Maximum Mean Discrepancy

Uri Shaham

March 4, 2024

1 Mean Embedding

1.1 Two Sasmple tests

Given two samples, $x_1, \ldots, x_n \sim P$, $y_1, \ldots, y_m \sim Q$ we are interested in the question whether $P = Q$. In one dimension, we can try methods like Kolmogorov Smirnov[1] which estimates the densities and checks the difference. But this is problematic in high dimensions, due to the curse of dimensionality.

1.2 Mean Embedding

The idea: choose a function class $\mathcal{F}$ and 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 on $\mathcal{X}$, 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 $\mathcal{H}$ be a universal RKHS and $\mathcal{F}$ be a unit ball in it, i.e., $\mathcal{F} = \{f \in \mathcal{H} \,|\, \|f\| \leq 1\}$. 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

$$\begin{aligned}
\mathbb{E}_{x \in P}[f^*(x)] - \mathbb{E}_{x \in Q}[f^*(x)] &= \mathbb{E}_{x \in P}[f(x)] - \mathbb{E}_{x \in Q}[f(x)] + \mathbb{E}_{x \in P}[f^*(x) - f(x)] - \mathbb{E}_{x \in Q}[f^*(x) - f(x)] \\
&\geq \mathbb{E}_{x \in P}[f(x)] - \mathbb{E}_{x \in Q}[f(x)] - 2\|f - f^*\|_\infty \\
&> \epsilon - 2\frac{\epsilon}{2} \\
&= 0.
\end{aligned}$$

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

[1] https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

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

2 Maximum Mean Discrepancy

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

$$\sup_{f \in \mathcal{H}, \|f\| \leq 1} \mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim P}[f(x)] = \sup_{f \in \mathcal{H}, \|f\| \leq 1} \langle \mu_P - \mu_Q, f \rangle = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2.$$

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

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.

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

□

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 is an unbiased estimator of expectation). This gives the sample MMD, defined as

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

We will now use a measure concentration result by Hoeffding [2] to get a convergence rate for the empirical MMD:

[2] `https://en.wikipedia.org/wiki/Hoeffding%27s_inequality`

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

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

Remark 2.4. *For example, with RBF kernel we have $r = 1$.*

This, together with the union bound [3] gives

Corollary 2.5 (MMD convergence)**.** *Let X, Y be samples of sizes m_x, m_y respectively, drawn from P, Q. Then*

$$\begin{aligned}
&\Pr\left(\left|\mathrm{MMD}^2(P,Q,\mathcal{F}) - \mathrm{MMD}^2(X,Y)\right| > \epsilon\right) > \\
&\Pr\left(\left|\mathbb{E}_{x,x'\sim P}\, k(x,x') - \frac{1}{m_x(m_x-1)}\sum_{i\neq j} k(x_i,x_j)\right| > \frac{\epsilon}{3}\right) + \\
&\Pr\left(\left|\mathbb{E}_{y,y'\sim Q}\, k(x,x') - \frac{1}{m_y(m_y-1)}\sum_{i\neq j} k(y_i,y_j)\right| > \frac{\epsilon}{3}\right) + \\
&\Pr\left(\left|\mathbb{E}_{x\sim P,y\sim Q}\, k(x,y) - \frac{1}{m_x m_y}\sum_{i,j} k(x_i,y_j)\right| > \frac{\epsilon}{3}\right) + \\
&\leq 6\exp\left(-\frac{m\epsilon^2}{9r^2}\right).
\end{aligned} \tag{1}$$

In words, we have a convergence rate exponential in $m = \min\{m_x, m_y\}$, i.e., the larger the samples are, the (exponentially) closer is the empirical MMD to the true MMD.

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 2.6 (HSIC)**.**

$$\mathrm{HSIC}^2\left(P_{XY}, P_X, P_Y\right) := \mathrm{MMD}^2(P_{XY}, P_X\otimes P_Y)$$

[3] https://en.wikipedia.org/wiki/Boole%27s_inequality

Let $\mathcal{F}$ be a $RKHS$ of functions on $\mathcal{X}$ with kernel k, and $\mathcal{G}$ be a $RKHS$ of functions on $\mathcal{Y}$ with kernel l. We use as a kernel

$$\kappa((x,y),(x',y') = k(x,x')l(y,y').$$

Proposition 2.7. *Prove that κ is a kernel*

Proof. Exercise □

We get:

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

However, in empirical estimation of HSIC we ancounter an issue, as we typically have only samples (x_i, y_i) from P_{XY}, we don't have samples from $P_X \otimes P_Y$. To tackle this, we estimate $P_X \otimes P_Y$ using samples (x_i, y_j) with $i \neq j$.

HSIC can be used to design independence tests, 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 proposition 2.7

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