Kernel Thinning

2/26/2024

Geometry @ UW

mmp

$\subseteq$ finding core-sets

$\neq$ data thinning !!! $\supset$ sample splitting

Problem: approximate distribution μ on $\mathcal{X}$ by a sample

— " — a large sample μ_N by smaller sample

$y_{1:N} \sim \text{iid}\ \mu$ $\rightarrow$ μ_N

empirical distribution

$Q = \{ x_{1:n} \in \mathcal{X} \} \longrightarrow \mu_Q$ = approximator

so that $\mu_Q \approx \mu_N \approx \mu$

Plan

- Kernel Thinning problem
- Ideas for solving it
  - Herding
  - K. Thinning
  - Weighted Quad.
- (Basic) RKHS facts ← geometric
  Important / useful
- The methods : H, KT, WQ (if time)

Problem: approximate distribution μ on $\mathcal{X}$ by a sample

— " — a large sample μ_N by smaller sample

$y_{1:N} \sim_{iid} \mu$

empirical distribution

$Q = \{x_{1:n} \in \mathcal{X}\} \longrightarrow \mu_Q$ = approximator

so that $\mu_Q \approx \mu_N \approx \mu$

More precisely

$\mu f \approx \mu_N f \approx \mu_Q f$ for all $f \in \mathcal{H}_k$

RKHS

uniformly

$$\boxed{\mu f \equiv E_\mu[f(x)]}$$

More precisely

- $\text{MMD}_k(\nu, \mu) = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} |\nu f - \mu f|$

Maximum Mean Discrepancy

Wanted $Q = \{x_{1:n}\}$

so that

$$\boxed{\text{MMD}_k(\mu_N, \mu_Q) < \varepsilon_n}$$

$$\text{MMD}_k(\mu, \mu_Q) < \varepsilon_n$$

Why RKHS ?

- compact specification of $\{f's\}$ - no need to enumerate!
- nice, elegant, well understood math :
  - complete
  - normed
  - "Euclidean" (intuitive)
  - tractable computations
  - non-parametric
  - general
  - .....

Wahba

- for ex:

  polynomials $k = (1 + x^T x')^d$

  kernel regr, classification

  k = Gaussian, Matérn, ...

  string, tree kernels

Why core-sets / thinning ?

want to do ML with large data $\Longrightarrow$ save computation time, tractability

- non-parametric
- f not known yet

$\leftarrow$ Kernel $k(\ )$ must match problem !!

Why expectations $E_\mu[f]$

many functionals (criteria of quality) are expectations

- P[err]
- $E[err^2]$
- quantile

Problem (rephrased)

given • $y_{1:N} \equiv \mu_N$ large sample

• kernel $k \equiv$ RKHS $\mathcal{H}_k$

wanted • $x_{1:n} \equiv \mu_Q$ small sample

so that $\text{MMD}_k(\mu_N, \mu_Q) \leq \varepsilon_n$

What is a reasonable ε_n?

• $\mu_N \sim \text{iid}\ \mu$

$\mu_N \approx \mu \Rightarrow$ want same rate for $\mu_Q \approx \mu_N$

• $E_\mu[x] = \mu_X$
$$\underset{\mathcal{H}_k}{\cap}$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y_i \rightarrow \text{Var}\,\hat{\mu} \propto \frac{1}{N}$$

$$\text{std}\,\hat{\mu} \propto \frac{1}{\sqrt{N}} = N^{-1/2}$$

benchmark rate

MMD_k rate no faster than $\hat{\mu}$ rate

Problem (rephrased)

given
- $y_{1:N} \equiv \mu_N$ large sample
- kernel $k \equiv$ RKHS $\mathcal{H}_k$

wanted
- $x_{1:n} \equiv \mu_Q$ small sample

so that $\quad MMD_k(\mu_N, \mu_Q) \leq \varepsilon_n$

Solutions

- grid on $\mathcal{X}$ ← optimal N^{-2r}
- MC = iid sampling $\quad N^{-1/2}$
- MCMC $\sim N^{-\alpha} \quad \alpha < \frac{1}{2}$
- Quasi MC $\sim N^{-\alpha'} \quad \alpha' > \alpha$
- Herding
- DPP (Determinantal Point P)
- Thinning
- Weighted Quadrature

dependent (not iid)

neg. correl.

