Finding religion: kernels and the Bayesian persuasion

Rev. Dr. Sayan Mukherjee

Department of Statistical Science
Institute for Genome Sciences & Policy
Department of Computer Science
Duke University

May 7, 2007

Not a Bayesian

What makes someone Bayesian

Is it Bayes rule ?

$$\mathbf{P}\text{rob}(\text{parameters}|\text{data}) = \frac{\text{Lik}(\text{data}|\text{paramaters}) \cdot \pi(\text{parameters})}{\mathbf{P}\text{rob}(\text{data})}.$$

What makes someone Bayesian

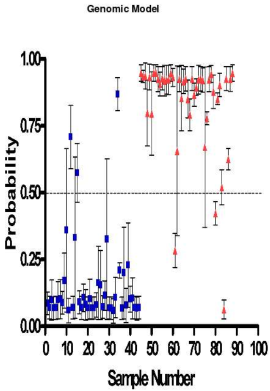
Is it Bayes rule ?

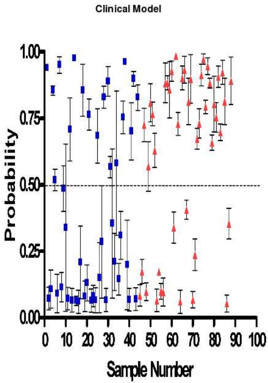
$$\mathbf{P}\mathrm{rob}(\mathrm{parameters}|\mathrm{data}) = \frac{\mathrm{Lik}(\mathrm{data}|\mathrm{paramaters}) \cdot \pi(\mathrm{parameters})}{\mathbf{P}\mathrm{rob}(\mathrm{data})}.$$

NO!!!!!!!!!!!!!!!!!!!!!! Necessary but no where near sufficient.

Why I am a Bayesian

Bayesian statistics is about embracing and formally modelling uncertainty.





A simple example

I draw points $x_1, ..., x_n$ from iid from a normal distribution and I want to know the mean and I know $\sigma = 1$.
My likelihood and prior are

$$
\begin{aligned}
\text{Lik}(x_1, ..., x_n|\mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} exp(-|x_i - \mu|^2/2) \\
\pi(\mu) &= \frac{1}{\sqrt{2\pi}} exp(-|\mu - 5|^2/2).
\end{aligned}
$$

The posterior can be computed closed form and it is a product of normals

$$
p(\mu|x_1, ..., x_n) = \frac{\text{Lik}(x_1, ..., x_n|\mu)\pi(\mu)}{\int_{-\infty}^{\infty} \text{Lik}(x_1, ..., x_n|\mu)\pi(\mu)d\mu}.
$$

Relevant papers

- Characterizing the function space for Bayesian kernel models. Natesh Pillai, Qiang Wu, Feng Liang, Sayan Mukherjee, Robert L. Wolpert. Journal Machine Learning Research, in press.
- Understanding the use of unlabelled data in predictive modelling. Feng Liang, Sayan Mukherjee, and Mike West. Statistical Science, in press.
- Non-parametric Bayesian kernel models. Feng Liang, Kai Mao, Ming Liao, Sayan Mukherjee and Mike West. Biometrika, submitted.

Table of contents

Regression

data $= \{L_i = (x_i, y_i)\}_{i=1}^n$ with $L_i \stackrel{iid}{\sim} \rho(X, Y)$.

$X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \subset \mathbb{R}$ and $p \gg n$.

A natural idea

$$f(x) = \mathbb{E}_Y[Y|x].$$

An excellent estimator

$$\hat{f}(x) = \underset{f \in \text{bs}}{\arg\min} \, [\text{error on data} + \text{smoothness of function}]$$

An excellent estimator

$$\hat{f}(x) = \arg\min_{f \in \text{bs}} [\text{error on data} + \text{smoothness of function}]$$

$$\begin{aligned}
\text{error on data} &= L(f, \text{data}) = (f(x) - y)^2 \\
\text{smoothness of function} &= \|f\|_K^2 = \int |f'(x)|^2 dx \\
\underline{\text{b}}\text{ig function }\underline{\text{s}}\text{pace} &= \text{reproducing kernel Hilbert space} = \mathcal{H}_K
\end{aligned}$$

An excellent estimator

$$\hat{f}(x) = \arg\min_{f \in \mathcal{H}_K} \left[L(f, \text{data}) + \lambda \|f\|_K^2 \right]$$

The kernel: $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ e.g. $K(u, v) = e^{(-\|u-v\|^2)}$.

The RKHS

$$\mathcal{H}_K = \overline{\left\{ f \;\middle|\; f(x) = \sum_{i=1}^{\ell} \alpha_i K(x, x_i), \;\; x_i \in \mathcal{X}, \; \alpha_i \in \mathbb{R}, \; \ell \in \mathbb{N} \right\}}.$$

Representer theorem

$$\hat{f}(x) = \arg \min_{f \in \mathcal{H}_K} \left[L(f, \text{data}) + \lambda \|f\|_K^2 \right]$$

$$\hat{f}(x) = \sum_{i=1}^{n} a_i K(x, x_i).$$

Great when $p \gg n$.

