UNIVERSITY OF CAMBRIDGE

Estimation of the Kernel Mean Embedding (with uncertainty)

Paul Rubenstein

University of Cambridge

Max-Planck Institute for Intelligent Systems, Tübingen

20th January 2016

RKHS theory

A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a kernel if given $x_1, \ldots, x_n \in \mathcal{X}$, $K_{ij} = k(x_i, x_k)$

K is symmetric and positive semi-definite (= is a valid covariance matrix)

Associated to k are:

- A Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \longrightarrow \mathbb{R}$
- A 'feature map' $\phi : \mathcal{X} \longrightarrow \mathbb{R}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$

RKHS theory

Suppose we are given:

- A random variable $X \sim \mathbb{P}$ taking value in $\mathcal{X}$
- A function $f : \mathcal{X} \longrightarrow \mathbb{R}$

and that we want to evaluate $\int f(x) d\mathbb{P}(x) = \mathbb{E}_X f(X)$. If $f \in \mathcal{H}$, then

$$\begin{aligned} \mathbb{E}_X f(X) &= \mathbb{E}_X \langle f, \phi(X) \rangle \\ &= \langle f, \mathbb{E}_X \phi(X) \rangle \end{aligned}$$

So if we know the *mean embedding of* X, $\mu_X := \mathbb{E}_X \phi(X)$, then we can calculate expectations of any function in $\mathcal{H}$ by taking an product.

RKHS theory

For certain k, the mapping $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective, ie

$$\mathbb{P} = \mathbb{Q} \iff \mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$$

We can exploit this to construct statistical tests of properties of distributions.

Two sample test: Given $\{X_i\} \sim \mathbb{P}$, $\{Y_i\} \sim \mathbb{Q}$, does $\mathbb{P} = \mathbb{Q}$?

Idea: estimate $\mu_{\mathbb{P}}$, $\mu_{\mathbb{Q}}$ and see how different they are

Independence testing: Given $\{(X_i, Y_i)\} \sim \mathbb{P}_{XY}$ does $\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_Y$?

Idea: estimate $\mathbb{P}_{XY}$, $\mathbb{P}_X \mathbb{P}_Y$ and see how different they are

Estimating μ_X

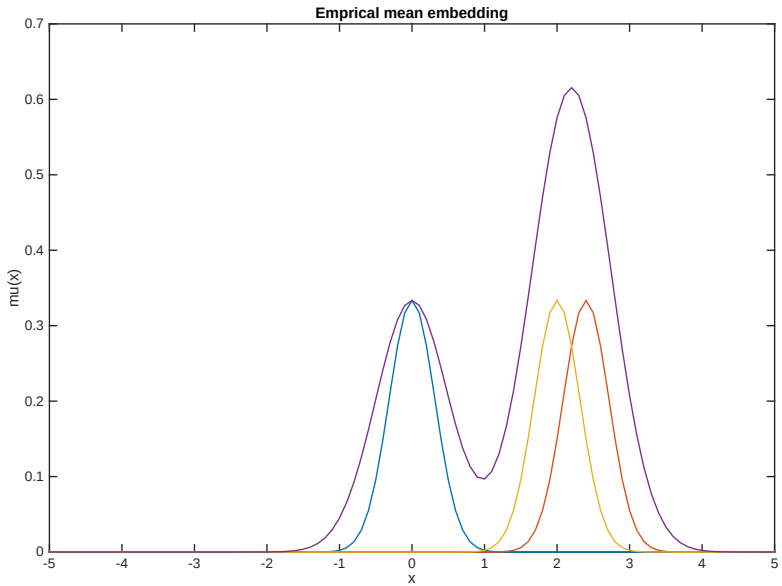
How do we estimate μ_X? $\mu_X = \mathbb{E}\phi(X) = \int k(x, \cdot) d\mathbb{P}(x)$

If $\{X_i\}_{i=1}^n \sim \mathbb{P}$, then

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} k(X_i, \cdot) \longrightarrow \mu_X$$

$$\mathbf{1}_n = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}, \Phi = \begin{pmatrix} k(X_1, \cdot) \\ \vdots \\ k(X_n, \cdot) \end{pmatrix}$$

$$\hat{\mu} = \mathbf{1}_n^\intercal \Phi$$



Estimating μ_X

In Muandet et al 2015(?) (Kernel Mean Shrinkage Estimators), the *risk* of an estimator $\hat{\mu}$ is defined:

$$\Delta = \mathbb{E}\|\hat{\mu} - \mu\|^2_{\mathcal{H}}$$

and estimators that minimise Δ are sought. Two proposals:

For particular α that can be estimated from observations,

$$\begin{aligned}\hat{\mu}_\alpha &= (1-\alpha)\hat{\mu} \\ &= (1-\alpha)\mathbf{1}_n^\intercal \Phi\end{aligned}$$

For λ estimated (by cross validation) from observations,

$$\hat{\mu}_\lambda = \Phi^\intercal (K + \lambda I)^{-1} K \mathbf{1}_n$$

(this looks like GP regression)

Bayesian estimation of μ_X

$$\hat{\mu}_\alpha = (1-\alpha)\mathbf{1}_n^\intercal \Phi \qquad\qquad \hat{\mu}_\lambda = \Phi^\intercal (K + \lambda I)^{-1} K \mathbf{1}_n$$

Kernel Ridge Regression $\iff$ MAP inference in GP regression.

Can we show that these estimators are the MAP solution to a Bayesian inference problem?

$$\mu \sim \mathcal{GP}(0, k) \qquad\qquad \mu \sim \mathcal{GP}(0, k)$$
$$\hat{\mu}|\mu \sim \mathcal{GP}(\mu, \gamma k) \qquad\qquad \hat{\mu}|\mu \sim \mathcal{GP}(\mu, \lambda \mathbb{I}_{x=x'})$$

Define 'pseudo-targets' $\hat{\mu}(\mathbf{x}) = K\mathbf{1}_n$ and then perform Bayesian inference

Deriving $\mu = (1-\alpha)\Phi\mathbf{1}_n$

Consider $\mu \sim \mathcal{GP}(0, k)$, $\hat{\mu}|\mu \sim \mathcal{GP}(\mu, \gamma k)$

Choose a previously unobserved z and consider distribution of $\begin{pmatrix} \mu(z) \\ \hat{\mu}(\mathbf{x}) \end{pmatrix}$

$$\begin{pmatrix} \mu(z) \\ \hat{\mu}(\mathbf{x}) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} k_{zz} & k_z^\intercal \\ k_z & (1+\gamma)K \end{pmatrix}\right)$$

$$\implies \mu(z)|\hat{\mu}(\mathbf{x}) \sim \mathcal{N}\left(\frac{1}{1+\gamma}k_z^\intercal \mathbf{1}_n, k_{zz} - \frac{1}{1+\gamma}k_z^\intercal K^{-1} k_z\right)$$

So if $\frac{1}{1+\gamma} = (1-\alpha) \iff \gamma = \frac{\alpha}{1-\alpha}$ then MAP solution is

$$\mu = (1-\alpha)\Phi\mathbf{1}_n$$

Deriving $\hat{\mu}_\lambda = \Phi^\intercal (K + \lambda I)^{-1} K \mathbf{1}_n$

Considering next $\mu \sim \mathcal{GP}(0, k)$, $\hat{\mu}|\mu \sim \mathcal{GP}(\mu, \lambda \mathbb{I}_{x=x'})$

$$\begin{pmatrix} \mu(z) \\ \hat{\mu}(\mathbf{x}) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} k_{zz} & k_z^\intercal \\ k_z & K + \lambda I \end{pmatrix}\right)$$

$$\implies \mu(z)|\hat{\mu}(\mathbf{x}) \sim \mathcal{N}\left(k_z^\intercal (K + \lambda I)^{-1} K \mathbf{1}_n, k_{zz} - k_z^\intercal (K + \lambda I)^{-1} k_z\right)$$

Thus the MAP solution is

$$\mu = \Phi^\intercal (K + \lambda I)^{-1} K \mathbf{1}_n$$

Some problems

Although we derive the same solution, most of the approach taken in the above doesn't really make sense:

- The prior over μ is not sensible
- The likelihood $\hat{\mu}$ is wrong - in fact, for large n, $\hat{\mu} \approx \mathcal{GP}(\mu, \frac{1}{n}[C_{XX} - \mu_X \otimes \mu_X)]$
- Uncertainty does not decay far away from observations as n grows.

Thanks!

Discussion?