Lecture 20: MMD in Practice

last lecture we looked at the basic properties of kernel mean embeddings of distributions

$$P(\mathcal{X}) \ni \mu \longmapsto \mu_K := \mathbb{E}_{\underline{x}\sim\mu} K(\underline{x},\cdot) \in \mathcal{H}_K$$
$$= \int K(\underline{\xi},\cdot)\,\mu(d\underline{\xi})$$

& the associated Maximum mean discrepancy

$$\mathrm{MMD}(\mu,\mu') := \| \mu_K - \mu'_K \|_K$$

The main take aways were:

- If $\int \sqrt{K(\underline{x},\underline{x})}\,\mu(d\underline{x}) < +\infty$ then μ_K is the Riesz. rep. of the bdd. lin operator

$$\phi \in \mathcal{H}_K^*, \quad \phi(f) = \int f(\underline{x})\,\mu(d\underline{x})$$
$$= \langle f, \mu_K \rangle_K$$

- MMD satisfies all but one property of a metric, ie, we may have in general

$$\mathrm{MMD}(\mu,\mu') = 0 \quad \text{while} \quad \mu \neq \mu'$$

As an example consider the linear kernel $K(\underline{x},\underline{z}) = \underline{x}^T\underline{z}$ then we have

$$\mu_k(\underline{z}) = \mathbb{E}_{\underline{x}\sim\mu} \underline{x}^T\underline{z} = \underline{m}^T\underline{z}$$

$$\mu'_k(\underline{z}) = \mathbb{E}_{\underline{x}\sim\mu'} \underline{x}^T\underline{z} = (\underline{m}')^T\underline{z}$$

$\underline{m}, \underline{m}'$ are mean of μ, μ'

& so, $MMD(\mu,\mu')^2 = \|\underline{m} - \underline{m}'\|^2$ so that $MMD(\mu,\mu') = 0$ so long as μ, μ' have the same mean!

- We introduced the idea of a Characteristic kernel, ie, a kernel k such that $MMD(\mu,\mu') = 0 \iff \mu = \mu'$ for all $\mu,\mu' \in \mathbb{P}_k(\mathcal{X})$.

eg: RBF kernel $k(\underline{x},\underline{x}') = \exp\left(-\frac{\|\underline{x}-\underline{x}'\|^2}{2\gamma^2}\right)$

Fourier $k(\underline{x},\underline{x}') = \exp(i\,\underline{x}^T\underline{x}')$

(see Table in Lec. 19).

In this lecture we want to consider more practical aspects of the MMD & kernel mean embeddings.

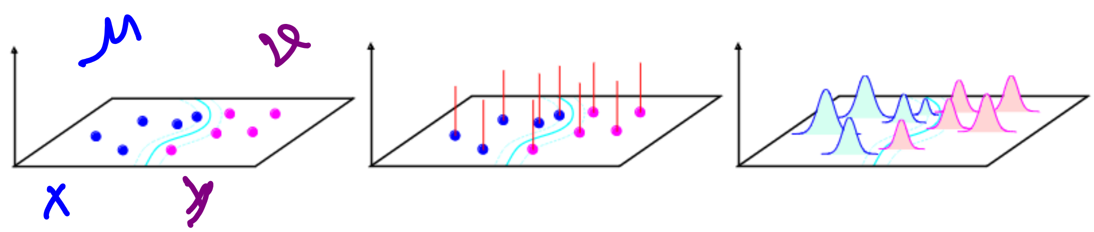
20.1 - empirical estimation of MMD

In most practical applications in statistics or data science we are given data sets

$$\mathcal{X} = \{\underline{x}_1, \dots, \underline{x}_n\} \subset \mathbb{R}^d$$

$$\mathcal{Y} = \{\underline{y}_1, \dots, \underline{y}_m\} \subset \mathbb{R}^d$$

& the general assumption is that these data sets are drawn (independently) from underlying measures μ, ν.



So, at best, assuming $\underline{x}_j$ & $\underline{y}_j$ are iid wrt μ, ν, we can approximate the MMD between the empirical distributions

$$\mu^n = \frac{1}{n} \sum_{j=1}^{n} \delta_{\underline{x}_j} \quad , \quad \nu^n = \frac{1}{m} \sum_{j=1}^{m} \delta_{y_j}$$

It turns out, that the reproducing property allows us to write a very simple expression for computing the empirical MMD.

$$\mu^n_K = \int K(\underline{x},\cdot)\mu^n(dx) = \frac{1}{n}\sum_{j=1}^{n} K(\underline{x}_j,\cdot)$$

$$\nu^m_K = \int K(\underline{x},\cdot)\nu^m(dx) = \frac{1}{m}\sum_{j=1}^{m} K(\underline{y}_j,\cdot)$$

& so,

$$\mathrm{MMD}(\mu^n,\nu^n)^2 = \|\mu^n_K - \nu^m_K\|_K^2$$

$$= \|\mu^n_K\|_K^2 + \|\nu^m_K\|_K^2 - 2\langle \mu^n_K, \nu^m_K\rangle$$

$$= \Big\|\frac{1}{n}\sum_{j=1}^{n} K(\underline{x}_j,\cdot)\Big\|_K^2 + \Big\|\frac{1}{m}\sum_{j=1}^{m} K(\underline{y}_j,\cdot)\Big\|_K^2$$

$$- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle K(\underline{x}_i,\cdot), K(\underline{y}_j,\ )\rangle$$

$$= \frac{1}{n^2}\sum_{i,j=1}^{n} K(\underline{x}_i,\underline{x}_j) + \frac{1}{m^2}\sum_{i,j=1}^{m} K(\underline{y}_i,\underline{y}_j)$$

$$- \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K(\underline{x}_i,\underline{y}_j) \qquad \text{(A)}$$

This convenient expression is one of the most attractive features of the MMD!