Very popular and useful

1. Support vector machines

$$\hat{f}(x) = \arg\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} |1 - y_i \cdot f(x_i)|_+ + \lambda \|f\|_K^2 \right],$$

2. Regularized Kernel regression

$$\hat{f}(x) = \arg\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} |y_i - f(x_i)|^2 + \lambda \|f\|_K^2 \right],$$

3. Regularized logistic regression

$$\hat{f}(x) = \arg\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{n} \ln\left(1 + e^{-y_i \cdot f(x_i)}\right) + \lambda \|f\|_K^2 \right].$$

Bayesian interpretation of RBF

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \text{No}(0, \sigma^2).$$

$$\text{Lik}(\text{data}|f) \propto \prod_{i=1}^{n} \exp(-(y_i - f(x_i))^2/2\sigma^2) \quad \pi(f) \propto exp(-\|f\|_K^2).$$

$$\mathbf{P}\text{rob}(f|\text{data}) \propto \text{Lik}(\text{data}|f) \cdot \pi(f).$$

Bayesian interpretation of RBF

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$$\mathrm{Lik}(\mathrm{data}|f) \propto \prod_{i=1}^{n} \exp(-(y_i - f(x_i))^2/2\sigma^2) \quad \pi(f) \propto exp(-\|f\|_K^2).$$

$$\mathbf{P}\mathrm{rob}(f|\mathrm{data}) \propto \mathrm{Lik}(\mathrm{data}|f) \cdot \pi(f).$$

Maximum a posteriori (MAP) estimator

$$\hat{f} = \arg\max_{f \in \mathcal{H}_K} \mathbf{P}\mathrm{rob}(f|\mathrm{data}).$$

I want the full posterior.

Priors via spectral expansion

$$\mathcal{H}_K = \left\{ f \;\middle|\; f(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \text{ with } \sum_{i=1}^{\infty} a_i^2/\lambda_i < \infty \right\},$$

$\phi_i(x)$ and $\lambda_i \geq 0$ are eigenfunctions and eigenvalues of K:

$$\lambda_i \phi_i(x) = \int_{\mathcal{X}} K(x,u)\phi_i(u) d\gamma(u).$$

Priors via spectral expansion

$$\mathcal{H}_K = \left\{ f \;\middle|\; f(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \text{ with } \sum_{i=1}^{\infty} a_i^2/\lambda_i < \infty \right\},$$

$\phi_i(x)$ and $\lambda_i \geq 0$ are eigenfunctions and eigenvalues of K:

$$\lambda_i \phi_i(x) = \int_{\mathcal{X}} K(x,u)\phi_i(u) d\gamma(u).$$

Specify a prior on $\mathcal{H}_K$ via a prior on $\mathcal{A}$

$$\mathcal{A} = \left\{ (a_k)_{k=1}^{\infty} \;\middle|\; \sum_k a_k^2/\lambda_k < \infty \right\}.$$

Priors via spectral expansion

$$\mathcal{H}_K = \left\{ f \;\middle|\; f(x) = \sum_{i=1}^{\infty} a_i \phi_i(x) \text{ with } \sum_{i=1}^{\infty} a_i^2/\lambda_i < \infty \right\},$$

$\phi_i(x)$ and $\lambda_i \geq 0$ are eigenfunctions and eigenvalues of K:

$$\lambda_i \phi_i(x) = \int_{\mathcal{X}} K(x,u)\phi_i(u) d\gamma(u).$$

Specify a prior on $\mathcal{H}_K$ via a prior on $\mathcal{A}$

$$\mathcal{A} = \left\{ (a_k)_{k=1}^{\infty} \;\middle|\; \sum_k a_k^2/\lambda_k < \infty \right\}.$$

Hard to sample and relies on computation of eigenvalues and eigenvectors.

Priors via duality

The duality between Gaussian processes and RKHS implies the following construction

$$f(\cdot) \sim GP(\mu_f, K),$$

where K is given by the kernel.

Priors via duality

The duality between Gaussian processes and RKHS implies the following construction

$$f(\cdot) \sim GP(\mu_f, K),$$

where K is given by the kernel.

$f(\cdot) \notin \mathcal{H}_K$ almost surely.

Integral operators

Integral operator $\mathcal{L}_K : \Gamma \to \mathcal{G}$

$$\mathcal{G} = \left\{ f \;\middle|\; f(x) := \mathcal{L}_K[\gamma](x) = \int_{\mathcal{X}} K(x,u)\, d\gamma(u), \quad \gamma \in \Gamma \right\},$$

with $\Gamma \subseteq \mathcal{B}(\mathcal{X})$.

Integral operators

Integral operator $\mathcal{L}_K : \Gamma \rightarrow \mathcal{G}$

$$\mathcal{G} = \left\{ f \;\middle|\; f(x) := \mathcal{L}_K[\gamma](x) = \int_{\mathcal{X}} K(x,u)\, d\gamma(u), \quad \gamma \in \Gamma \right\},$$

with $\Gamma \subseteq \mathcal{B}(\mathcal{X})$.

A prior on Γ implies a prior on $\mathcal{G}$.

Equivalence with RKHS

For what Γ is $\mathcal{H}_K = \text{span}(\mathcal{G})$?

