L5: Two-sample problem and MMD

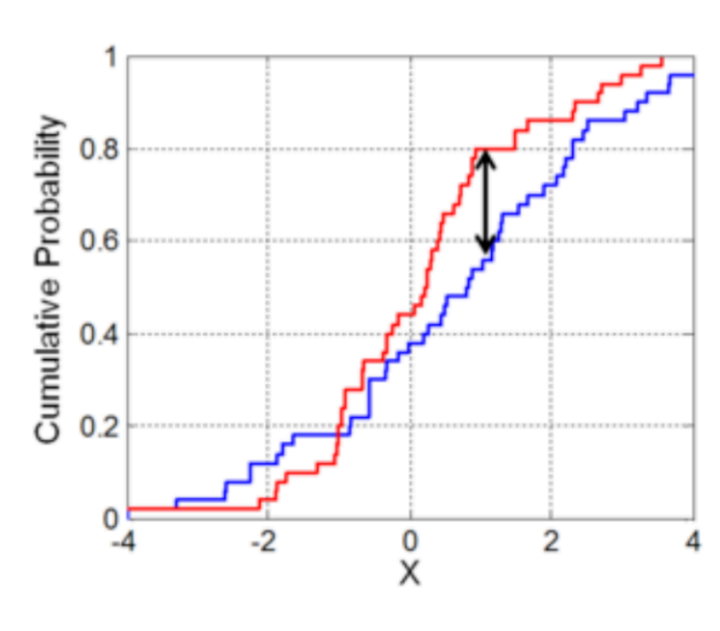
Two-sample problem

given $X = \{x_i\}_{i=1}^{n_X}$, $Y = \{y_i\}_{i=1}^{n_Y}$,
$x_i \sim P$, $y_i \sim Q$, i.i.d., $X \perp\!\!\!\perp Y$
Can we tell if $P = Q$ from the data?
When $P \neq Q$, where do they differ?

Rk. A problem in the GAN loss. Related to the estimation of distance / divergence btw two (unknown) distributions from finite sample.

Standard procedure:

- Compute test statistic $T(X,Y)$
- Specify a threshold value τ
- Accept $\mathcal{H}_0$ if $T(X,Y) < \tau$, reject otherwise



In 1D: Kolmogorov-Smirnov, Wald-Wolfowitz runs, ...

We consider general P and Q in $\mathbb{R}^d$.

- MMD (maximum mean discrepancy)
  Integral Probability Metrics (IPM)

$$D_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)|$$

Eg $\mathcal{F} = \{ f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k} \le 1 \}$ kernel MMD.

Eg $dP = p\,dx$, $dQ = q\,dx$, p, q densities

$$W_1(p,q) = \sup_{Lip(f) \le 1} \left| \int f(p-q) \right|$$

$$Lip(f) := \sup_{x \neq y \in \mathcal{X}} \frac{|f(x) - f(y)|}{\|x-y\|} \quad \text{Lipschitz semi-norm}$$

Wasserstein-1 distance (Earth-Mover distance)

- f-divergence: $dP = p\,dx$, $dQ = q\,dx$,

$f: \mathbb{R}_+ \to \mathbb{R}$, convex, lower-semicont', $f(1) = 0$

$$D_f(P \| Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

Eg $f(u) = u \log u$. KL-divergence $\int p \log \frac{p}{q}$

Eg $f(u) = (\sqrt{u} - 1)^2$, Square Hellinger $\int (\sqrt{p} - \sqrt{q})^2$

Other: Jensen-Shannon Divergence
Pearson χ^2

Rk. density ration p/q is only defined where $q(x) > 0$. Thus mathematically, MMD can handle when supp(p) and supp(q) differ.

- MMD is a pseudometric on $\mathcal{P}(\mathcal{X})$ the space of prob. measures.

