Nonparametric Bayesian Kernel Models

Feng Liang

Joint work with Ming Liao, Sayan Mukherjee, Mike West

Institute of Statistics and Decision Sciences
Duke University

NISS Affiliate Annual Meeting, Athens GA, 2006.

Outline

Outline

- Assumptions: $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim P(\boldsymbol{x}, y)$.
- Find a function $f(\boldsymbol{x}) \longrightarrow y$.
- Risk : $R[f] = \mathbb{E}L(Y, f(\boldsymbol{X}))$

$$f^*(\boldsymbol{x}) = \text{argmin}_f R[f]$$

e.g., $f^*(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$ under squared error loss.

- Empirical risk minimization

$$\hat{f}_n(\boldsymbol{x}) = \text{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i))$$

Regularization

- Regularization:

$$\min_{f \in \mathcal{F}} \hat{R}_n[f] + \lambda \Omega[f]$$

- Consider a linear regression model: $f(\boldsymbol{x}) = \boldsymbol{x}^t \beta$,

$$\hat{\beta}_{OLS} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_i (y_i - \boldsymbol{x}_i^t \beta)^2.$$

- Large p small n: $p >> n$.
- $\Omega(\beta) = \|\beta\|^2$ (ridge); $|\beta|$ (LASSO).

Reproducing Kernel Hilbert Space (RKHS)

- $\mathcal{H}_K = \{f(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$.
- $K(\cdot, \cdot)$ is a semi-positive definite bivariate symmetric function defined on $\mathcal{X} \times \mathcal{X}$, i.e.

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0, \quad \forall a_i \in \mathbb{R}, \forall \boldsymbol{x}_i \in \mathcal{X}, \forall m \in \mathbb{N}.$$

- Denote $k_{\boldsymbol{u}}(\cdot) = K(\boldsymbol{u}, \cdot)$.

$$\mathcal{H}_K = \overline{\text{span}\{k_{\boldsymbol{u}}(\cdot), \boldsymbol{u} \in \mathcal{X}\}}.$$

- Reproducing property: $\forall f \in \mathcal{H}_K, \forall \boldsymbol{x}_0 \in \mathcal{X}$

$$f(\boldsymbol{x}_0) = \langle f, k_{\boldsymbol{x}_0} \rangle = \langle f, K(\boldsymbol{x}_0, \cdot) \rangle.$$

A Simple Examples of RKHS

- Consider all lineal functions in $\mathbb{R}^2$ passing the origin, i.e.

$$\mathcal{H}_K = \{f_\theta(\boldsymbol{x}) = \theta^6 \boldsymbol{x} = \theta_1 x_1 + \theta_2 x_2,\ \theta \in \mathbb{R}^2\}$$

  with

$$\langle f_\theta, f_\lambda \rangle = \theta^t \lambda, \quad K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^t \boldsymbol{x}' = x_1 x_1' + x_2 x_2'.$$

- Reproducing property : $\forall f_\theta$,

$$f_\theta(\boldsymbol{x}_0) = \theta^t \boldsymbol{x}_0 = \langle f_\theta, f_{\boldsymbol{x}_0} \rangle.$$

- Semi-positive definite :

$$\sum_i \sum_j a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Big(\sum_i a_i \boldsymbol{x}_i\Big)^t \cdot \Big(\sum_j a_j \boldsymbol{x}_j\Big) = \|\sum_i a_i \boldsymbol{x}_i\|^2$$

More Examples of RKHS

- Any RKHS is a space of *linear* functions and $K(\cdot, \cdot)$ represents a dot product

More Examples of RKHS

- Any RKHS is a space of *linear* functions and $K(\cdot,\cdot)$ represents a dot product in the feature space.

More Examples of RKHS

- Any RKHS is a space of *linear* functions and $K(\cdot,\cdot)$ represents a dot product in the feature space.
- Feature mapping ϕ: $\boldsymbol{x} \in \mathbb{R}^p \longrightarrow \phi(\boldsymbol{x}) \in \mathbb{R}^{\mathbb{N}}$.

$$(x_1, x_2) \in \mathbb{R}^2 \longrightarrow (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2) \in \mathbb{R}^6$$

- $\mathcal{H}_K = \{f_\theta(\boldsymbol{x}) = \theta^t \cdot \phi(\boldsymbol{x})\}$ and $K(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^t \cdot \phi(\boldsymbol{x}')$.
- Examples
  - Polynomial kernels: $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^t \boldsymbol{x}')^d$.
  - Gaussian radial basis kernels:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\Big(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\sigma^2}\Big).$$