What is $\mathcal{L}_K^{-1}(\mathcal{H}_K) =??$. This is hard to characterize.

Equivalence with RKHS

For what Γ is $\mathcal{H}_K = \text{span}(\mathcal{G})$?

What is $\mathcal{L}_K^{-1}(\mathcal{H}_K) =$??. This is hard to characterize.

The candidates for Γ will be

1. square integrable functions
2. integrable functions
3. discrete measures
4. the union or integrable functions and discrete measures.

Square integrable functions are too small

Proposition

For every $\gamma \in L^2(\mathcal{X})$, $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$. *Consequently,* $L^2(\mathcal{X}) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Square integrable functions are too small

Proposition

For every $\gamma \in L^2(\mathcal{X})$, $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$. *Consequently,* $L^2(\mathcal{X}) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Corollary

If $\Lambda = \{k : \lambda_k > 0\}$ *is a finite set, then* $\mathcal{L}_K(L^2(\mathcal{X})) = \mathcal{H}_K$ *otherwise* $\mathcal{L}_K(L^2(\mathcal{X})) \subsetneq \mathcal{H}_K$. *The latter occurs when the kernel K is strictly positive definite, the RKHS is infinite-dimensional.*

Signed measures are (almost) just right

Measures: The class of functions $L^1(\mathcal{X})$ are signed measures.

Proposition

For every $\gamma \in L^1(\mathcal{X})$, $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$. *Consequently,* $L^1(\mathcal{X}) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Signed measures are (almost) just right

Measures: The class of functions $L^1(\mathcal{X})$ are signed measures.

Proposition

For every $\gamma \in L^1(\mathcal{X})$, $\mathcal{L}_K[\gamma] \in \mathcal{H}_K$. Consequently, $L^1(\mathcal{X}) \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Discrete measures:

$$\mathcal{M}_D = \left\{ \mu = \sum_{i=1}^{n} c_i \delta_{x_i} : \sum_{i=1}^{n} |c_i| < \infty, x_i \in \mathcal{X}, \ \ n \in \mathbb{N} \right\}.$$

Proposition

Given the set of finite discrete measures, $\mathcal{M}_D \subset \mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Signed measures are (almost) just right

Nonsingular measures: $\mathcal{M} = L^1(\mathcal{X}) \cup \mathcal{M}_D$

Proposition

$\mathcal{L}_K(\mathcal{M})$ is dense in $\mathcal{H}_K$ with respect to the RKHS norm.

Signed measures are (almost) just right

Nonsingular measures: $\mathcal{M} = L^1(\mathcal{X}) \cup \mathcal{M}_D$

Proposition

$\mathcal{L}_K(\mathcal{M})$ *is dense in* $\mathcal{H}_K$ *with respect to the RKHS norm.*

Proposition

$\mathcal{B}(\mathcal{X}) \subsetneqq \mathcal{L}_K^{-1}(\mathcal{H}_K(\mathcal{X}))$.

The implication

Take home message – need priors on signed measures.

A function theoretic foundation for random signed measures such as Gaussian, Dirichlet and Lévy process priors.

Bayesian kernel model

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where $Z(du) \in \mathcal{M}(\mathcal{X})$ is a signed measure on $\mathcal{X}$.

Bayesian kernel model

$$y_i = f(x_i) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathrm{No}(0, \sigma^2).$$

$$f(x) = \int_{\mathcal{X}} K(x, u) Z(du)$$

where $Z(du) \in \mathcal{M}(\mathcal{X})$ is a signed measure on $\mathcal{X}$.

$$\pi(Z|\text{data}) \propto L(\text{data}|Z)\ \pi(Z),$$

this implies a posterior on f.

Lévy processes

A stochastic process $Z := \{Z_u \in \mathbb{R} : u \in \mathcal{X}\}$ is called a Lévy process if it satisfies the following conditions:

1. $Z_0 = 0$ almost surely.
2. For any choice of $m \geq 1$ and $0 \leq u_0 < u_1 < ... < u_m$, the random variables $Z_{u_0}, Z_{u_1} - Z_{u_0}, ..., Z_{u_m} - Z_{u_{m-1}}$ are independent. (Independent increments property)
3. The distribution of $Z_{s+u} - Z_s$ is independent of Z_s (Temporal homogeneity or stationary increments property).
4. Z has càdlàg paths almost surely.

Lévy processes

Theorem (Lévy-Khintchine)

If Z is a Lévy process, then the characteristic function of $Z_u : u \geq 0$ has the following form:

$$\mathbb{E}[e^{i\lambda Z_u}] = \exp\left\{ u \left[i\lambda a - \frac{1}{2}\sigma^2\lambda^2 + \int_{\mathbb{R}\setminus\{0\}} [e^{i\lambda w} - 1 - i\lambda w 1_{\{w:|w|<1\}}(w)]\nu(dw) \right] \right\},$$

where $a \in \mathbb{R}$, $\sigma^2 \geq 0$ and ν is a nonnegative measure on $\mathbb{R}$ with $\int_{\mathbb{R}} (1 \wedge |w|^2)\nu(dw) < \infty$.

