Nonparametric Bayesian Methods

Debdeep Pati
Florida State University

October 2, 2014

Large spatial datasets (Problem of big n)

- Large observational and computer-generated datasets:
- Often have spatial and temporal aspects.
- Goal: Make inference on underlying spatial processes from observations at n locations where n is large.

Computational bottleneck

- The posterior predictive involves $(C(X, X) + \sigma^2 I)^{-1}$
- The covariance matrix $C(X, X)$ is large: $n \times n$ for n locations. unstructured: irregular spaced locations. dense: non-negligible correlations.
- Cholesky decomposition of $n \times n$ matrices Generally requires $O(n^3)$ computations and $O(n^2)$ memory.

Options for large n

- Use models that reduce computations and/or storage. Use approximate methods.
- Compactly supported covariance functions.
- Reduced rank covariance functions.
- Leads Statistical and computational efficiency.

Covariance Tapering (Furrer et al 2006)

- Covariance tapering: $\tilde{C}(x, x') = C(x, x') \circ T(x, x'; \gamma)$,
- $T(x, x'; \gamma)$: an isotropic correlation function of compact support, i.e., $T(x, x'; \gamma) = 0$ for $|x - x'| \geq \gamma$.
- Assumptions: The covariance function has compact support. Its range is sufficiently small.
- The tapered covariance matrix $\tilde{C}$ retains the property of positive definiteness, zero at large distances.
- Minimal distortion to C for nearby locations.
- Efficient sparse matrix algorithms can be used. Also saves storage.

Reduced Rank approximations

- Find reduced rank covariance function representation, Banerjee et al. (2008), JRSSB: proposed Gaussian predictive processes $\tilde{f}(x)$ to replace $f(x)$ by projecting $f(x)$ onto a m-dimension (lower) subspace $\tilde{f}(x) = E(f(x) \mid f(x_1^*), \ldots, f(x_m^*))$.
- Cressie and Johannesson (2008), JRSSB proposed a reduced rank approach by defining a low rank process $\tilde{f}(x) = B^T(x)\eta_{m\times 1}$, where B is a vector consisting of m basis functions and $var(\eta) = G$.
- Have computational advantages but also limitations. (Stein, 2013, Spatial Statistics).
- Low rank+tapering: Sang and Huang (2011), JRSSB

Why Projections help

- For both predictive process and the basis function truncation approach, $\tilde{C}(X, X)$ is of the form $\tilde{C}(X, X) = B'GB$ where B is an $m \times n$ matrix, $m \ll n$.
- Need to invert $\sigma^2 I + \tilde{C}(X, X) = \sigma^2 I + B'GB$
- Use Woodbury Inversion formula

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1} + VA^{-1}U\right)^{-1}VA^{-1}$$

- Requires inverting $m \times m$ matrices !!!

RKHS of Gaussian processes

- In np Bayes, want priors to place positive probability around arbitrary neighborhoods of a large class of parameter values (large support property)
- The prior concentration plays a key role in determining the rate of posterior contraction
- The reproducing kernel Hilbert space (RKHS) of a Gaussian process determines the prior support and concentration
- Intuitively, a space of functions that are similar to the covariance kernel in terms of smoothness

Applications of Gaussian processes- Classification

- Let $\{(X_i, Y_i), i = 1, \ldots, n\}$, be i.i.d random pairs of observations, where $X_i \in [0,1]^d$ and $Y_i \in \{0,1\}$. Let $\mathcal{Z} = [0,1]^d \times \{0,1\}$.
- Denote by P_X, the probability distribution of X_i and by $P_{X,Y}$ the joint distribution of (X_i, Y_i) and $P^{\otimes n}$ the joint distribution of $\{(X_i, Y_i), i = 1, \ldots, n\}$ and $E^{\otimes n}$ denotes the expectation w.r.t $P^{\otimes n}$.
- The goal of a classification is to predict the label Y given the value of X, i.e. to provide a decision rule $f : [0,1]^d \to \{0,1\}$. The class of decision rules is denoted by $\mathcal{F}$.
- The performance of a decision rule f is measured by the misclassification error

$$R(f) := P(Y \neq f(X))$$

and corresponding empirical version

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{Y_i \neq f(X_i)}.$$

Applications of Gaussian processes- Classification

- Of particular interest is the optimal decision rule $f^*(X) = 1_{\{\eta(X) \geq 1/2\}}$ where $\eta(x) = P(Y = 1 \mid X = x)$.
- The parameter η is usually estimated from the data necessitating the next definition. An empirical classifier is a random mapping $\hat{f}_n : \mathcal{Z}^n \to \mathcal{F}$. Its accuracy can be characterized by excess risk

$$\mathcal{E}(\hat{f}_n) = E^{\otimes n}\{R(\hat{f}_n) - R(f^*)\}.$$

Theorem

The decision rule f^* is a minimizer of the risk $R(f)$ over all decision rules $f \in \mathcal{F}$.

Applications of Gaussian processes- Classification

Lemma

For any empirical decision rule $\hat{f}_n$,
$\mathcal{E}(\hat{f}_n) = E^{\otimes n} \int_{[0,1]^d} |2\eta(x) - 1| I_{\{\hat{f}_n(x) \neq f^*(x)\}} P_X(x) dx.$

- We define two metrics on $\mathcal{F}$,
- $d(f, f^*) = \int_{[0,1]^d} I_{\{f(x) \neq f^*(x)\}} P_X(x) dx$
- $d_\eta(f, f^*) = \int_{[0,1]^d} |2\eta(x) - 1| I_{\{f(x) \neq f^*(x)\}} P_X(x) dx$. $d_\eta(f, f^*)$ is actually a pseudo-metric as it satisfies all the axioms except that $d(f_1, f_2) = 0 \implies f_1 = f_2$.

Modeling $\eta(x)$

- We will consider $\mathcal{Z} = [0,1]^d \times \{0,1\}$. For $\eta : [0,1]^d \to [0,1]$, consider

$$y_i \mid x_i \sim \text{Ber}\{\eta(x_i)\},$$

- Assume $\eta(x) = \Phi(f(x))$ and $f \sim GP(0, c)$. Consider three different classifiers based on the posterior distribution of η.
  1. **Plug-in classifiers:** $\hat{f}(x) = 1_{\hat{\eta}(x)>1/2}$, where $\hat{\eta}(x)$ is posterior mean / median.
  2. **Hybrid Plug-in Empirical Risk Minimizer (ERM) classifiers:** $\hat{\eta}_{ERM}$ is the maximizer of the posterior density $R_n(f) \mid Y^n, X^n$, $\hat{f}(x) = 1_{\hat{\eta}_{ERM}(x)>1/2}$
  3. **Bayes estimate** with respect to loss $d_\eta(f, f^*)$: $\hat{f}(x) = 1_{\{\Pi(\eta(x)>1/2|Y^n,X^n)>1/2\}}$.