(i) $D_{\mathcal{F}}(P,P) = 0$

(ii) $D_{\mathcal{F}}(P,Q) = D_{\mathcal{F}}(Q,P)$

(iii) Triangle inequality (Ex)
$$D_{\mathcal{F}}(P,R) \le D_{\mathcal{F}}(P,Q) + D_{\mathcal{F}}(Q,R)$$

However, generally, $D_{\mathcal{F}}(P,Q) = 0 \not\Rightarrow P = Q$

- RKHS-MMD (kernel MMD)

$\mathcal{H}_k$ the RKHS associated with PSD kernel k on $\mathcal{X}$, $(\mathcal{X}, dx)$ the measure space
P and Q are prob. distributions on $\mathcal{X}$, have densities.

kernel MMD: (the absolute value in def can removed)

$$MMD(p,q) = \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1} \int f(p-q)$$

def Mean embedding of P is $\mu_p(x) = \mathbb{E}_{y\sim p} k(x,y)$

Lemma If $\mathbb{E}_{x\sim p} \sqrt{k(x,x)} < \infty$, then $\mu_p \in \mathcal{H}$.

pf. $\mu_p(x)$ is well defined b/c (by C-S)
$|k(x,y)| \le \sqrt{k(x,x)}\sqrt{k(y,y)}$,
thus $k(x,y)$ is integrable under $\mathbb{E}_{y\sim p}$.

$\mu_p = \mathbb{E}_{y\sim p} \phi(y)$, then

$$\|\mu_p\|_{\mathcal{H}}^2 = \langle \mu_p, \mu_p \rangle_{\mathcal{H}} = \langle \mathbb{E}_{y\sim p} \phi(y), \mathbb{E}_{x\sim p} \phi(x) \rangle_{\mathcal{H}}$$
$$= \mathbb{E}_{x\sim p, y\sim p} \underbrace{\langle \phi(y), \phi(x) \rangle_{\mathcal{H}}}_{= k(x,y)}$$
$$\overset{(C-S)}{\le} \mathbb{E}_{x\sim p, y\sim p} \sqrt{k(x,x)}\sqrt{k(y,y)} < \infty \quad \#$$

Prop
$$MMD(p,q)^2 = \|\mu_p - \mu_q\|_{\mathcal{H}}^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)(p-q)(x)(p-q)(y)\,dx\,dy$$

pf. For $f \in \mathcal{H}$, $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}$, then

$$(\mathbb{E}_{x\sim p} - \mathbb{E}_{x\sim q}) f(x) = \langle f, \mathbb{E}_{x\sim p}\phi(x) - \mathbb{E}_{x\sim q}\phi(x) \rangle_{\mathcal{H}}$$
$$= \langle f, \mu_p - \mu_q \rangle_{\mathcal{H}},$$

thus $\sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \le 1} \langle f, \mu_p - \mu_q \rangle_{\mathcal{H}}$ is achieved when

$$f = \frac{\mu_p - \mu_q}{\|\mu_p - \mu_q\|_{\mathcal{H}}}, \quad \text{if } \|\mu_p - \mu_q\|_{\mathcal{H}} > 0.$$

and then $MMD(p,q) = \|\mu_p - \mu_q\|_{\mathcal{H}}$.
(When $\|\mu_p - \mu_q\|_{\mathcal{H}} = 0$, $MMD = 0$)

Thus $MMD(p,q)^2 = \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_{\mathcal{H}}$
$$= \langle (\mathbb{E}_{x\sim p} - \mathbb{E}_{x\sim q})\phi(x), (\mathbb{E}_{y\sim p} - \mathbb{E}_{y\sim q})\phi(y) \rangle_{\mathcal{H}}$$
$$= (\mathbb{E}_{x\sim p} - \mathbb{E}_{x\sim q})(\mathbb{E}_{y\sim p} - \mathbb{E}_{y\sim q}) k(x,y) \quad \#$$

Rk The pf shows that the maximizer equals $(\mu_p - \mu_q)$ up to a normalising constant. By definition,

$$(\mu_p - \mu_q)(x) = \int_{\mathcal{X}} k(x,y)(p-q)(y)\,dy =: w(x)$$

witness function of kernel MMD.

Q. kernel k, $MMD_k(P,Q) = 0 \overset{?}{\Rightarrow} p = q$

Let $\mathcal{X}$ be a compact metric space, $M_b(\mathcal{X})$ the set of finite signed measures on $\mathcal{X}$,

def Kernel k is universal if for any $\mu \in M_b(\mathcal{X})$,

$$\int_{\mathcal{X}} k(\cdot, x)\,d\mu(x) = 0 \Rightarrow \mu = 0$$

Prop The def is equivalent to that

$$\int_{\mathcal{X}\times\mathcal{X}} k(x,y)\,d\mu(x)\,d\mu(y) > 0, \quad \forall \mu \in M_b(\mathcal{X}) \setminus \{0\}. \quad (*)$$

Ref. Sriperumbudur et al. On the relation between universality, characteristic kernels and RKHS embedding of measures. AISTATS 2010.

