Least-Squares Two-Sample Test

Masashi Sugiyama[1], Taiji Suzuki[2], Yuta Itoh[1],
Takafumi Kanamori[3], Manabu Kimura[1]

[1]Tokyo Institute of Technology
[2]University of Tokyo [3]Nagoya University

Presented by
Wittawat Jitkrittum

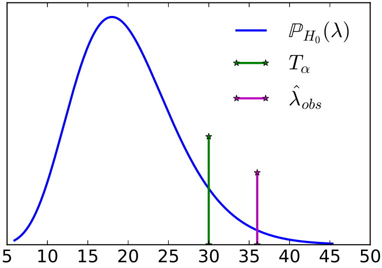
Gatsby tea talk
22 July 2016

Two-Sample Test

- $\mathsf{X} = \{\mathbf{x}_i\}_{i=1}^{m} \overset{i.i.d.}{\sim} P$ and $\mathsf{Y} = \{\mathbf{y}_i\}_{i=1}^{n} \overset{i.i.d.}{\sim} Q$.
- $\mathsf{X}, \mathsf{Y} \subset \mathbb{R}^d$. Unknown P, Q.
- Using X, Y, test hypotheses

$$H_0 : P = Q$$
$$H_1 : P \neq Q.$$

- Reject H_0 if test statistic $\hat{\lambda}_{obs} > T_\alpha$ (rejection threshold).
- $T_\alpha = (1-\alpha)$-quantile of the null distribution $\mathbb{P}_{H_0}(\lambda)$.
- Significance level α.



This Paper

Least-Squares Two-Sample Test
Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, Manabu Kimura
Neural Networks, 2011.

- Proposed to use the Pearson divergence as the test statistic

$$\mathrm{PE}(P, Q) := \frac{1}{2} \int \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1 \right)^2 q(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

- $\mathrm{PE}(P, Q) = 0 \iff P = Q$.
- Use a density ratio estimator to estimate $r(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$.
- **Advantage**: Can cross validate kernel parameters with the loss used by the density ratio estimator.
- **Disadvantage**: $O(m^3)$. Expensive.

Density Ratio Estimation

- Observe $\mathsf{X} = \{\mathbf{x}_i\}_{i=1}^m \overset{i.i.d.}{\sim} P$ and $\mathsf{Y} = \{\mathbf{y}_i\}_{i=1}^n \overset{i.i.d.}{\sim} Q$ in $\mathbb{R}^d$.
- **Goal**: Estimate the density ratio $r(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$.
- **Don't want**: Estimate $\hat{p}(\mathbf{x}), \hat{q}(\mathbf{x})$ separately. Ratio $\hat{p}(\mathbf{x})/\hat{q}(\mathbf{x})$. Inefficient.
- Many approaches to estimate $r(\mathbf{x})$ in one shot.
- We will use unconstrained least-squares importance fitting (uLSIF) [Kanamori et al., 2009].

uLSIF: A Density Ratio Estimator

- Estimate $r(\mathbf{x})$ with a linear model

$$
\begin{aligned}
\hat{r}(\mathbf{x}) &= \alpha_0 + \sum_{i=1}^{m} \alpha_i K(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\alpha}^\top \mathbf{k}(\mathbf{x}), \\
\boldsymbol{\alpha} &:= (\alpha_0, \ldots, \alpha_m)^\top \in \mathbb{R}^{m+1}, \\
\mathbf{k}(\mathbf{x}) &:= (1, K(\mathbf{x}, \mathbf{x}_1), \ldots, K(\mathbf{x}, \mathbf{x}_m))^\top \in \mathbb{R}^{m+1}, \\
K(\mathbf{x}, \mathbf{x}') &= \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2\right).
\end{aligned}
$$

- Find $\boldsymbol{\alpha}$ to minimize a squared-loss

$$
J(\boldsymbol{\alpha}) := \frac{1}{2} \int \left(\hat{r}(\mathbf{x}) - r(\mathbf{x})\right)^2 q(\mathbf{x})\, \mathrm{d}\mathbf{x}.
$$

Analytic Solution

$$J(\boldsymbol{\alpha}) := \frac{1}{2}\int \left(\hat{r}(\mathbf{x}) - r(\mathbf{x})\right)^2 q(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$= \frac{1}{2}\int \hat{r}(\mathbf{x})^2 q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \int \hat{r}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \overbrace{\frac{1}{2}\int r(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}}^{\text{constant}}$$

$$= \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{H}\boldsymbol{\alpha} - \mathbf{h}_p^\top \boldsymbol{\alpha} + \text{constant},$$

where $\mathbf{H} := \int \mathbf{k}(\mathbf{y})\mathbf{k}(\mathbf{y})^\top q(\mathbf{y})\,\mathrm{d}\mathbf{y} \in \mathbb{R}^{(m+1)\times(m+1)}$, and $\mathbf{h}_p := \int \mathbf{k}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

- With a regularization term

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^\top \hat{\mathbf{H}}\boldsymbol{\alpha} - \hat{\mathbf{h}}_p^\top \boldsymbol{\alpha} + \frac{\lambda}{2}\boldsymbol{\alpha}^\top \boldsymbol{\alpha} = \left(\hat{\mathbf{H}} + \lambda\mathbf{I}\right)^{-1} \hat{\mathbf{h}}_p.$$

- Unconstrained. $\hat{\boldsymbol{\alpha}}$ may contain negative entries.
- Can use $J(\hat{\boldsymbol{\alpha}})$ to select Gaussian width σ by cross validation.