Problem (rephrased)

given • $y_{1:N} \equiv \mu_N$ large sample

• kernel k $\equiv$ RKHS $\mathcal{H}_k$

wanted • $x_{1:n} \equiv \mu_Q$ small sample

so that $\mathrm{MMD}_k(\mu_N, \mu_Q) \leq \varepsilon_n$

• all select subsample Q of $y_{1:N}$

• $n = \sqrt{N}$

rate $\sim \underbrace{n^{-1} = N^{-1/2}}$ but with $\sqrt{N}$ points

Solutions

• grid on $\mathcal{X}$ ← optimal N^{-2r}

• MC = iid sampling $N^{-1/2}$

• MCMC $\sim N^{-\alpha}$ $\alpha < \frac{1}{2}$

• Quasi MC $\sim N^{-\alpha'}$ $\alpha' > \alpha$

seq uential greedy

• Herding [Yutian Chen, Welling, Smola '12]

• DPP (Determinantal Point P) [Belhadji ...]

recursive vector balancing

• Thinning [Dwivedi, Mackey 21]

(paired comparisons)

• Weighted Quadrature

+ square root kernel

weighted sample [Hayakawa, Ober hauser, Lyons 21]

RKHS approximation*

+ cubature algorithm (Caratheodory)*

What makes thinning pb. hard? $\varepsilon_n \sim \underline{\underline{C}}\, n^{-1}$
core-set

rate of decay of kernel, μ

For ex:

μ compact support
$\wedge$
μ subgaussian
$\wedge$
μ heavy tailed

- similar for k : compact support, ...

Weighted Quadrature : $(\sigma_{1,2,\dots}, e_{1,2,\dots}(x))$ = e-values, e-functions of operator

$$(kf)(x) = \int_{\mathcal{X}} f(x')\, k(x', x)\, d\mu(x')$$

typically convolution op.

$\Rightarrow$ err depends on σ_n, $r_{n+1} = \sum_{j=n+1}^{\infty} \sigma_j$ (residual spectrum)

Kernel thinning

kernel decay radius

- $R_{k,n} = \inf r$
  $\sup_{\|x-y\|_2 \geq r} |k(x,y)| \leq \frac{1}{n} \|k\|_\infty$

- $\tau_k(r) = \sup_x \int_{\|y\|_2 \geq r} k^2(x, x-y)\, dy$ ← tail weight of $k^2(\,)$ at r

Ex: $k(x,y) = N(x-y; 0, \sigma^2 I) \Rightarrow \tau_k(r) = \int_{\|y\| \geq r} e^{-2\frac{\|x-(x-y)\|^2}{2\sigma^2}} dy$ = tail prob for $N(0, \frac{\sigma^2}{2} I)$ at $r \cdot \sqrt{2}$

Gaussian kernel

for $\sigma = 1$: $\|k\|_\infty = \frac{1}{(2\pi)^{d/2}}$

$$\|x-y\| \geq r \Rightarrow k(x,y) \leq \frac{1}{(2\pi)^{d/2}} e^{-\frac{r^2}{2}}$$

$$\Rightarrow e^{-\frac{R^2}{2}} = \frac{1}{n} \Rightarrow \boxed{R_{k,n} = \sqrt{2 \ln n}}$$

It helps if these are known analytically ← otherwise they must be approximated, majorized, ...

or exact, tractably

- kernel tail probabilities $\tau_k(r)$ computable

  decay radii $R_{k,n}$

  (+ decay rates of μ) (Thinning)

- $\bar{\varphi} = E_\mu[\varphi(x)]$ or $\langle z, \bar{\varphi} \rangle_k \in \mathbb{R}$