Representer Theorem

(Kimeldorf and Wahba, 1971)

$$f^*(\boldsymbol{x}) \quad = \quad \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_K}$$

Representer Theorem

(Kimeldorf and Wahba, 1971)

$$
\begin{aligned}
f^*(\boldsymbol{x}) &= \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \\
&= \sum_{i=1}^{n} \alpha_i k_{\boldsymbol{x}_i}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)
\end{aligned}
$$

Representer Theorem

(Kimeldorf and Wahba, 1971)

$$
\begin{aligned}
f^*(\boldsymbol{x}) &= \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \\
&= \sum_{i=1}^{n} \alpha_i k_{\boldsymbol{x}_i}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i)
\end{aligned}
$$

Proof : Let $\mathcal{H}_1 = \text{span}\{k_{\boldsymbol{x}_1}, \ldots, k_{\boldsymbol{x}_n}\}$ and $\mathcal{H}_2 = \mathcal{H}_1^{\perp}$,

$$f = f_1 + f_2, \quad \mathcal{H}_K = \mathcal{H}_1 \oplus \mathcal{H}_2$$

- $\|f\|^2 \geq \|f_1\|^2$
- $f(\boldsymbol{x}_i) = f_1(\boldsymbol{x}_i)$ because

$$\langle f, k_{\boldsymbol{x}_i} \rangle = \langle f_1 + f_2, k_{\boldsymbol{x}_i} \rangle = \langle f_1, k_{\boldsymbol{x}_i} \rangle$$

Outline

Primer on Bayesian Analysis

- Data and a parametric family: $\boldsymbol{y} \sim p(\cdot \mid \theta)$

Primer on Bayesian Analysis

- Data and a parametric family: $\boldsymbol{y} \sim p(\cdot \mid \theta)$
- Prior on θ: $\pi(\theta)$

Primer on Bayesian Analysis

- Data and a parametric family: $\boldsymbol{y} \sim p(\cdot \mid \theta)$
- Prior on θ: $\pi(\theta)$
- Joint distribution on $(\boldsymbol{y}, \theta)$: $p(\boldsymbol{y} \mid \theta)\pi(\theta)$

Primer on Bayesian Analysis

- Data and a parametric family: $\boldsymbol{y} \sim p(\cdot \mid \theta)$
- Prior on θ: $\pi(\theta)$
- Joint distribution on $(\boldsymbol{y}, \theta)$: $p(\boldsymbol{y} \mid \theta)\pi(\theta)$
- Posterior inference:

$$\pi(\theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \theta)\pi(\theta)}{\int p(\boldsymbol{y} \mid \theta)\pi(\theta)d\theta} \propto p(\boldsymbol{y} \mid \theta)\pi(\theta)$$

Connection to Regularization

- Log posterior $= \log p(\boldsymbol{y} \mid \theta) + \log \pi(\theta) + \ldots$

$$\text{Posterior Mode} = \text{argmin} \sum_i (y_i - f_\theta(\boldsymbol{x}_i))^2 - c \log \pi(\theta)$$

- Regularization

$$\text{argmin} \sum_{i=1} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \Omega[f]$$

Connection to Regularization

- Log posterior $= \log p(\boldsymbol{y} \mid \theta) + \log \pi(\theta) + \ldots$

$$\text{Posterior Mode} = \text{argmin} \sum_i (y_i - f_\theta(\boldsymbol{x}_i))^2 - c \log \pi(\theta)$$

- Regularization

$$\text{argmin} \sum_{i=1} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \Omega[f]$$

- $\pi(\theta) \propto \exp\{-\lambda \Omega[f]\}$. For example, when $f_\theta = \theta^t \boldsymbol{x}$,
  - Ridge ($\|\theta\|^2$) $\Longleftrightarrow$ normal
  - LASSO ($|\theta|$) $\Longleftrightarrow$ double exponential

Previous Work

For example, Tipping 2001, Chakraborty et al. 2005, and others.

- Start with the finite representation from the representer Theorem:

$$\sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) \quad (*)$$

- Specify priors on the coefficients α_i's.
- Their models change when sample size changes without a coherent argument.
- Can we justify $(*)$ using the connection between regularization and posterior mode?

