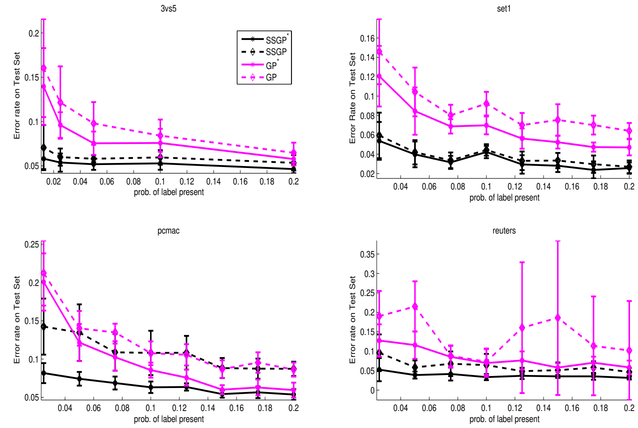

Figure 2: Comparison of SSGP with GP on Test Data.

all datasets, except PCMAC where the gap seems to converge faster. (b) Performance based on parameters chosen by evidence based model selection is much closer to the optimal performance for SSGP than for GP. The performance curves for SSGP have lesser variance.

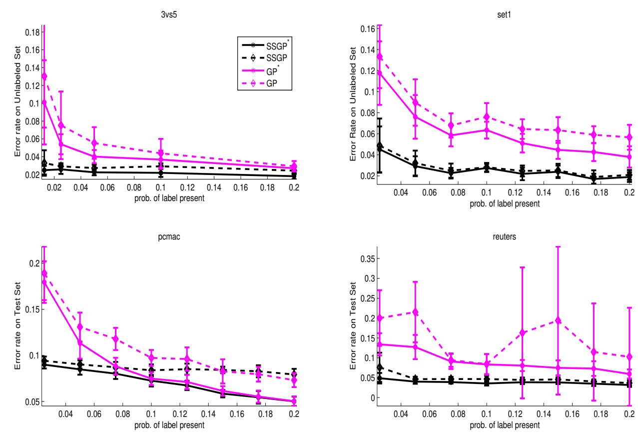

Figure 3: SSGP vs GP on Unlab. Data (Transduction).

2. EVIDENCE MAXIMIZATION VERSUS CV: We next compare evidence-based model selection in SSGP with CV based model selection in Laplacian SVM. In Figure 4, we plot the mean error rates on the test sets for the four datasets as a function of the number of labeled examples for SSGP and Laplacian SVM, both of which utilize unlabeled data through the same kernel. Figure 5 shows the corresponding curves for performance on the unlabeled data. The solid curves show the mean and standard deviation of the minimum error rates for SSGP and Laplacian SVM over the set of parameters $\sigma, \gamma_A, \gamma_I$; and the dashed curves show corresponding curves at a parameter setting chosen by evidence-maximization and CV for SSGP and Laplacian SVM respectively. The mean and standard deviations are computed over 10 random draws for every choice of the amount of labeled data. We used 5-fold CV for each labeled set, except those with 10 or fewer examples, in which case leave-one-out cross validation was used. It is important to note that our CV protocol for Laplacian SVMs only suppresses labels but does not exclude the labeled examples from the graph. We make

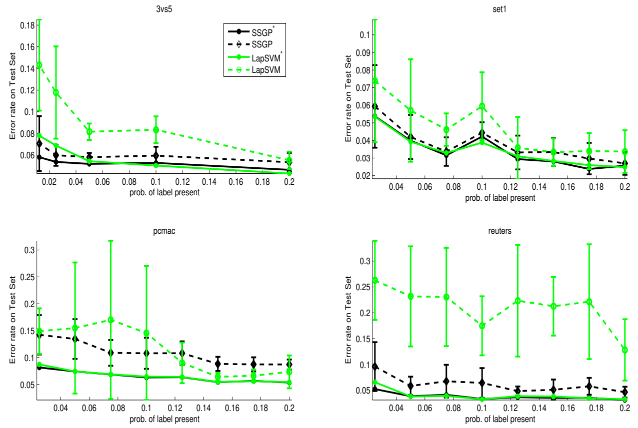

Figure 4: Comparison of SSGP with LapSVM on Test Data.

