Distance covariance and friends

Xun Zheng
xzheng1@andrew.cmu.edu

1 Distance covariance and friends

- General strategy for measuring dependency between X and Y:
  1. Define a distance (semi-)metric between distributions
  2. Measure distance between P_{XY} and $P_X P_Y$

  Then it naturally follows that distance $= 0 \iff X \perp Y$.
- Mutual information is dependency measure induced by KL divergence.
- Energy distance is a distance between distributions, induced by a distance metric of random variables.

$$\mathrm{ED}(P, Q) = \mathbb{E}[-d(A, A')] - 2\mathbb{E}[-d(A, B)] + \mathbb{E}[-d(B, B')] \tag{1}$$

  where expectation is taken over $A, A' \sim P$ and $B, B' \sim Q$.

  We would like ED to be *positive definite*, *i.e.*, $\mathrm{ED} \geq 0$ and $\mathrm{ED} = 0 \iff P = Q$. However, not all choice of d induces positive definite ED. Those nice behaving d are said to have "strong negative type", which includes the simple Euclidean distance $d(A, B) = \|A - B\|$ used in the original paper.
- Distance covariance is the dependency measure induced by energy distance.

$$\begin{aligned} \mathrm{dCov}(X, Y) &= \mathrm{ED}(P_{XY}, P_X P_Y) && (2) \\ &= \mathbb{E}[d(X, X')d(Y, Y')] - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'} d(X, X') E_{Y'} d(Y, Y')] + \mathbb{E}[d(X, X')]\mathbb{E}[d(Y, Y')] && (3) \end{aligned}$$

- Maximum mean discrepancy (MMD) is another distance between distributions, induced by a Mercer kernel k.

$$\begin{aligned} \mathrm{MMD}(P, Q) &= \mathbb{E}[k(A, A')] - 2\mathbb{E}[k(A, B)] + \mathbb{E}[k(B, B')] && (4) \\ &= \|\mu_P - \mu_Q\|_{\mathcal{H}} && (5) \end{aligned}$$

  where $\mu_P = \int k(\cdot, x) \, \mathrm{d}P(x)$ is the mean embedding of P in $\mathcal{H}$.

  MMD is positive definite by construction.
- Hilbert-Schmidt independence criterion (HSIC) is the dependency measure induced by MMD.

$$\begin{aligned} \mathrm{HSIC}(X, Y) &= \mathrm{MMD}(P_{XY}, P_X P_Y) && (6) \\ &= \mathbb{E}[k(X, X')k(Y, Y')] - 2\mathbb{E}_{X,Y}[\mathbb{E}_{X'} k(X, X') \mathbb{E}_{Y'} k(Y, Y')] + \mathbb{E}[k(X, X')]\mathbb{E}[k(Y, Y')] && (7) \end{aligned}$$

- The overall picture:

function of two r.v.s	distance of two distributions	dependency of two r.v.s
	KL divergence	mutual information
distance metric	energy distance	distance covariance
Mercer kernel	MMD	HSIC

- Distance $\to$ kernel: Any semi-metric d induces a kernel via

$$k(x, y) = d(x, z) + d(z, y) - d(x, y) \tag{8}$$

  through an arbitrary fixed point z. Also,

$$k \text{ is positive definite kernel} \iff d \text{ is "negative type"} \tag{9}$$

$$k \text{ is characteristic } (k = 0 \iff x = y) \iff d \text{ is "strong negative type"} \tag{10}$$

- Kernel $\to$ distance: Any non-degenerate kernel k induces a semi-metric via

$$d(x, y) = k(x, x) - 2k(x, y) + k(y, y) \tag{11}$$

  Similar result hold.

- The above equivalence between kernel and distance hold for population statistics. However, more (translation invariance and bijectivity) has to be defined for sample equivalence.