  $\in \mathcal{H}_k$ $\qquad$ $\cap$ $\mathcal{H}_k$

  $\max_x \langle z, \varphi(x) \rangle$ (Herding)

  $\cap$ $\mathcal{H}_k$

- (Bochner) Fourier repr of $k(\cdot,\cdot)$ → Thinning $k_{1/2}$

  Spectrum of operator k $\quad (\sigma_{1,2,\dots m}, e_{1,2,\dots m}(x))$

  → weighted Quadrature

$\boxed{\text{Bochner}}$

$\boxed{\text{Caratheodory}}$

$\boxed{\text{Universal kernel}}$

$\boxed{\text{Weight optimization}}$

$\boxed{\text{Computing } MMD_k}$

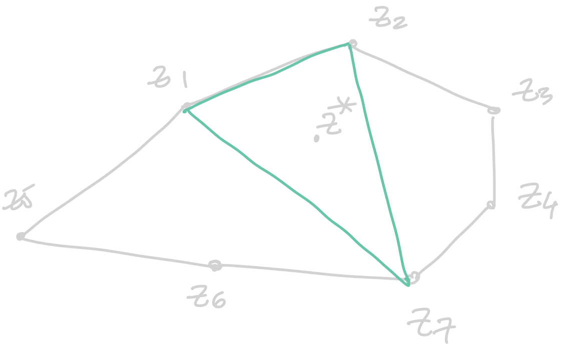
Bochner

Caratheodory

<u>Thm</u> [Caratheodory] $z_{1:N} \in \mathbb{R}^{d-1}$, $z^* \in \text{convex hull}(z_{1:N}) \Longrightarrow$

$z^* \in \text{conv. hull}(z'_1, \dots z'_d)$ where

$z'_{1:d} \subset \{z_{1:N}\}$



$$\Longrightarrow z^* \in \text{conv}(\underbrace{z_1, z_2, z_7}_{d+1})$$

$d=2 \Longrightarrow$ $d+1$ points sufficient

For us:

$$\left.\begin{array}{l} \text{functions } \varphi_{1:n-1} \\ \text{points } y_{1:N} \end{array}\right\} \Rightarrow \text{define } z_i = \begin{bmatrix} \varphi_1(y_i) \\ \vdots \\ \varphi_{n-1}(y_i) \end{bmatrix} \in \mathbb{R}^{n-1} \qquad \bar{z} = \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \varphi_j(y_i) \\ \vdots \end{bmatrix} \leftarrow \bar{z}_j$$

$$\Rightarrow \left[\begin{array}{l} \exists\, x_{1:n} \in \{y_{1:N}\} \text{ so that } \bar{z}_j = \sum_{i'=1} w_{i'}\, \varphi_j(x_{i'}) \text{ for } j = 1:n-1 \\ w_{1:n} \geq 0, \quad \sum w_{i'} = 1 \end{array}\right.$$

$\boxed{\text{Universal kernel}}$

If k universal kernel $\not\Longrightarrow |\mu f - \nu f| \leq \|f\|_{\mathcal{H}} \|\mu - \nu\|_{\mathcal{H}}$

[Koksma Hlavka inequality]

Computing MMD_k

from Weighted Quadrature (14)

$$MMD_k(\mu_Q, \mu)^2 = \| \mu_Q \varphi - \mu \varphi \|_{\mathcal{H}}^2$$

$\varphi : \mathcal{X} \to \mathcal{H}$ feature map

$$= w^T k(X,X) w - 2 E_y [w^T k(X,y)] + E_{y,y'} [k(y,y')]$$

weights $w_{1:n}$

or $1/n$

$\mu = \mu_N$

or need to know analythically

Weight optimization

Quadratic in $w \in \mathbb{R}^n \Rightarrow \min_w MMD^2$ s.t. $w \geq 0$, $\sum w_i = 1$ with $x_{1:n}, y_{1:N}$ given

- optimizes any core-set (e.g. from QMC, Thinning, ...)