the following observations: (a) In terms of best performance over the range of parameters (i.e the solid curves), SSGP and Laplacian SVM return nearly identical mean error rates, except in 3VS5, where when given very few labeled examples, SSGP outperforms LapSVM. (b) By looking at the performance curves at parameters chosen by their corresponding model selection strategies (i.e the dashed curves), the superiority of evidence maximization in SSGP over CV in Laplacian SVM becomes quite evident. This is true for both test and unlabeled set performance. (c) The quality of CV based performance of Laplacian SVM is significantly better over unlabeled data as compared to that over test data, indicating that CV drives the model selection towards parameters that favor transduction at the expense of semi-supervised induction. These experiments show that evidence maximization returns significantly better performance than CV on both the test and unlabeled sets.

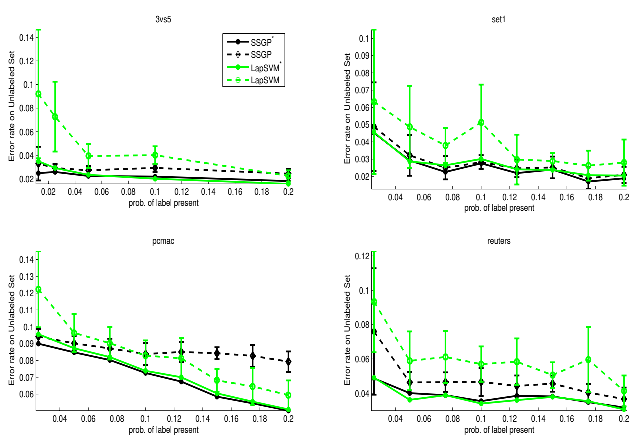

Figure 5: SSGP vs LapSVM on Unlab. Data (Transduction).

3. COMPARISON WITH OTHER METHODS: Finally, in Figure 6, we superimpose the SSGP performance curves over the results of the Null Category Noise Model (NCNM), Informative Vector Machine (IVM), SVM, and transductive SVM (TSVM) plotted in [Lawrence and Jordan, 2004]. The same experimental protocol and data splits were used. We are encouraged to see that SSGP outperforms all algorithms tested in this experiment.

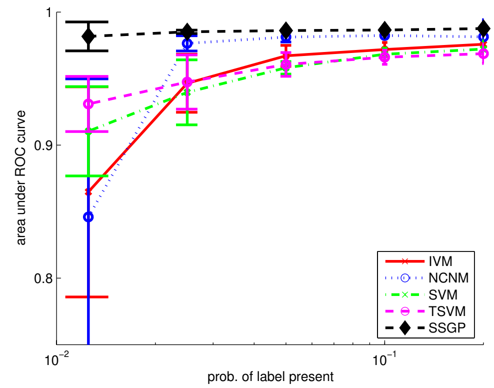

Figure 6: SSGP versus NCNM, TSVM, IVM and SVM

5 Conclusion

To conclude, we have presented semi-supervised Gaussian process classifiers based on a data-dependent covariance function that is adapted to the geometry of unlabeled data. We have empirically demonstrated the utility of SSGP on several learning tasks, and observed the benefits of evidence-based model selection in comparison to cross-validation techniques in Laplacian SVM/RLS.

References

[Belkin *et al.*, 2006] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* (to appear), 2006.

[Chu *et al.*, 2006] W. Chu, V. Sindhwani, Z. Gharamani, and S. S. Keerthi. Relational learning with gaussian processes. In *Neural Information Processing Systems 19*, 2006.

[Kapoor *et al.*, 2005] A. Kapoor, Y. Qi, H. Ahn, and R. Picard. Hyperparameter and kernel learning for graph-based semi-supervised classification. In *Neural Information Processing Systems 18*, 2005.

[Krishnapuram *et al.*, 2004] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. A. T. Figueiredo. On semi-supervised classification. In *Neural Information Processing Systems 17*, 2004.

[Lawrence and Jordan, 2004] N. D. Lawrence and M. Jordan. Semi-supervised learning via Gaussian processes. In *Neural Information Processing Systems 17*, 2004.

[Rasmussen and Williams, 2006] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[Sindhwani *et al.*, 2005] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *International Conference on Machine Learning*, 2005.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning: From Gaussian fields to Gaussian processes. Technical Report CMU-CS-03-175, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 2003.