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.

\[
\text{ED}(P; Q) = \mathbb{E}\left[d(A; A')d(B; B')\right]
\]

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

We would like ED to be positive definite, i.e., \(ED \geq 0\) and \(ED = 0 \iff P = Q\). 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.

\[
d\text{Cov}(X, Y) = \text{ED}(P_{XY}, P_X P_Y) \\
= \mathbb{E}[d(X, X')d(Y, Y')] - 2\mathbb{E}[d(X, X')d(Y, Y')] + \mathbb{E}[d(X, X')d(Y, Y')]
\]

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

\[
\text{MMD}(P, Q) = \mathbb{E}[k(A, A')] - 2\mathbb{E}[k(A, B)] + \mathbb{E}[k(B, B')] \\
= \|\mu_P - \mu_Q\|_H
\]

where \(\mu_P = \int k(\cdot, x) dP(x)\) is the mean embedding of \(P\) in \(H\).

MMD is positive definite by construction.

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

\[
\text{HSIC}(X, Y) = \text{MMD}(P_{XY}, P_X P_Y) \\
= \mathbb{E}[k(X, X')k(Y, Y')] - 2\mathbb{E}_{X,Y}[k(X, X')E_Yk(Y, Y')] + \mathbb{E}[k(X, X')E[Yk(Y, Y')]
\]

The overall picture:
<table>
<thead>
<tr>
<th>function of two r.v.s</th>
<th>distance of two distributions</th>
<th>dependency of two r.v.s</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance metric</td>
<td>KL divergence</td>
<td>mutual information</td>
</tr>
<tr>
<td>Mercer kernel</td>
<td>energy distance</td>
<td>distance covariance</td>
</tr>
<tr>
<td></td>
<td>MMD</td>
<td>HSIC</td>
</tr>
</tbody>
</table>

- **Distance → kernel**: Any semi-metric $d$ induces a kernel via
  \[ k(x, y) = d(x, z) + d(z, y) - d(x, y) \]  
  through an arbitrary fixed point $z$. Also,
  \[ k \text{ is positive definite kernel } \iff d \text{ is “negative type”} \]  
  \[ k \text{ is characteristic } (k = 0 \iff x = y) \iff d \text{ is “strong negative type”} \]

- **Kernel → distance**: Any non-degenerate kernel $k$ induces a semi-metric via
  \[ d(x, y) = k(x, x) - 2k(x, y) + k(y, y) \]
  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.