An Orthonormal Representation of $\mathcal{H}_K$

- For Mercer kernels,

$$K(\boldsymbol{x}, \boldsymbol{u}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{u}),$$

  where ϕ_j is a sequence of orthonormal functions and $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

- (Cucker and Smale, 2001) $\forall f \in \mathcal{H}_K$,

$$f(\boldsymbol{x}) = \sum_j \alpha_j \phi_j(\boldsymbol{x}), \quad \sum_j \frac{\alpha_j^2}{\lambda_j} < \infty.$$

  That is, $\mathcal{H}_K$ can be parameterized by

$$\mathcal{A} = \{(\alpha_j)_{j=1}^{\infty} : \sum_j \alpha_j^2 / \lambda_j < \infty\}.$$

An Overcomplete Representation of $\mathcal{H}_K$

- Recall

$$\mathcal{H}_K = \overline{\text{span}\{K(\cdot, \boldsymbol{u}), \boldsymbol{u} \in \mathcal{X}\}}$$

- Start with a larger space

$$f(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{u}) d\Gamma(\boldsymbol{u}),$$

where $\Gamma(\boldsymbol{u})$ is a sign measure on $\mathcal{X}$.

- In this talk, we focus on the following representation

$$f(\boldsymbol{x}) = \int_{\mathcal{X}} w(\boldsymbol{u}) K(\boldsymbol{x}, \boldsymbol{u}) dF(\boldsymbol{u}),$$

where $w(\boldsymbol{u})$ denotes the coefficient at location $\boldsymbol{u}$ and F denotes the distribution function of the location $\boldsymbol{u}$.

An Overcomplete Representation of $\mathcal{H}_K$

- Recall

$$\mathcal{H}_K = \overline{\text{span}\{K(\cdot, \boldsymbol{u}), \boldsymbol{u} \in \mathcal{X}\}}$$

- Start with a larger space

$$f(\boldsymbol{x}) = \int_{\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{u}) d\Gamma(\boldsymbol{u}),$$

where $\Gamma(\boldsymbol{u})$ is a sign measure on $\mathcal{X}$.

- In this talk, we focus on the following representation

$$f(\boldsymbol{x}) = \int_{\mathcal{X}} w(\boldsymbol{u}) K(\boldsymbol{x}, \boldsymbol{u}) dF_{\boldsymbol{X}}(\boldsymbol{u}),$$

where $w(\boldsymbol{u})$ denotes the coefficient at location $\boldsymbol{u}$ and $F_{\boldsymbol{X}}$ denotes the distribution function of explanatory variable $\boldsymbol{X}$.

Dirichlet Process Priors

- Beta (α, β) on $x \in [0,1]$

$$f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$$

- Dir $(\alpha_1, \ldots, \alpha_k)$ on $(x_1, \ldots, x_k)$ where $x_i \in [0,1]$ and $\sum_i x_i = 1$

$$f(x) \propto x_1^{\alpha_1-1} \cdots x_k^{\alpha_k-1}$$

- DP(α_0, F_0) on $\mathbf{F}$ (note $\mathbf{F}$ is a random distribution on $\mathcal{X}$): for any measurable partition of $\mathcal{X}$,

$$\mathbf{F}(B_1), \mathbf{F}(B_2), \ldots \mathbf{F}(B_k) \sim \text{Dir}(\alpha_1, \ldots, \alpha_k),$$

where $\alpha_i = \alpha_0 F_0(B_i)$.

A Bayesian Representer Theorem

Given $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots, x_n \sim \mathbf{F}$, with DP$(\alpha_0, F_0)$ prior on $\mathbf{F}$, the posterior distribution of $\mathbf{F}$ is DP$(\alpha_0 + n, \alpha_0 F_0 + \sum \delta_{\boldsymbol{x}_i})$. So

$$\mathbb{E}[\mathbf{F} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] = \frac{\alpha_0}{\alpha_0 + n} F_0 + \frac{1}{\alpha_0 + n} \sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}.$$

Bayesian Representer Theorem

For $f(x) = \int w(\boldsymbol{u}) K(\boldsymbol{x}, \boldsymbol{u}) d\mathbf{F}_X(\boldsymbol{u})$, under a Dirichlet prior on $\mathbf{F}_{\boldsymbol{X}}$,

$$\mathbb{E}[f(\boldsymbol{x})] \approx \sum_{i=1}^{n} w_i \; K(\boldsymbol{x}, \boldsymbol{x}_i).$$