Rk. If k is universal, then $MMD_k = 0 \Rightarrow p = q$.

Rk The def $(*)$ of universal kernel can be extended to non-compact $\mathcal{X}$.

For translation invariant kernel on $\mathbb{R}^d$,

$$k(x,y) = h(x-y), \quad h: \mathbb{R}^d \to \mathbb{R}, \quad h(x) = h(-x),$$

$\hat{h}(\xi)$ the Fourier transform of h,

$$h(x) = \int_{\mathbb{R}^d} e^{i x^T \xi} \hat{h}(\xi)\,d\xi, \quad \hat{h}(\xi) \in \mathbb{R}$$

Prop The kernel $h(x-y)$ is universal iff. $\mathrm{Supp}(\hat{h}) = \mathbb{R}^d$

Pf $\forall \mu \in M_b(\mathbb{R}^d)$, $\mu \neq 0$, b/c $\int_{\mathbb{R}^d} |d\mu| < \infty$,

$$\hat{\mu}(\xi) := \int_{\mathbb{R}^d} e^{-i x^T \xi}\,d\mu(x) \quad \text{is finite } \forall \xi \in \mathbb{R}^d.$$

$$\iint h(x-y)\,d\mu(x)\,d\mu(y) = \iiint e^{i(x-y)^T\xi} \hat{h}(\xi)\,d\xi\,d\mu(x)\,d\mu(y)$$
$$= \int_{\mathbb{R}^d} \left(\int e^{i x^T \xi} d\mu(x)\right)\left(\int e^{-i y^T \xi} d\mu(y)\right) \hat{h}(\xi)\,d\xi$$
$$= \int_{\mathbb{R}^d} |\hat{\mu}(\xi)|^2 \hat{h}(\xi)\,d\xi$$

thus if $\mathrm{Supp}(\hat{h}) = \mathbb{R}^d$, then r.h.s $= 0$ only when $\hat{\mu}(\xi) = 0 \Rightarrow \mu = 0$.
This proves the "$\Rightarrow$" direction of the prop. the other direction "$\Leftarrow$" as Ex. #.

Eg Universal kernels in $\mathbb{R}^d$:

$e^{-\|x-y\|^2/\epsilon}$ Gaussian kernel
$e^{-\|x-y\|_1}$ Laplace kernel

- Estimation of MMD from data

$MMD(p,q)$ is called population MMD, given finite samples $X = \{x_i\}_{i=1}^{n_X}$, $Y = \{y_i\}_{i=1}^{n_Y}$,
$x_i \sim p$ i.i.d., $y_i \sim q$ i.i.d., $X \perp\!\!\!\perp Y$,
How to estimate MMD from X and Y?

Denote $T = MMD(p,q)^2 = \iint k(x,y)(p-q)(x)(p-q)(y)\,dx\,dy$

$$\hat{p} := \frac{1}{n_X}\sum_{i=1}^{n_X} \delta_{x_i}, \quad \hat{q} := \frac{1}{n_Y}\sum_{i=1}^{n_Y} \delta_{y_i},$$

empirical distributions of X and Y.

$$\hat{T} := \iint k(x,y)(\hat{p}-\hat{q})(x)(\hat{p}-\hat{q})(y)\,dx\,dy$$
$$= \frac{1}{n_X^2}\sum_{i,j=1}^{n_X} k(x_i,x_j) - \frac{2}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} k(x_i,y_j) + \frac{1}{n_Y^2}\sum_{i,j=1}^{n_Y} k(y_i,y_j)$$

	n_X	n_Y
n_X	K_{XX}	K_{XY}
n_Y	K_{XY}^T	K_{YY}

$=: K$

Rk. $\hat{T}$ is an biased estimator of T. An unbiased estimator is to exclude the diagonal entries in K_{XX}, K_{YY}

$$\hat{T}_u = \frac{1}{n_X(n_X-1)}\sum_{i\neq j,1}^{n_X} k(x_i,x_j) + \frac{1}{n_Y(n_Y-1)}\sum_{i\neq j,1}^{n_Y} k(y_i,y_j) - \frac{2}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} k(x_i,y_j)$$