Lévy processes

- drift term a
- variance of Brownian motion σ^2
- $\nu(dw)$ the jump process or Lévy measure.

$$\exp\left\{u\left[i\lambda a - \tfrac{1}{2}\sigma^2\lambda^2\right]\right\}$$
$$\exp\left\{u\int_{\mathbb{R}\setminus\{0\}}\left[e^{i\lambda w} - 1 - i\lambda w 1_{\{w:|w|<1\}}(w)\right]\nu(dw)\right\}$$

Two approaches to Gaussian processes

Two modelling approaches

1. prior directly on the space of functions by sampling from paths of the Gaussian process defined by K;
2. Gaussian process prior on $Z(du)$ implies on prior on function space via integral operator.

Prior on random measure

A Gaussian process prior on $Z(du)$ is a signed measure so $\text{span}(\mathcal{G}) \subset \mathcal{H}_{\mathcal{K}}$.

Direct prior elicitation

Theorem (Kallianpur)

If $\{Z_u, u \in \mathcal{X}\}$ is a Gaussian process with covariance K and mean $m \in \mathcal{H}_K$ and $\mathcal{H}_K$ is infinite dimensional, then

$$\mathbf{P}(Z_\bullet \in \mathcal{H}_K) = 0.$$

The sample paths are almost surely outside $\mathcal{H}_K$.

A bigger RKHS

Theorem (Lukić and Beder)

Given two kernel functions R and K, R dominates K ($R \succ K$) if $\mathcal{H}_K \subseteq \mathcal{H}_R$. Let $R \succ K$. Then

$$\|g\|_R \leq \|g\|_K, \quad \forall g \in \mathcal{H}_K.$$

There exists a unique linear operator $L : \mathcal{H}_R \rightarrow \mathcal{H}_R$ whose range is contained in $\mathcal{H}_K$ such that

$$\langle f, g \rangle_R = \langle Lf, g \rangle_K, \quad \forall f \in \mathcal{H}_R, \forall g \in \mathcal{H}_K.$$

In particular

$$LR_u = K_u, \quad \forall u \in \mathcal{X}.$$

As an operator into $\mathcal{H}_R$, L is bounded, symmetric, and positive.
Conversely, let $L : \mathcal{H}_R \rightarrow \mathcal{H}_R$ be a positive, continuous, self-adjoint operator then

$$K(s,t) = \langle LR_s, R_t \rangle_R, \quad s,t \in \mathcal{X}$$

defines a reproducing kernel on $\mathcal{X}$ such that $K \leq R$.

A bigger RKHS

If L is nucular (an operator that is compact with finite trace independent of basis choice) then we have nucular dominance $R \succcurlyeq K$.

A bigger RKHS

If L is nucular (an operator that is compact with finite trace independent of basis choice) then we have nucular dominance $R \succcurlyeq K$.

Theorem (Lukić and Beder)

Let K and R be two reproducing kernels. Assume that the RKHS $\mathcal{H}_R$ is separable.
A necessary and sufficient condition for the existence of a Gaussian process with covariance K and mean $m \in \mathcal{H}_R$ and with trajectories in $\mathcal{H}_R$ with probability 1 is that $R \succcurlyeq K$.

A bigger RKHS

If L is nucular (an operator that is compact with finite trace independent of basis choice) then we have nucular dominance $R \succcurlyeq K$.

Theorem (Lukić and Beder)

Let K and R be two reproducing kernels. Assume that the RKHS $\mathcal{H}_R$ is separable.
A necessary and sufficient condition for the existence of a Gaussian process with covariance K and mean $m \in \mathcal{H}_R$ and with trajectories in $\mathcal{H}_R$ with probability 1 is that $R \succcurlyeq K$.

Characterize $\mathcal{H}_R$ by $\mathcal{L}_K^{-1}(\mathcal{H}_K)$.

Dirichlet distribution

Multinomial distribution

$$g(x_1, ..., x_k | n, p_1, ..., p_k) = \frac{n!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_k^{x_k}, \quad \sum_{i=1}^{k} x_i = n, x_i \geq 0.$$

Dirichlet distribution

Multinomial distribution

$$g(x_1, ..., x_k | n, p_1, ..., p_k) = \frac{n!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_k^{x_k}, \quad \sum_{i=1}^{k} x_i = n, x_i \geq 0.$$

Dirichlet distribution

$$f(p_1, \ldots, p_k | \alpha_1, \ldots, \alpha_k) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \quad \sum_{i=1}^{k} p_i = 1, p_i \geq 0,$$

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}.$$

Conjugacy

If $\mathbf{P}\mathrm{rob}(\theta|\mathrm{data})$ and $\pi(\theta)$ belong to the same family they are conjugate.

Conjugacy

If $\mathbf{P}\mathrm{rob}(\theta|\mathrm{data})$ and $\pi(\theta)$ belong to the same family they are conjugate.