Prior Specification

- Likelihood: $\boldsymbol{Y} \mid \mathcal{N}(w_0 + \mathbf{K}\boldsymbol{w}, \sigma^2 I_n)$ where $\mathbf{K}_{n\times n}$ be the centered kernel matrix at the n data points.
- Non-informative prior on (w_0, σ^2),

$$\pi(w_0, \sigma^2) \propto 1/\sigma^2$$

- Generalized g-prior on $\boldsymbol{w}$ (West 2002)

$$\begin{aligned} \boldsymbol{w} \mid T &\sim \mathcal{N}(0, U\Delta^{-1}T^{-1}\Delta^{-1}U^t) \\ \tau_1, \ldots, \tau_n &\sim \text{Gamma}\left(\frac{s_0}{2}, \frac{s_0}{2}v\right), \quad v \sim \text{Exp}(\alpha_0) \end{aligned}$$

  where U and Δ come from $\mathbf{K} = U\Delta U^t$.
- $\mathbf{K}\boldsymbol{w} = U\beta$, then the prior on $\boldsymbol{w}$ corresponds to a student t distribution on β.

Model Fitting via MCMC

- Gibbs sampling
  - $w_0 = \ldots\ldots$
  - Draw $\beta \sim \mathcal{N}(\ , \ )$
  - Draw $T = \text{diag}(\tau_1, \ldots, \tau_n) \sim Ga(\ , \ )$
  - Draw $v \sim Ga(\ , \ )$
- Probit model: $(\boldsymbol{x}, y) \longrightarrow (\boldsymbol{x}, y, z)$ and

$$P(y = 1) = \Phi(z).$$

Outline

Semi-supervised Learning

- Supervised learning (labelled data):

$$(\boldsymbol{x}_i, y_i) \sim P(\boldsymbol{x}, y)$$

- Unsupervised learning (unlabelled data):

$$(\boldsymbol{x}_i) \sim P_{\boldsymbol{X}}(\boldsymbol{x})$$

- Semi-supervised learning:

$$(\boldsymbol{x}_i, y_i)_{i=1}^{n} \sim P, \quad (\boldsymbol{x}_i)_{i=n+1}^{n+m} \sim P_{\boldsymbol{X}}$$

- How's it different from missing data?

The Role of Unlabelled Data

- Data $D = (\boldsymbol{X}, Y, \boldsymbol{X}^U)$ where

$$
\begin{aligned}
(\boldsymbol{X}, Y) &\sim P(\boldsymbol{x}, y) = p(\boldsymbol{x} \mid \phi)\, p(y \mid \boldsymbol{x}, \theta) \\
\boldsymbol{X}^U &\sim P_{\boldsymbol{X}}(\boldsymbol{x}) = p(\boldsymbol{x} \mid \phi)
\end{aligned}
$$

- Prediction of $\boldsymbol{y}^*$ at a new location $\boldsymbol{x}^*$,

$$
\begin{aligned}
\mathbb{E}[Y^* \mid x^*, D] &= \int y^* p(y^* \mid x^*, D) dy^* \\
&= \int y^* p(y^* \mid \boldsymbol{x}^*, \theta) \pi(\theta \mid D) d\theta dy^*.
\end{aligned}
$$

- The key to understand the role of $\boldsymbol{X}^m$ is

$$
\begin{aligned}
\pi(\theta \mid D) &= \int \pi(\theta, \phi \mid D) d\phi \\
&\propto \int p(\boldsymbol{X}, \boldsymbol{X}^U \mid \phi) p(Y \mid \boldsymbol{X}, \theta) \pi(\theta, \phi) d\phi
\end{aligned}
$$

An Intimate Relationship

- $P(\boldsymbol{x}, y) = p(\boldsymbol{x} \mid \phi) p(y \mid \boldsymbol{x}, \theta)$
- Recall that $y \mid \boldsymbol{x} \sim \mathcal{N}(f(\boldsymbol{x}), \sigma^2)$ where

$$f(\boldsymbol{x}) = \int w(\boldsymbol{u}) K(\boldsymbol{x}, \boldsymbol{u}) dF_{\boldsymbol{X}}(\boldsymbol{u}).$$

  So in our model, $\theta = (\phi, \dots)$. Therefore unlabelled data will be relevant and should be incorporated into prediction.
- By our Bayesian representor Theorem, given $(\boldsymbol{x}_i, y_i)_{i=1}^n$ and $(\boldsymbol{x}_j)_{j=n+1}^{n+m}$,

$$f(\boldsymbol{x}) \approx \sum_{i=1}^{n} w_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \sum_{j=n+1}^{n+m} w_j K(\boldsymbol{x}, \boldsymbol{x}_j).$$

Connection to Regularization

- Transductive SVM (TSVM) (Joachims, 1999)

