Maximum Mean Discrepancy

Uri Shaham

October 11, 2022

1 Mean Embedding

1.1 Two Sasmple tests

Given two samples, $x_1, \ldots, x_n \sim P$, $y_1, \ldots, y_m \sim Q$ the questions is whether $P = Q$. Some plug-in options: Kolmogorov Smirnov (in 1 dim). We can estimate the densities and check the difference. But this is problematic in high dimensions.

1.2 Mean Embedding

The idea: Look for a function $f \in \mathcal{F}$ that can distinguish between P and Q through means

$$D(P, Q, \mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{x \in P}[f(x)] - \mathbb{E}_{x \in Q}[f(x)]$$

Definition 1.1 (Universal kernel)**.** *A kernel k is called universal if its corresponding RKHS $\mathcal{H}$ is dense in $\mathcal{C}(\mathcal{X})$ (i.e., if for every bounded continuous function of $\mathcal{X}$, there is a sequence of functions in $\mathcal{H}$ converging to it pointwise.*

For example, the RBF kernel is known to be universal.

Theorem 1.2 (Stainwart 2001, Smola et al.2006)**.** *Let H be a universal RKHS and $\mathcal{F}$ be a unit ball in it, i.e., $\mathcal{F} = \{f \in \mathcal{F} | \|f\| \leq 1\}$. Then $D(P, Q, \mathcal{F}) = 0$ iff $P = Q$.*

Proof. (informal) The direction $\Leftarrow$ is obvious. If $P \neq Q$, there exists a continuous and bounded f, such that $\mathbb{E}_{x \in P}[f(x) - \mathbb{E}_{x \in Q} f(x) = \epsilon > 0$. Then since $\mathcal{H}$ is universal, we can find $f^* \in \mathcal{H}$ such that $\|f - f^*\|_\infty = \frac{\epsilon}{2}$. Then

$$\mathbb{E}_{x \in P}[f^*(x) - \mathbb{E}_{x \in Q} f^*(x) > \mathbb{E}_{x \in P}[f(x) - \mathbb{E}_{x \in Q} f(x) - 2\|f - f^*\|_\infty > 0.$$

finally, we can rescale f to fit into the unit ball. $\square$

Let $\mathcal{H}$ be a RKHS with kernel k, and let $f \in \mathcal{H}$. Recall that by the reproducing property, $f(x) = \langle k(\cdot, x), f \rangle$. Then by linearity of the inner product and the fact that $\phi(x)$ is integrable,

$$\mathbb{E}_{x \in P}[f(x)] = \mathbb{E}_{x \in P}[\langle k(\cdot, x), f \rangle] = \langle \mathbb{E}_{x \in P}[k(\cdot, x)], f \rangle .$$

Definition 1.3 (mean embedding)**.** *The mean embedding of a distribution P in an RKHS $\mathcal{H}$ with kernel k is $\mu_P := \mathbb{E}_{x \in P}[k(\cdot, x)]$.*

Note that similar to the reproducing property that gives $f(x) = \langle k(\cdot, x), f \rangle$, the mean embedding gives $\mathbb{E}_{x \in P}[f(x)] = \langle \mu_P, f \rangle$.

Definition 1.4 (sample mean embedding)**.** *The sample mean embedding of a sample $X = \{x_1, \ldots, x_n\}$ in an RKHS $\mathcal{H}$ with kernel k is $\mu_X := \frac{1}{n} \sum_{i=1}^n [k(\cdot, x_i)]$.*

2 Maximum Mean Discrepancy

We are looking to distibguish between P and Q. The optimization problem is

$$\sup_{f\in\mathcal{H},\|f\|\leq 1} \mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim P}[f(x)] = \sup_{f\in\mathcal{H},\|f\|\leq 1} \langle \mu_P - \mu_Q, f\rangle = \|\mu_P - \mu_Q\|_{\mathcal{H}}.$$

Definition 2.1 (MMD)**.** *The MMD between two distributions is the distance between their mean embeddings* $\mathrm{MMD}^2(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$.

Theorem 2.2. $\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}[k(x,x')] + \mathbb{E}_{y,y'\sim Q}[k(y,y')] - 2\mathbb{E}_{x,\sim P}\mathbb{E}_{y\sim Q}[k(x,y)]$.