Let $x = \{x_1, ..., x_k\}$ and $p = \{p_1, ..., p_k\}$

$$\begin{aligned} p &\sim \mathrm{Dir}(\alpha) \\ x|p &\sim \mathrm{Mult}(p) \\ p|x &\sim \mathrm{Dir}(p+\alpha). \end{aligned}$$

Dirichlet process prior

Given distribution function F and a specified distribution F_0 with the same support on a space $\mathcal{X}$.
Dirichlet process $\mathrm{DP}(\alpha, F_0)$ implies that for any partition of the space $B_1, ..., B_K$

$$F(B_1), ..., F(B_k) \sim \mathrm{Dir}(\alpha(F_0(B_1)), ..., \alpha(F_0(B_k))).$$

Dirichlet process prior

$$f(x) = \int_{\mathcal{X}} K(x,u) Z(du) = \int_{\mathcal{X}} K(x,u) w(u) F(du)$$

$F(du)$ is a distribution and $w(u)$ a coefficient function.

Dirichlet process prior

$$f(x) = \int_{\mathcal{X}} K(x,u) Z(du) = \int_{\mathcal{X}} K(x,u) w(u) F(du)$$

$F(du)$ is a distribution and $w(u)$ a coefficient function.

Model F using a Dirichlet process prior: $\mathrm{DP}(\alpha, F_0)$

Bayesian representer theorem

Given $X_n = (x_1, ..., x_n) \overset{iid}{\sim} F$

$$F \mid X_n \sim \mathrm{DP}(\alpha + n, F_n), \quad F_n = (\alpha F_0 + \sum_{i=1}^{n} \delta_{x_i})/(\alpha + n).$$

$$\mathbb{E}[f \mid X_n] = a_n \int K(x,u)\, w(u)\, dF_0(u) + n^{-1}(1 - a_n) \sum_{i=1}^{n} w(x_i)\, K(x, x_i),$$

$a_n = \alpha/(\alpha + n)$.

Bayesian representer theorem

Taking $\lim \alpha \to 0$ to represent a non-informative prior:

Theorem (Bayesian representor theorem)

$$\hat{f}_n(x) = \sum_{i=1}^{n} w_i \, K(x, x_i),$$

$w_i = w(x_i)/n.$

Likelihood

$$y_i = f(x_i) + \varepsilon_i = w_0 + \sum_{j=1}^{n} w_j K(x_i, x_j) + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\varepsilon_i \sim \mathrm{No}(0, \sigma^2)$.

$$Y \sim \mathrm{No}(w_0 \iota + K w, \sigma^2 I).$$

where $\iota = (1, \ldots, 1)'$.

Prior specification

Factor: $K = F \Delta F'$ with $\Delta := \text{diag}(\lambda_1^2, ..., \lambda_n^2)$ and $w = F \Delta^{-1} \beta$.

Prior specification

Factor: $K = F \Delta F'$ with $\Delta := \mathrm{diag}(\lambda_1^2, ..., \lambda_n^2)$ and $w = F\Delta^{-1}\beta$.

$$
\begin{aligned}
\pi(w_0, \sigma^2) &\propto 1/\sigma^2 \\
\tau_i^{-1} &\sim \mathrm{Ga}(a_\tau/2, b_\tau/2) \\
T &:= \mathrm{diag}(\tau_1, ..., \tau_n) \\
\beta &\sim \mathrm{No}(0, T) \\
w | K, T &\sim \mathrm{No}(0, F\Delta^{-1} T \Delta^{-1} F').
\end{aligned}
$$

Prior specification

Factor: $K = F \Delta F'$ with $\Delta := \text{diag}(\lambda_1^2, ..., \lambda_n^2)$ and $w = F\Delta^{-1}\beta$.

$$\begin{aligned} \pi(w_0, \sigma^2) &\propto 1/\sigma^2 \\ \tau_i^{-1} &\sim \text{Ga}(a_\tau/2, b_\tau/2) \\ T &:= \text{diag}(\tau_1, ..., \tau_n) \\ \beta &\sim \text{No}(0, T) \\ w|K, T &\sim \text{No}(0, F\Delta^{-1} T \Delta^{-1} F'). \end{aligned}$$

Standard Gibbs sampler simulates $p(w, w_0, \sigma^2|\text{data})$.

Sampling from posterior

Objective: sample from $p(w, w_0, \sigma^2 | \text{data})$.

Sampling from posterior

Objective: sample from $p(w, w_0, \sigma^2 | \text{data})$.

Given for $\{w^{(i)}, w_0^{(i)}, p_i(w, w_0)\}_{i=1}^{T}$ we have T functions can compute Bayes average and variance pointwise

$$
\begin{aligned}
\bar{f}(x) &= \sum_{i=1}^{T} p_i(w, w_0) \left[w_0^{(j)} + \sum_{j=1}^{n} K(x, x_j) w_j^{(i)} \right] \\
\operatorname{var}[f(x)] &= \sum_{i=1}^{T} p_i(w, w_0) \left[\bar{f}(x) - f_i(x) \right]^2 .
\end{aligned}
$$

Markov chain Monte Carlo

It may be difficult to sample from $p(w, w_0, \sigma^2|\text{data})$, for example high dimensions. Also the normalizing constant $\mathcal{Z}$ is unavailable

$$p(w, w_0, \sigma^2|\text{data}) = \frac{\text{Lik}(\text{data}|w, w_0, \sigma^2) \cdot \pi(w, w_0, \sigma^2)}{\mathcal{Z}}$$

$$\mathcal{Z} = \int \text{Lik}(\text{data}|w, w_0, \sigma^2) \cdot \pi(w, w_0, \sigma^2) dw_0 \, dw \, d\sigma.$$