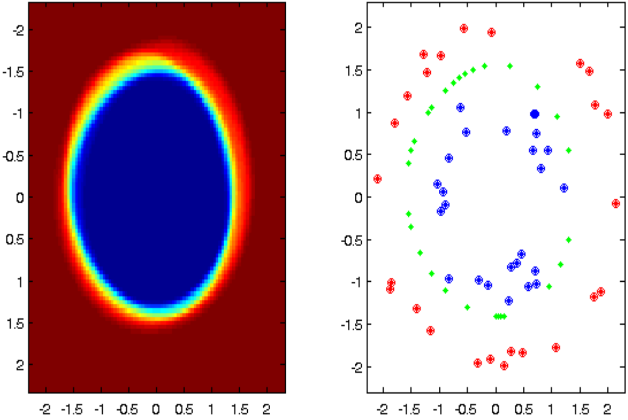
$$\text{argmin } C\sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + C^* \sum_{i=n+1}^{n+m} L(y_i, f(\boldsymbol{x}_i)) + \|f\|^2_{\mathcal{H}_K}$$

- Manifold regularization (Belkin et al, 2005)

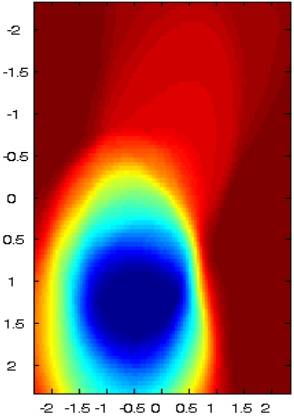
$$\text{argmin} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda_1 \|f\|^2_{\mathcal{H}_K} + \lambda_2 \|f\|^2_I,$$

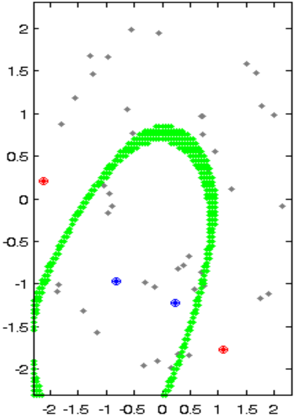
where $\|f\|^2_I$ measures the intrinsic structure of $F_{\boldsymbol{X}}$ and is approximated on all the data (include the unlabelled ones) using graph Laplacian.

