Bayesian Kernel Regression with Feature Selection

Zhi Ouyang, Merlise A. Clyde, Robert L. Wolpert

Institution of Statistics and Decision Sciences
Duke University

Salt Lake City, Joint Statistical Meetings
July 30, 2007
Motivating Problems

- Feature Selection
  - e.g. discovering bio-markers
- Supervised Learning
  - e.g. making predictions for a new observation
Response: $y_1, \ldots, y_n \in \{0, 1\}$.
Predictors: $x_1, \ldots, x_n \in \mathbb{R}^d$. 
Support Vector Machine

\[
f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, \mathbf{c}_j, \Lambda_j)
\]

- \( J = n \).
- \( \mathbf{c}_j = x_j \).
- \( \Lambda_j = \Lambda \).

\[
\min_f \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{Cost}(y_i, f(x_i)) + \lambda \| f \|_{\mathcal{H}} \right\}
\]
**Relevance Vector Machine**

\[ f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, c_j, \Lambda_j) \]

- \( J = n \).
- \( c_j = x_j \).
- \( \Lambda_j = \Lambda \).
- \( y_i | x_i \sim \text{N}(f(x), \sigma^2), \quad \beta_j \sim \text{Cauchy}, \quad \sigma^{-2} \sim \text{Ga}(0, 0) \).
Bayesian Kernel Regression

\[ f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, c_j, \Lambda_j) \]

- \( J \)? Poisson?
- \( \beta_j \)? Normal/T? Double Exponential? Point mass mixture?
- \( c_j \)?
- \( \Lambda_j \)?
Lévy Adaptive Regression Kernel (LARK)

- \( f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, c_j, \Lambda_j) = \int_{\Omega} K(x, \omega) \mathcal{L}(d\omega). \)
- \( \mathcal{L} \) is a Lévy measure, and we can construct \( \mathcal{L}(\beta) \) from a Poisson process. Suppose \( \int_{\mathbb{R}} (1 \wedge |\beta|) \nu(d\beta) < \infty \), then
  \[
  \mathcal{L}(B) = \int_{B} \beta N(d\beta) = \int_{j=0}^{J} \beta_j 1_{\{\beta_j \in B\}} 1_{\beta_j}(d\beta_j).
  \]
- In particular, we chose a symmetric stable Lévy measure, \( \nu_{\alpha, \gamma}(d\beta) = \frac{\alpha \gamma}{\pi} \Gamma(\alpha) \sin \left( \frac{\pi \alpha}{2} \right) |\beta|^{-\alpha-1} d\beta \), and approximate it with \( \nu_{\alpha, \gamma, \epsilon}(d\beta) = \frac{\alpha \gamma}{\pi} \Gamma(\alpha) \sin \left( \frac{\pi \alpha}{2} \right) |\beta|^{-\alpha-1} 1_{|\beta| > \epsilon} d\beta \).
Suppose we have proper prior for $c$ and $\Lambda$, then

- $J \sim \text{Po} \left( \frac{2 \gamma}{\pi e^\alpha} \Gamma(\alpha) \sin \frac{\pi \alpha}{2} \right)$.
- $\beta_j \overset{\text{i.i.d.}}{\alpha} |\beta_j|^{-\alpha-1} 1(|\beta_j|>\varepsilon)$.

If we put a Gamma hyper prior on $\gamma$, then we obtain a Negative Binomial construction on $J$.

- $J|\mu \sim \text{Po}(\mu)$, $\mu \sim \text{Ga}(\alpha_J, \alpha_J/\lambda_J)$.
- $J \sim \text{NB}(\lambda_J, \alpha_J)$ with mean $\lambda_J$ and variance $\lambda_J + \lambda_J^2/\alpha_J$. 
Negative Binomial Construction - Illustration

Model Dimensions

Poisson Prior

Model Dimensions

Negative Binomial Prior

Ouyang, Clyde, Wolpert

JSM talk
Bayesian Kernel Regression - Locations

\[ f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, c_j, \Lambda_j) \]

- \(J\) is Negative Binomial constructed through a truncated \(\alpha\)-stable Lévy process prior on \(\beta\).
- \(c_j\)?
  - Random in \(\mathbb{R}^d\)? Dirichlet Process?
  - Only chosen from the data?
- \(\Lambda_j\)?
Bayesian Kernel Regression - Kernel Shapes

\[ f(x) = \beta_0 + \sum_{j=1}^{J} \beta_j K(x, c_j, \Lambda_j) \]

- \( J \sim \text{NB} \left( \frac{2\gamma}{\pi\epsilon^\alpha} \Gamma(\alpha) \sin \frac{\pi\alpha}{2}, \text{size} \right), \beta_j \text{ i.i.d. } \propto |\beta_j|^{\alpha-1} 1(|\beta_j|>\epsilon) \).
- \( c_j \sim \text{Unif}\{x_1, \ldots, x_n\} \).
- \( \Lambda_j ? \)
  - All \( \Lambda_j \)'s are equal, but random; or not equal?
  - Gaussian Kernel, reduce the number of parameters?
Kernel Shapes

\[
K(x, c, \Lambda) = \prod_{l=1}^{r} \frac{\lambda_{l}^{1/2}}{(2\pi)^{r/2}} \exp \left\{ -\frac{1}{2} (x - c)^T \Lambda_{d \times r} \left( \text{diag} \lambda_{r \times 1} \right) U_{r \times d}^T (x - c) \right\},
\]

\[
\Lambda_{d \times d} = U_{d \times r} \left( \text{diag} \lambda_{r \times 1} \right) U_{r \times d}^T, \quad U^T U = I_r.
\]

- \( U \sim \text{Unif}\{\text{Stiefel}(d, r)\} \).
- \( \lambda_{j,l} \overset{\text{i.i.d.}}{\sim} \text{LN}(\mu_{\lambda}, \phi_{\lambda}^{-1}) \), \( \mu_{\lambda} \sim \text{Normal}, \phi_{\lambda} \sim \text{Gamma} \).
von Mises-Fisher Proposal

- \( \text{St}(d, r) = \{ U_{d \times r} | U^T U = I_r \} \).
- \( U \sim \text{VMF}(A) \), with density 
  \( p(U) \propto \exp\{ tr(A^T U) \} \).
- \( U = [U_1, U_2], A = [A_1, A_2] \), then 
  \( U_1 | U_2 \sim \text{VMF} \).
- Special case: \( r = 1 \).
We implement this with reversible jump MCMC.

- Birth step, propose new parameters from the prior.
- Death step, kill one kernel at random (uniformly), and ... 
- Update step
  - Regular updates 
  - When proposed $\beta_j$ is out of the Pareto domain, we propose to kill the corresponding kernel.

- Hyper parameters can be updated conjugately.
Simulation studies

- 2 dimensional test signals - checker board, rotated blocks and concentric circles.
- 2 dimensional signals plus 10 dimensions of noises.
- 2 dimensional signals hidden in 5 dimensional data.
  \[ s_1 = x_1 + x_2 + x_3 + x_4 + x_5, \quad s_2 = x_1 - x_2 + x_3 - x_4 + x_5. \]

Real data analysis

- Ionosphere data, 32 dimensions, 351 observations.
Summary:
- Fully Bayesian approach of kernel regression models.
- Use Lévy random measures to control model complexity.
- Make inference for $U$ on the Stiefel manifold.

Future Work:
- Improve MCMC convergence - gradient method based on geodesics, adaptive MCMC ...
- Move to large $p$ small $n$ problems.