$\hat{T}$ can be computed from the kernel matrix K computed from pairwise measurements in $X \cup Y$, which has $O(n^2)$ complexity (memory and computation)

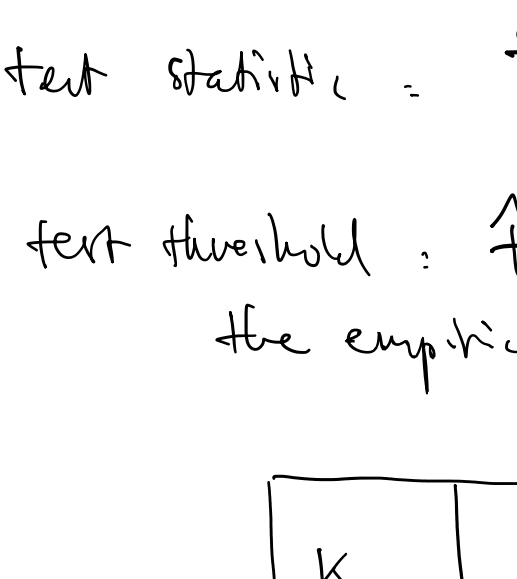
- $O(n)$ MMD statistic

Suppose $n_X = n_Y = n$, $X = \{x_i\}_{i=1}^n$, $Y = \{y_i\}_{i=1}^n$

consider $z_i = (x_i, y_i)$, $i = 1, \dots, n$

$$h(z_i, z_j) := k(x_i,x_j) + k(y_i,y_j) - k(x_i,y_j) - k(x_j,y_i)$$

$$\hat{T}_{lin} := \frac{1}{\lfloor n/2 \rfloor}\sum_{i=1}^{\lfloor n/2 \rfloor} h(z_{2i-1}, z_{2i})$$



$\hat{T}_{lin}$ is a consistent estimator of MMD^2.

$\hat{T}_{lin}$ can be computed using $O(n)$ time / memory.

- kernel two-sample test.

given X and Y, $x_i \sim p$, $y_i \sim q$,

test statistic: $\hat{T}(X,Y)$ kernel MMD.
test threshold: $\hat{t}$ computed by bootstrap the empirical 0.95-quantile of $\hat{T}_{null}$

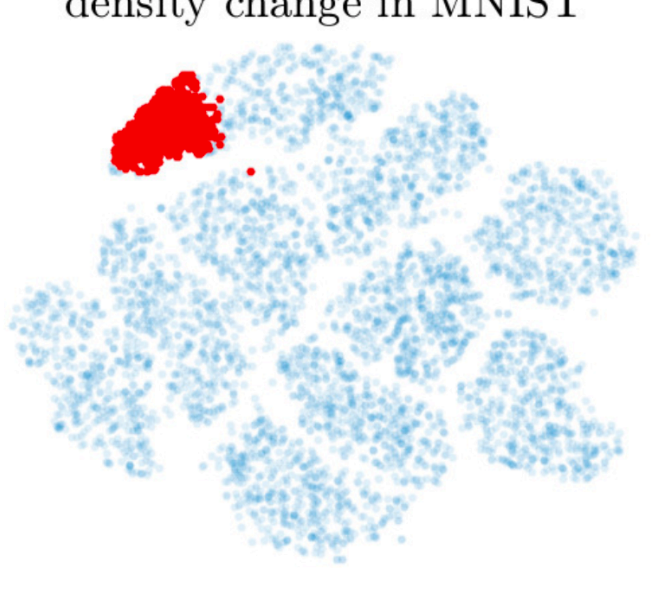
$K =$	K_{XX}	K_{XY}
	K_{XY}^T	K_{YY}

X: label 0
Y: label 1

$\hat{T}_{null}$ computed $K(ind, ind)$, ind is randomly permuted indices $1:n_X+n_Y$

Ref. Arcones and Gine. On the bootstrap of u and v statistics. The Annals of Statistics (1992).

density change in MNIST



(tSNE embedding in 2D)

MMD statistic (red cross) and bootstrapped distribution under null (blue histogram)