A Toydata

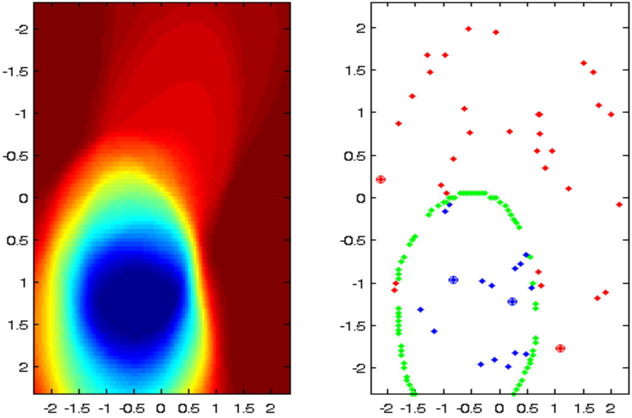

-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

A Toydata

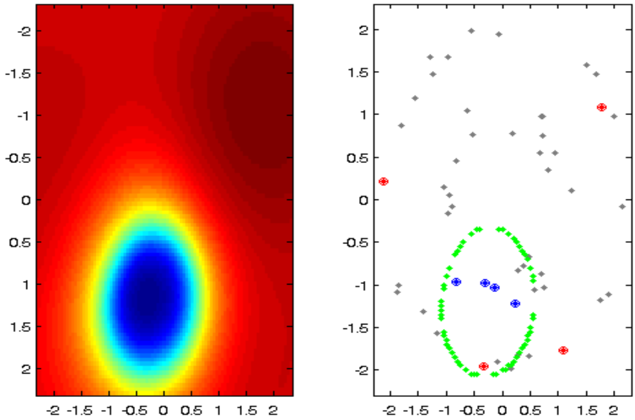

-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

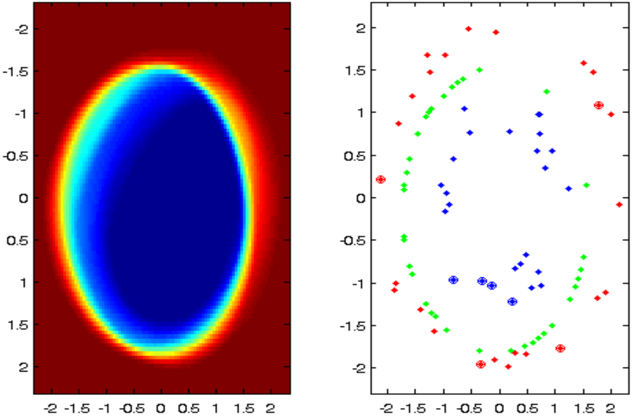
A Toydata


-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

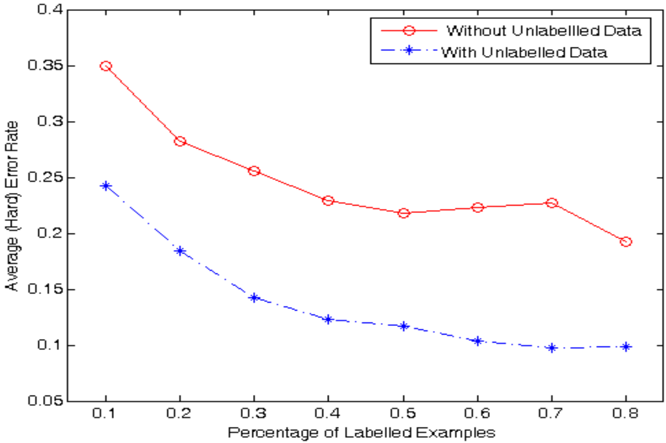
A Toydata


-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

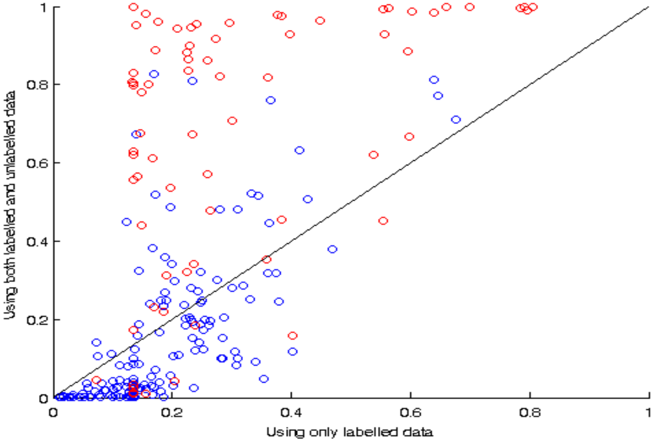
A Toydata


-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2

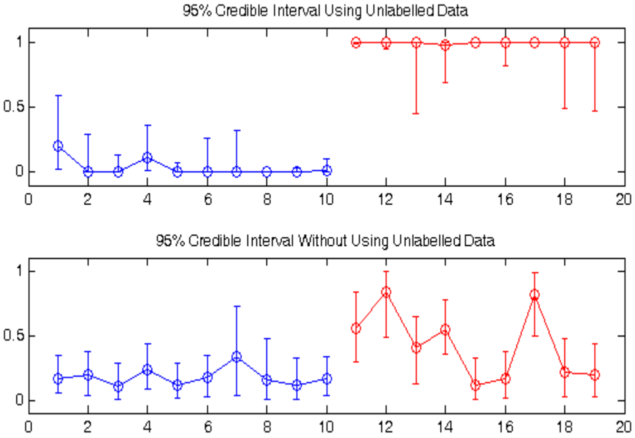
Cancer Data


Without Unlabellled Data
With Unlabelled Data
Average (Hard) Error Rate
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Percentage of Labelled Examples

Cancer Data


Using both labelled and unlabelled data
Using only labelled data
0
0.2
0.4
0.6
0.8
1

Cancer Data


95% Credible Interval Using Unlabelled Data
1
0.5
0
0 2 4 6 8 10 12 14 16 18 20
95% Credible Interval Without Using Unlabelled Data
1
0.5
0
0 2 4 6 8 10 12 14 16 18 20

Outline

Conclusion

- A coherent Bayesian approach on RKHS:
  - characterize RKHS using an overcomplete representation;
  - specify priors on the whole RKHS;
  - incorporate the relevant information from the unlabelled data.

Conclusion

- A coherent Bayesian approach on RKHS:
  - characterize RKHS using an overcomplete representation;
  - specify priors on the whole RKHS;
  - incorporate the relevant information from the unlabelled data.
- Future work:
  - Other choice of priors and sensitivity study
  - Feature selection