Proof.

$$\begin{aligned}
\mathrm{MMD}^2(P,Q) &= \|\mu_P - \mu_Q\|^2_{\mathcal{H}} \\
&= \langle \mu_p, \mu_p\rangle + \langle \mu_q, \mu_q\rangle - 2\langle \mu_p, \mu_q\rangle \\
&= \mathbb{E}_{x\sim P}[\mu_P(x)] + \mathbb{E}_{y\sim Q}[\mu_Q(y)] - 2\mathbb{E}_{x\sim P}[\mu_Q(X)] \\
&= \mathbb{E}_{x\sim P}[\langle \mu_P, k(\cdot,x)\rangle] + \mathbb{E}_{y\sim Q}[\langle \mu_Q, k(\cdot,y)\rangle] - 2\mathbb{E}_{x\sim P}[\langle \mu_Q, k(\cdot,x)\rangle] \\
&= \mathbb{E}_{x,x'\sim P}[k(x,x')] + \mathbb{E}_{y,y'\sim Q}[k(y,y')] - 2\mathbb{E}_{x,\sim P}\mathbb{E}_{y\sim Q}[k(x,y)].
\end{aligned}$$

□

2.1 Empirical estimation of MMD

We can estimate $\mathbb{E}_{x,x'\sim P}[k(x,x')]$ by

$$\frac{1}{n(n-1)} \sum_{i,j=1,i\neq j}^{n} k(x_i, x_j).$$

This is an unbiased estimation (as average are unbiased estimator of expectations). This gives the sample MMD, defined as

$$\mathrm{MMD}(X,Y) = \frac{1}{n}(n-1) \sum_{i,j=1,i\neq j}^{n} k(x_i,x_j) + \frac{1}{m(m-1)} \sum_{i,j=1,i\neq j}^{m} k(y_i,y_j) - 2\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x_i,y_j).$$

We will now use a measure concentration result by Hoeffding to get a convergence rate for the empirical MMD:

Theorem 2.3 (Hoeffding)**.** *Let k be a kernel with $|k(x,x')| < r$, and let X be a samples of size m drawn from P. Then*

$$\Pr\left(\left|\mathbb{E}_{x,x'\sim P}\, k(x,x') - \frac{1}{m(m-1)}\sum_{i\neq j} k(x_i,x_j)\right| > \epsilon\right) \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right).$$

which gives

Corollary 2.4 (MMD convergence)**.** *Let X, Y be samples of size m drawn from P, Q, respectively. Then*

$$\Pr\left(|\mathrm{MMD}(P,Q,\mathcal{F}) - \mathrm{MMD}(X,Y)| > \epsilon\right) \leq 2\exp\left(-\frac{m\epsilon^2}{r^2}\right).$$

2.2 Applications

1. Generative models: MMD can be used as a differentiable loss term to encourage generated samples to be similar to training samples from a given distribution

2. Statistical hypothesis testing: use MMD as a test statistic. Null hypothesis: $P = Q$. The distribution under the null can be estimated using permutations (more on this later on in this course).

2.2.1 Hilbert-Schmidt Independence Criterion (HSIC) - MMD for independence

Let P_X, P_Y be marginal distributions of a joint distribution P_{XY} over $\mathcal{X} \times \mathcal{Y}$. Let $\mu_{P_{XY}}, \mu_{P_X}, \mu_{P_Y}$ be the corresponding mean embeddings.

Definition 2.5 (HSIC)**.**

$$\mathrm{HSIC}\,(P_{XY}, P_X P_Y) := \|\mu_{P_{XY}} - \mu_{P_X}\mu_{P_Y}\|^2.$$

Let $\mathcal{F}$ be a $RKHS$ on $\mathcal{X}$ with kernel k, and $\mathcal{G}$ be a $RKHS$ on $\mathcal{Y}$ with kernel l. Then

$$\begin{aligned}\mathrm{HSIC}\,(P_{XY}, P_X P_Y) &= \mathbb{E}_{(x,y),(x',y')\sim P_{XY}}[k(x,x')l(y,y')] + \mathbb{E}_{x,x'\sim P_X}[k(x,x')\,\mathbb{E}_{y,y'\sim P_Y}\,k(y,y') \\ &\quad - 2\,\mathbb{E}_{(x,y)\sim P_{XY}}\left[\mathbb{E}_{x\sim P_X}\,k(x,x')\,\mathbb{E}_{y\sim P_Y}\,l(y,y')\right]. \end{aligned} \tag{1}$$

HSIC can be used to design independence test, similar to the MMD usage in two sample test. In addition, it can be used as a differential objective function for disentanglement models.

Homework

1. Prove equation 1

2. Design an experiment to verify the empirical MMD convergence rate.