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

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

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

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

RKHS

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

uniformly

More precisely

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

Maximum Mean Discrepancy

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

so that

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

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

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 $\longleftarrow$ 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 $\quad \mathrm{MMD}_k(\mu_N, \mu_Q) \leq \varepsilon_n$

What is a reasonable ε_n ?

- $\mu_N \sim \text{iid } \mu$

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

- $E_\mu[x] = \mu_X$

  $\mu_X \in \mathcal{H}_k$

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

  $\mathrm{std}\,\hat{\mu} \propto \frac{1}{\sqrt{N}} = N^{-1/2}$ ← 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 $\mathrm{MMD}_k(\mu_N, \mu_Q) \leq \varepsilon_n$

dependent (not iid)

neg. corr.

Solutions

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

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

sequential greedy

recursive vector balancing (paired comparisons)

+ square root kernel

Solutions

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

• MC = iid sampling $N^{-1/2}$

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

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

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

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

• Thinning [Dwivedi, Mackey 21]

• Weighted Quadrature

weighted sample [Hayakawa, Oberhauser, 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

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

- $\tau_k(r) = \sup_{x} \int_{\|y\|_2 \geq r} k^2(x, x-y)\, dy \leftarrow$ 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, ...

- kernel tail probabilities $\tau_k(r)$ computable or exact, tractably
  decay radii $R_{k,n}$
  (+ decay rates of μ) (Thinning)
- $\bar{\varphi} = E_\mu[\varphi(x)]$ or $\langle z, \bar{\varphi} \rangle_k \in \mathbb{R}$
  $\bar{\varphi} \in \mathcal{H}_k$, $z \in \mathcal{H}_k$

  $\max_x \langle z, \varphi(x) \rangle$, $z \in \mathcal{H}_k$ (Herding)

- (Bochner) Fourier repr of $k(,)$ → Thinning $k_{1/2}$
  Spectrum of operator k $(\sigma_{1,2,..m}, e_{1,2,..m}(x))$
  → weighted Quadrature

$\boxed{\text{Bochner}}$

$\boxed{\text{Caratheodory}}$

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

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

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

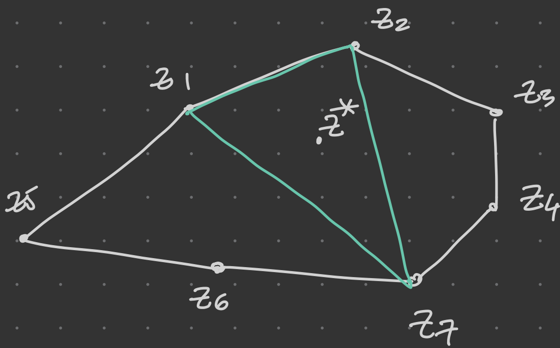
Bochner

Caratheodory

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

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

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



$\Longrightarrow z^* \in \text{conv}(z_1, z_2, z_7)$

$d=2 \Longrightarrow$ $\underbrace{d+1}$ points sufficient

For us:

functions $\varphi_{1:n-1}$, points $y_{1:N}$ $\Big\} \Rightarrow$ define $z_i = \begin{bmatrix} \varphi_1(y_i) \\ \vdots \\ \varphi_{n-1}(y_i) \end{bmatrix} \in \mathbb{R}^{n-1}$

$$\bar{z} = \begin{bmatrix} \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \varphi_j(y_i) \\ \vdots \end{bmatrix} \leftarrow \bar{z}_j$$

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

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

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

$\boxed{\text{[Koksma Hlavka inequality]}}$

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

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

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, ...)