Journal Club

Consistent Kernel Mean Estimation for Functions of Random Variables (Simon-Gabriel et. al)

CMAP, Ecole Polytechnique

February 16th, 2017

Plan

1. Reminders and Notations
2. Motivations
3. Reduced Set Methods
4. Main Result
5. Probabilistic Programming

Reminders and Notations

Reminders

- $X \in \mathcal{X}, Y \in \mathcal{Y}$ Random Variables
- Function f: $\mathcal{X} \mapsto \mathcal{Z}$
- Positive definite and bounded kernel k: $\mathcal{X} x \mathcal{X} \mapsto \mathbb{R}$
- Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_k$ induced by k:
  - $\mathcal{H}$ Hilbert space
  - inner product $\langle ., . \rangle_{\mathcal{H}_k}$
  - k follows the reproducing property: $f(x) = \langle f(.), k(x, .) \rangle_{\mathcal{H}_k}$

Reminders

Several ways to find approximate representation for random variables

- Monte Carlo approach
  - Generate weighted samples $\{(x_i, w_i), 1 < i < n\}$ from $P(X)$
  - Approximate $E[X]$ by $\frac{\sum_i w_i x_i}{\sum_i w_i}$
  - No notion of 'Best representation'
- Kernel Mean Embeddings
  - Generate weighted samples $\{(x_i, w_i), 1 < i < n\}$ from $P(X)$
  - Represent $P(X)$ by its KME μ_X and $\{(x_i, w_i), 1 < i < n\}$ by its KME $\hat{\mu_X}$

$$\mu_X = \int k(x,.)\,\mathrm{d}P(x) \text{ and } \hat{\mu_X} = \sum_i w_i k(x_i,.) \tag{1}$$

  - μ_X and $\hat{\mu_X}$ belong to the RKHS $\mathcal{H}$ induced by k
  - norm and inner product of that space makes optimization easier

Motivations

Motivations

- Represent/Approximate the distribution f(X)
- Time-accuracy trade-off
- Extend the assumptions under which current results hold
- Have theoretical results for PPL

Reduced Set Methods

Reduced Set Methods (Schölkopf et al.)

- If X and Y requires N samples then $f(X, Y)$ requires N^2 (exponential cost)
- Need to find a way to reduce the sample size while keeping a good approximate for X, Y and $f(X, Y)$
- Several ways
  - min $||\hat{\mu_{X'}} - \hat{\mu_X}||$ under a certain threshold ϵ
  - Sequential Kernel Herding (Lacost-Julien et al.): minimize error ϵ_K at each iteration conditionned on the past samples: $\epsilon_K = ||\mu - \sum_{i=1}^{K} w_i k(x_i, .)||$
- Of course we lose i.i.d. property (of the samples and the weights depending on them) and results of Smola et al. on consistency of estimators does not hold

Reduced Set Methods (Schölkopf et al.)

- Suggests reducing the size of X and Y and estimate $f(X, Y)$ by $\sum_{i,j=1}^{n} w_i u_j k(f(x_i, y_j))$
- Improvements compared to KME of the output distribution of higher complexity ($\mathcal{O}(n^2)$)
- Nevertheless, here, the reduced set methods held the i.i.d. property.

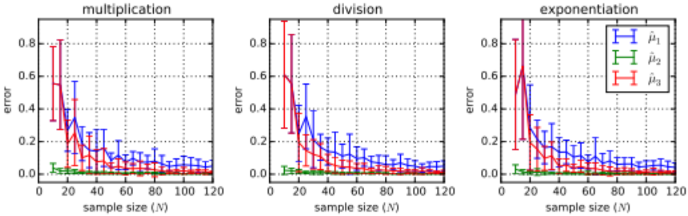

Figure 1: Error of kernel mean estimators for basic arithmetic functions of two variables, $X \cdot Y$, X/Y and X^Y, as a function of sample size N. The U-statistic estimator $\hat{\mu}_2$ works best, closely followed by the proposed estimator $\hat{\mu}_3$, which outperforms the diagonal estimator $\hat{\mu}_1$.

Main Result

Main Result

- Two main results
  1. Consistency of the estimator of the KME of the function of a random variable for non iid samples (quite general setting)
  2. Convergence rate for Matern Kernels with finite samples and smooth function

Consistency of the estimator

- Smola et al, 2007 already showed consistency of KME of X implies consistency of KME of f(X) if the samples are iid
- Assumptions
  1. f: $\mathcal{X} \mapsto \mathcal{Z}$ $\mathcal{C}^0$ with two kernels k_x, c_0-universal and k_z continuous
  2. The weights have to be bounded
- Theorem

  A consistent KME of X leads to a consistent KME of f(X). Even though the samples are no longer i.i.d.

$$\hat{\mu}_X^{k_x} \mapsto \mu_X^{k_x} \Longrightarrow \hat{\mu}_{f(X)}^{k_x} \mapsto \mu_{f(X)}^{k_x} \tag{2}$$

Consistency of the estimator

- Need for a new kernel k'_x such as $\forall (x, x') \in \mathcal{X}, k'_x(x, x') = k_z(f(x), f(x'))$
- Two propositions needed
    - (a) Convergence of KME means weak convergence of distributions
    - (b) Weak convergence of distributions means convergence of KME for kernels defined on bounded sets of $\mathcal{X}$
- (a) and (b) shows $\hat{\mu}_X^{k_x} \mapsto \mu_X^{k_x} \implies \hat{\mu}_X^{k'_x} \mapsto \mu_X^{k'_x}$
- we now need to show $\hat{\mu}_X^{k'_x} \mapsto \mu_X^{k'_x} \implies \hat{\mu}_{f(X)}^{k_z} \mapsto \mu_{f(X)}^{k_z}$

Consistency of the estimator

- Remember $\{(f(x_i), w_i)\}_n$ weighted samples of f(X) and $k_z(f(x_i), .) = k'_x(x_i, .)$

$$
\begin{aligned}
||\hat{\mu}^{k_z}_{f(X)} - \mu^{k_z}_{f(X)}||_{\mathcal{H}_{k_z}} &= ||\sum_{i=1}^{n} w_i k_z(f(x_i), .) - \mathbb{E}[k_z(f(X), .)]||_{\mathcal{H}_{k_z}} \\
&= ||\sum_{i,j=1}^{n} w_i w_j k_z(f(x_i), f(x_j)) - 2\sum_{i=1}^{n} w_i \mathbb{E}[k_z(f(X), f(x_i))) \\
&+ \mathbb{E}[k_z(f(X), f(X'))]||_{\mathcal{H}_{k_z}} \\
&= ||\sum_{i=1}^{n} w_i k'_x(x_i, .) - \mathbb{E}[k'_x(X, .)]||_{\mathcal{H}_{k'_x}} \\
&= ||\hat{\mu}^{k'_x}_{X} - \mu^{k'_x}_{X}||_{\mathcal{H}_{k'_x}} \to 0
\end{aligned}
\tag{3}
$$

Probabilistic Programming

Probabilistic Programming

- Using abstractions of inference algorithm to build short and efficient algorithms with X as input and $f(X)$ as output
- Focus on Bayesian Inference (computing the posterior distribution)
  - Alone, the results are not enough to do inference in probabilistic programs
  - We have to know the KME of X, in other words how to sample from the posterior distribution
  - Kanawaga et. al developed a Kernel Monte Carlo filtering where the KME of the posterior is computed via the Kernel Bayes Rule (Fukumizu et al.)
  - In this method, the samples are generated conditionned on the past samples
  - With this, the results can be used (since no need for i.i.d.) to do inference

Thank you