Markov chain Monte Carlo

Say we want to sample $p(\theta|\text{data})$ but its hard.
Say we have a Markov chain (aperiodic, irreducible, detailed balance)

$$\begin{aligned} q(\theta^*|\theta) &= \mathbf{P}\text{rob}(\theta^*|\theta) \\ \mathbf{P}\text{rob}(\theta)q(\theta^*|\theta) &= \mathbf{P}\text{rob}(\theta^*)q(\theta|\theta^*). \end{aligned}$$

Markov chain Monte Carlo

Metropolis-Hastings

1. Given $\theta^{(t)}$ sample θ^* from $q(\theta^*|\theta^{(t)})$
2. Accept, $\theta^{(t+1)} = \theta^*$ with probability

$$\mathcal{A} = \min\left[1, \frac{p(\theta^*)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)})q(\theta^*|\theta^{(t)})}\right]$$

otherwise $\theta^{(t+1)} = \theta^{(t)}$.

Gibbs sampling

Given d-dimensional θ with known conditional

$$p(\theta_j | \theta_{-j}) = p(\theta_j | \theta_1, \ldots \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d).$$

Gibbs sampling

Given d-dimensional θ with known conditional

$$p(\theta_j|\theta_{-j}) = p(\theta_j|\theta_1, ...\theta_{j-1}, \theta_{j+1}, ..., \theta_d).$$

Proposal distribution

$$q(\theta^*|\theta^{(t)}) = p(\theta_j^*|\theta_{-j}^{(t)}).$$

Gibbs sampling

Acceptance probability

$$
\begin{aligned}
\mathcal{A} &= \min \left[1, \frac{p(\theta^*) q(\theta^{(t)} | \theta^*)}{p(\theta^{(t)}) q(\theta^* | \theta^{(t)})} \right] \\
&= \min \left[1, \frac{p(\theta^*) p(\theta_j^{(t)} | \theta_{-j}^{(t)})}{p(\theta^{(t)}) q(\theta_j^* | \theta_{-j}^*)} \right] \\
&= \min \left[1, \frac{p(\theta_{-j}^*)}{p(\theta_{-j}^{(t)})} \right].
\end{aligned}
$$

Gibbs sampling example

We want to sample from $x = 1, 2, 3, ..., n$ and $y \in [0, 1]$

$$p(x, y|n, \alpha, \beta) = \frac{n!}{(n-x)!x!} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}.$$

Gibbs sampling example

We want to sample from $x = 1, 2, 3, \ldots., n$ and $y \in [0, 1]$

$$p(x, y | n, \alpha, \beta) = \frac{n!}{(n-x)!x!} y^{x+\alpha-1} (1-y)^{n-x+\beta-1}.$$

Conditionals

$$\begin{aligned} x|y &\sim \mathrm{Bin}(n, y) = \frac{n!}{(n-x)!} y^x (1-y)^{(n-x)} \\ y|x &\sim \mathrm{Be}(x+\alpha, n-x+\beta) \propto y^{x+\alpha} (1-y)^{n-x+\beta} \end{aligned}$$

Gibbs sampling example

1. given y_t

Gibbs sampling example

1. given y_t
2. draw $x_{t+1} \sim \text{Bin}(n, y_t)$

Gibbs sampling example

1. given y_t
2. draw $x_{t+1} \sim \text{Bin}(n, y_t)$
3. draw $y_{t+1} \sim \text{Be}(x_{t+1} + \alpha, n - x_{t+1} + \beta)$

Gibbs sampling example

1. given y_t
2. draw $x_{t+1} \sim \text{Bin}(n, y_t)$
3. draw $y_{t+1} \sim \text{Be}(x_{t+1} + \alpha, n - x_{t+1} + \beta)$
4. return to (2).

Kernel model extension

$$K_\nu(x,u) = K(\sqrt{\nu} \otimes x, \sqrt{\nu} \otimes u)$$

with $\nu = \{\nu_1, ..., \nu_p\}$ with $\nu_k \in [0, \infty)$ as a scale parameter.

$$\begin{aligned}
k_\nu(x,u) &= \sum_{k=1}^{p} \nu_k \, x_k \, u_k, \\
k_\nu(x,u) &= \left(1 + \sum_{k=1}^{p} \nu_k \, x_k \, u_k\right)^d, \\
k_\nu(x,u) &= \exp\left(-\sum_{k=1}^{p} \nu_k (x_k - u_k)^2\right).
\end{aligned}$$

Prior specification

$$\begin{aligned} \nu_k &\sim (1-\gamma)\delta_0 + \gamma \ \mathrm{Ga}(a_\nu, a_\nu s), \quad (k = 1, \dots, p), \\ s &\sim \mathrm{Exp}(a_s), \quad \gamma \sim \mathrm{Be}(a_\gamma, b_\gamma) \end{aligned}$$

Prior specification

$$
\begin{aligned}
\nu_k &\sim (1-\gamma)\delta_0 + \gamma \ \mathrm{Ga}(a_\nu, a_\nu s), \quad (k = 1, \ldots, p), \\
s &\sim \mathrm{Exp}(a_s), \quad \gamma \sim \mathrm{Be}(a_\gamma, b_\gamma)
\end{aligned}
$$

Standard Gibbs sampler does not work: Metropolis-Hastings.