The question arises as to how good of an approx. is $\mathrm{MMD}(\mu^n, \nu^m)$?

Thm: Let k be a continuous & PDS kernel on a separable metric space $\mathcal{X}$ such that $\sup_{\underline{x} \in \mathcal{X}} k(\underline{x}, \underline{x}) < C_k < +\infty$. Then for any $\delta \in (0,1)$ with prob. at least $1-\delta$ we have

$$\mathrm{MMD}(\mu, \mu^n) \leq \sqrt{\frac{C_k}{n}} + \sqrt{\frac{2 C_k \log \frac{1}{\delta}}{n}}$$

By triangle ineq. we have

$$\mathrm{MMD}(\mu, \nu) \leq \mathrm{MMD}(\mu, \mu^n) + \mathrm{MMD}(\mu^n, \nu)$$

$$\leq \mathrm{MMD}(\mu, \mu^n) + \mathrm{MMD}(\mu^n, \nu^n) + \mathrm{MMD}(\nu^n, \nu)$$

$$\Rightarrow \left| \mathrm{MMD}(\mu, \nu) - \mathrm{MMD}(\mu^n, \nu^n) \right| = O\left(\sqrt{\tfrac{1}{n}} \vee \sqrt{\tfrac{1}{m}}\right)$$

with high prob.

In other words, convergence happens at monte carlo rate!

Computing MMD using the formula ⊛ is in general an $\mathcal{O}(n^2 d)$ operation (assuming $m = \mathcal{O}(n)$). This can be prohibitive in some cases. Luckily there are many work arounds, such as random features.

Recall if we have a stationary kernel $K(\underline{x}, \underline{x}') = K(\underline{x} - \underline{x}')$ then we could approx our kernel with

$$K^N(\underline{x}, \underline{x}') = \frac{1}{N} \sum_{j=1}^{N} \varphi_j(\underline{x}) \varphi_j(\underline{x}')$$

where φ_j where the random features of our kernel. We also showed that the RKHS of K^N consists of functions $f = \sum_{j=1}^{N} \frac{\alpha_j}{\sqrt{N}} \varphi_j$ with RKHS norm $\|f\|^2 = \|\underline{\alpha}\|_2^2$.

Based on this we can immediately see that

$$\mathrm{MMD}_k(\mu,\nu)^2 \approx \mathrm{MMD}_{K^N}(\mu,\nu)^2$$

$$= \| \mu_{K^N} - \nu_{K^N} \|_{K^N}^2$$

$$= \left\| \int K^N(\underline{x},\cdot)\,\mu(d\underline{x}) - \int K^N(\underline{x},\cdot)\,\nu(d\underline{x}) \right\|_{K^N}^2$$

$$= \left\| \int \sum_{j=1}^{N} \frac{1}{\sqrt{N}} \varphi_j(\underline{x}) \frac{1}{\sqrt{N}} \varphi_j(\cdot)\,\mu(d\underline{x}) - \int \sum_{j=1}^{N} \frac{1}{\sqrt{N}} \varphi_j(\underline{x}) \frac{1}{\sqrt{N}} \varphi_j(\cdot)\,\mu(d\underline{x}) \right\|_{K^N}^2$$

$$= \left\| \sum_{j=1}^{N} \left(\int \frac{1}{\sqrt{N}} \varphi_j(\underline{x})\,\mu(d\underline{x}) - \int \frac{1}{\sqrt{N}} \varphi_j(\underline{x})\,\nu(d\underline{x}) \right) \frac{1}{\sqrt{N}} \varphi_j \right\|_{K^N}^2$$

$$= \frac{1}{N} \sum_{j=1}^{N} \left| \int \varphi_j(\underline{x})\,\mu(d\underline{x}) - \int \varphi_j(\underline{x})\,\nu(d\underline{x}) \right|^2$$

$$\approx \frac{1}{N} \sum_{j=1}^{N} \left| \frac{1}{n} \sum_{l=1}^{n} \varphi_j(\underline{x}_l) - \frac{1}{m} \sum_{l=1}^{m} \varphi_j(\underline{y}_l) \right|^2$$

for $\underline{x}_l \sim \mu$, $\underline{y}_l \sim \nu$.

Observe that this formula not only involves the mean of N random features. It is also highly parallelizable since the $\varphi_j(\underline{x}_i)$ can be computed independently for each feature.

Also observe that the above calculation can be repeated for any other low-rank approx. to our kernel.

$$K(\underline{x}, \underline{x}') \approx \sum_{j=1}^{N} \psi_j(\underline{x}) \psi_j(\underline{x}')$$