Kernel Mean Estimation

Krikamol Muandet

Empirical Inference Department
Max Planck Institute for Intelligent Systems

Joint work with
Kenji Fukumizu (ISM)
Bharath Sriperumburdur (Cambridge)
Arthur Gretton (Gatsby, UCL)
Bernhard Schölkopf (MPI-IS)

September 26, 2013

Kernel Mean Estimator

A kernel mean of a distribution P over a measurable space $\mathscr{X}$:

$$\text{true mean} \qquad \mu_P := \int k(x,\cdot)\,dP(x) \in \mathscr{H}$$

Kernel Mean Estimator

A kernel mean of a distribution P over a measurable space $\mathscr{X}$:

$$\begin{aligned} \text{true mean} \qquad \mu_P &:= \int k(x,\cdot)\,dP(x) \in \mathscr{H} \\ \text{standard est.} \qquad \widehat{\mu}_P &:= \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot), \quad x_i \sim P \end{aligned}$$

Kernel Mean Estimator

A kernel mean of a distribution P over a measurable space $\mathscr{X}$:

true mean	μ_P	$:=$	$\int k(x,\cdot)\,dP(x) \in \mathscr{H}$
standard est.	$\widehat{\mu}_P$	$:=$	$\frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot), \quad x_i \sim P$
proposed est.	$\widehat{\mu}_P^{\lambda}$	$:=$	$\sum_{i=1}^{n} \beta_i k(x_i,\cdot), \quad \boldsymbol{\beta} := (K+\lambda I)^{-1} K \mathbf{1}_n$

Kernel Mean Estimator

A kernel mean of a distribution P over a measurable space $\mathscr{X}$:

true mean	μ_P	$:=$	$\int k(x,\cdot)\,dP(x) \in \mathscr{H}$
standard est.	$\widehat{\mu}_P$	$:=$	$\frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot), \quad x_i \sim P$
proposed est.	$\widehat{\mu}_P^{\lambda}$	$:=$	$\sum_{i=1}^{n} \beta_i k(x_i,\cdot), \quad \boldsymbol{\beta} := (K+\lambda I)^{-1} K \mathbf{1}_n$

Conjecture

Under the risk functional

$$R(\mu,\widehat{\mu}) = \mathbb{E}_X[\|\mu - \widehat{\mu}\|_{\mathscr{H}}^2],$$

there exist $\lambda > 0$ for which

$$R(\mu,\widehat{\mu}_P^{\lambda}) \leq R(\mu,\widehat{\mu}_P).$$

Step I

We have $\widehat{\mu}^{\lambda} = \sum_{i=1}^{n} \beta_i \phi(x_i) = \Phi^{\top}(K + \lambda I)^{-1} K \mathbf{1}_n = \Phi^{\top} \underbrace{(I + \lambda K^{-1})^{-1}}_{G_{\lambda}} \mathbf{1}_n$.

Rewriting G_{λ}:

$$\begin{aligned} G_{\lambda} = (I + \lambda K^{-1})^{-1} &= (I + \lambda K^{-1})^{-1}(I + \lambda K^{-1} - \lambda K^{-1}) \\ &= I - (I + \lambda K^{-1})^{-1}(\lambda K^{-1}) \\ &= I - \lambda (K + \lambda I)^{-1} \end{aligned}$$

Hence, we have

$$\begin{aligned} \widehat{\mu}^{\lambda} &= \Phi^{\top}(I - \lambda(K + \lambda I)^{-1})\mathbf{1}_n \\ &= \Phi^{\top}\mathbf{1}_n - \lambda \Phi^{\top}(K + \lambda I)^{-1}\mathbf{1}_n \\ &= \widehat{\mu} - \lambda \Phi^{\top}(K + \lambda I)^{-1}\mathbf{1}_n \\ &= \widehat{\mu} - \lambda A_{\lambda} \end{aligned}$$

where $A_{\lambda} = \Phi^{\top}(K + \lambda I)^{-1}\mathbf{1}_n$.

Step II

Substitute back $\widehat{\mu}^\lambda = \widehat{\mu} - \lambda A_\lambda$.

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mu}^\lambda - \mu\|^2] &= \mathbb{E}[\|\widehat{\mu} - \lambda A - \mu\|^2] \\
&= \mathbb{E}[\|(\widehat{\mu} - \mu) - \lambda A_\lambda\|^2] \\
&= \mathbb{E}[\|\widehat{\mu} - \mu\|^2] - 2\lambda\mathbb{E}[\langle\widehat{\mu} - \mu, A_\lambda\rangle] + \lambda^2\mathbb{E}[\|A_\lambda\|^2]
\end{aligned}$$

Thus, we have

$$\underbrace{\mathbb{E}[\|\widehat{\mu}^\lambda - \mu\|^2] - \mathbb{E}[\|\widehat{\mu} - \mu\|^2]}_{R(\mu,\widehat{\mu}_P^\lambda) - R(\mu,\widehat{\mu}_P)} = -2\lambda\mathbb{E}[\langle\widehat{\mu} - \mu, A_\lambda\rangle] + \lambda^2\mathbb{E}[\|A_\lambda\|^2] =: F(\lambda)$$

We need to show that the r.h.s is negative for some λ's.

Step III

Recall that

$$\begin{aligned} F(\lambda) &= -2\lambda\mathbb{E}[\langle\widehat{\mu}-\mu, A_\lambda\rangle] + \lambda^2\mathbb{E}[\|A_\lambda\|^2] \\ F(0) &= 0 \end{aligned}$$

Thus, it remains to show that $F'(0) < 0$. Taking the derivative of F w.r.t. λ gives

$$F'(0) = \frac{2}{n}\left(1 - \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}^{-1} b_j\right]\right)$$

where

$$b_j = \mathbb{E}_X[k(x_j, X)].$$