Problem setup

$$\text{Labelled data : } (Y^p, X^p) = \{(y_i^p, x_i^p);\ i = 1 : n_p\} \overset{iid}{\sim} \rho(Y, X|\phi, \theta).$$

$$\text{Unlabelled data: } X^m = \{x_i^m,\ i = (1) : (n_m)\} \overset{iid}{\sim} \rho(X|\theta).$$

Problem setup

Labelled data : $(Y^p, X^p) = \{(y_i^p, x_i^p);\ i = 1 : n_p\} \stackrel{iid}{\sim} \rho(Y, X|\phi, \theta)$.

Unlabelled data: $X^m = \{x_i^m,\ i = (1) : (n_m)\} \stackrel{iid}{\sim} \rho(X|\theta)$.

How can the unlabelled data help our a predictive model ?

data $= \{Y, X, X^m\}$

$$p(\phi, \theta|\text{data}) \propto \pi(\phi, \theta) p(Y|X, \phi) p(X|\theta) p(X^m|\theta).$$

Need very strong dependence between θ and ϕ.

Bayesian kernel model

Result of DP prior

$$\hat{f}_n(x) = \sum_{i=1}^{n_p+n_m} w_i \, K(x, x_i).$$

Bayesian kernel model

Result of DP prior

$$\hat{f}_n(x) = \sum_{i=1}^{n_p+n_m} w_i \, K(x, x_i).$$

Same as in Belkin and Niyogi but without

$$\min_{f \in \mathcal{H}_K} \left[L(f, \text{data}) + \lambda_1 \|f\|_K^2 + \lambda_2 \|f\|_I^2 \right].$$

Bayesian kernel model

Result of DP prior

$$\hat{f}_n(x) = \sum_{i=1}^{n_p+n_m} w_i \, K(x, x_i).$$

1. $\theta = F(\cdot)$ so that $p(x|\theta)dx = dF(x)$ - the parameter is the full distribution function itself;
2. $p(y|x, \phi)$ depends intimately on $\theta = F$; in fact, $\theta \subseteq \phi$ in this case and dependence of θ and ϕ is central to the model.

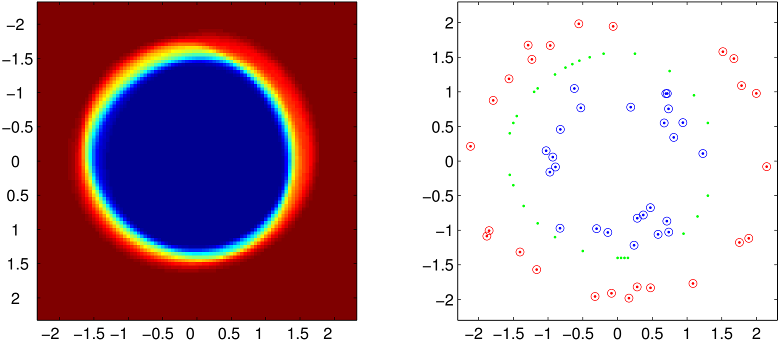
Bayesian kernel model

Result of DP prior

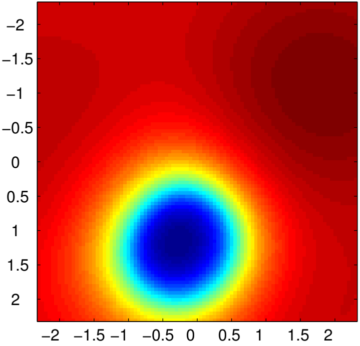
$$\hat{f}_n(x) = \sum_{i=1}^{n_p+n_m} w_i \, K(x, x_i).$$

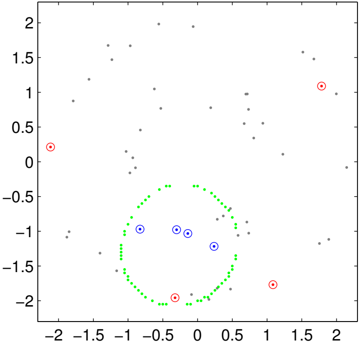
1. $\theta = F(\cdot)$ so that $p(x|\theta)dx = dF(x)$ - the parameter is the full distribution function itself;
2. $p(y|x, \phi)$ depends intimately on $\theta = F$; in fact, $\theta \subseteq \phi$ in this case and dependence of θ and ϕ is central to the model.