Pearson Divergence as the Test Statistic

$$
\begin{aligned}
\mathrm{PE}(P, Q) &:= \frac{1}{2} \int \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} - 1 \right)^2 q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&= \frac{1}{2} \int r(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int r(\mathbf{y}) q(\mathbf{x}) \, \mathrm{d}\mathbf{y} + \frac{1}{2}. \\
\hat{\mathrm{PE}}(\mathsf{X}, \mathsf{Y}) &\approx \frac{1}{2m} \sum_{i=1}^{m} \hat{r}(\mathbf{x}_i) - \frac{1}{n} \sum_{i=1}^{n} \hat{r}(\mathbf{y}_i) + \frac{1}{2}
\end{aligned}
$$

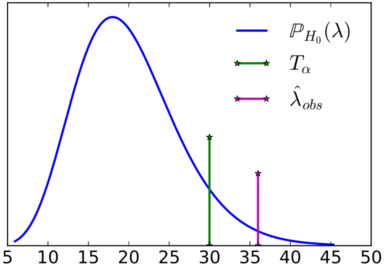
- Use the permutation test.

To approximate the null distribution,

1. Randomly divide $\mathsf{X} \cup \mathsf{Y}$ into disjoint X' and Y'.
2. Compute $\hat{\mathrm{PE}}(\mathsf{X}', \mathsf{Y}')$.
3. Repeat to get a histogram of $\hat{\mathrm{PE}}(\mathsf{X}', \mathsf{Y}')$.

Permutation Test

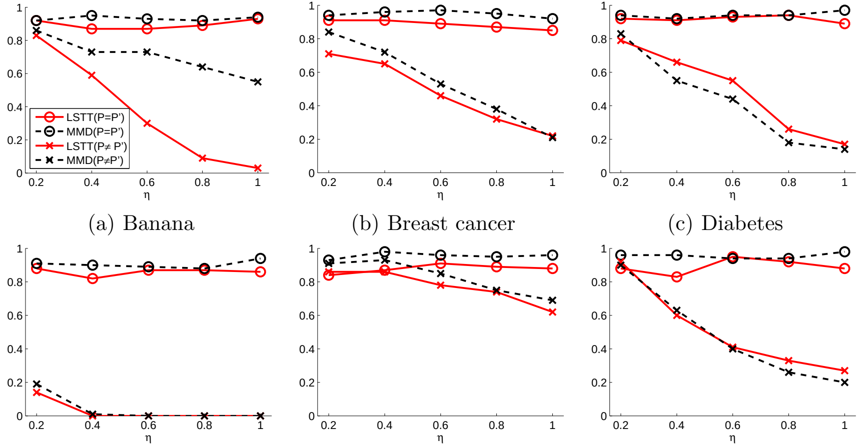
- $\lambda :=$ random variable representing the test statistic
- Use the permutation test, when $F(t) := \mathbb{P}_{H_0}(\lambda < t)$ is unknown.



$$
\begin{aligned}
&\text{reject } H_0 \text{ if } \hat{\lambda}_{obs} > T_\alpha \\
\iff & F(\hat{\lambda}_{obs}) > F(T_\alpha) \\
\iff & 1 - F(\hat{\lambda}_{obs}) < 1 - F(T_\alpha) \\
\iff & \text{p-value} < \alpha.
\end{aligned}
$$

- p-value $= \mathbb{P}_{H_0}(\lambda > \hat{\lambda}_{obs}) = \mathbb{E}_{\lambda \sim \mathbb{P}_{H_0}} I(\lambda > \hat{\lambda}_{obs}) \approx \frac{1}{B} \sum_{i=1}^{B} I(\hat{\lambda}_i > \hat{\lambda}_{obs})$.
- Assume samples X, Y have the same size.
- $\mathsf{X}', \mathsf{Y}' \leftarrow$ permute$(\mathsf{X} \cup \mathsf{Y})$. We expect $\mathsf{X}', \mathsf{Y}' \sim 0.5p(\mathbf{x}) + 0.5q(\mathbf{x})$.
- As X', Y' have the same distribution (H_0 is true), $\hat{\lambda}_i = \hat{\lambda}(\mathsf{X}', \mathsf{Y}') \sim \mathbb{P}_{H_0}$.

Experiments on Binary Classification Data



(a) Banana (b) Breast cancer (c) Diabetes

(d) Flare solar (e) German (f) Heart

- $P =$ data from $+$ class. $Q =$ data from $-$ class.
- Mix both classes to simulate $P = Q$ case.
- Report $\mathbb{P}(\text{not reject } H_0)$. Set $\alpha = 0.05$.
- $\eta =$ proportion of the sample size. Each problem has a different full size.

Questions?

Thank you

References I

Kanamori, T., Hido, S., and Sugiyama, M. (2009).
A least-squares approach to direct importance estimation.
J. Mach. Learn. Res., 10:1391–1445.