Simulated data – semi-supervised

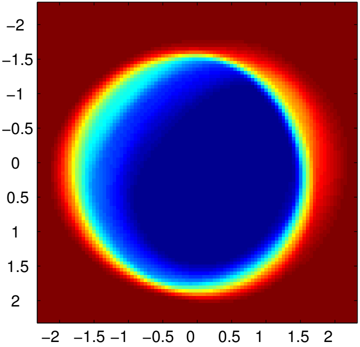


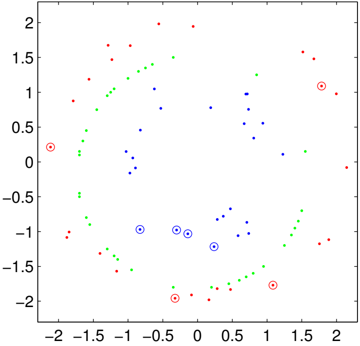
Simulated data – semi-supervised



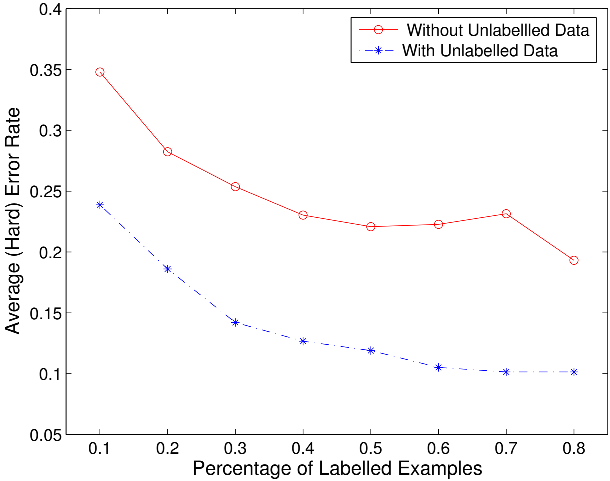


Simulated data – semi-supervised

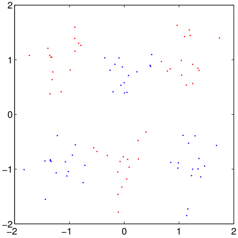


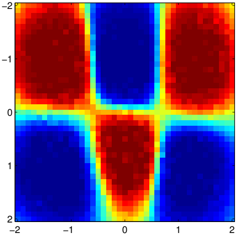


Cancer classification – semi-supervised

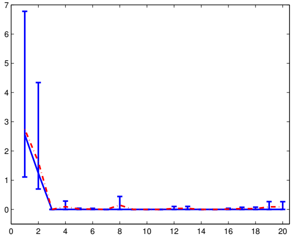


Simulated data – feature selection

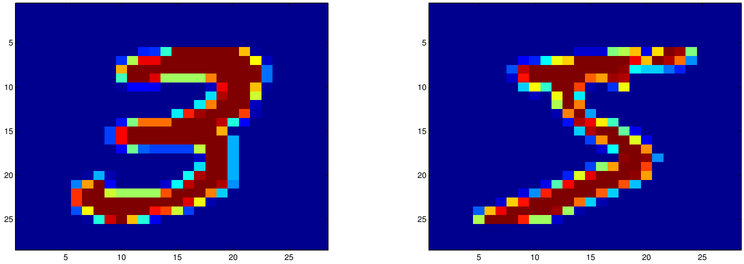




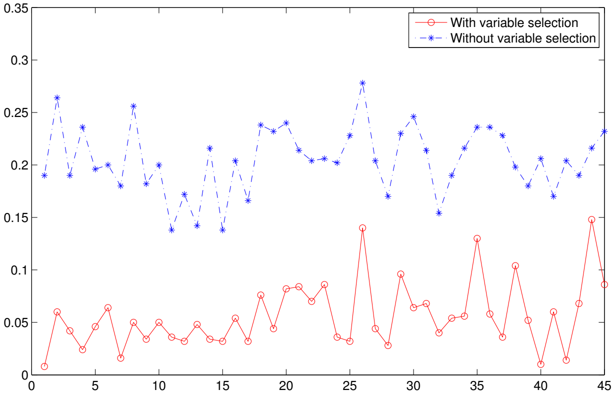
Simulated data – feature selection



MNIST digits – feature selection



MNIST digits – feature selection



Discussion

Lots of work left:

- Further refinement of integral operators and priors in terms of Sobolev spaces.

Discussion

Lots of work left:

- Further refinement of integral operators and priors in terms of Sobolev spaces.
- Semi-supervised setting: relation of kernel model and priors with Laplace-Beltrami and graph Laplacian operators.

Discussion

Lots of work left:

- Further refinement of integral operators and priors in terms of Sobolev spaces.
- Semi-supervised setting: relation of kernel model and priors with Laplace-Beltrami and graph Laplacian operators.
- Semi-supervised setting: Duality between diffusion processes on manifolds and Markov chains.

Discussion

Lots of work left:

- Further refinement of integral operators and priors in terms of Sobolev spaces.
- Semi-supervised setting: relation of kernel model and priors with Laplace-Beltrami and graph Laplacian operators.
- Semi-supervised setting: Duality between diffusion processes on manifolds and Markov chains.
- Bayesian variable selection: Efficient sampling and search in high-dimensional space.

Discussion

Lots of work left:

- Further refinement of integral operators and priors in terms of Sobolev spaces.
- Semi-supervised setting: relation of kernel model and priors with Laplace-Beltrami and graph Laplacian operators.
- Semi-supervised setting: Duality between diffusion processes on manifolds and Markov chains.
- Bayesian variable selection: Efficient sampling and search in high-dimensional space.
- Numeric stability and statistical robustness.

Summary

Its extra work but it pays to be Bayes :)