カーネル法による共部分構造の教師なし学習

持橋大地　　数理・推論研究系 准教授　　*daichi@ism.ac.jp*

本研究は, 横井祥氏 (東北大学 D1) との共同研究です。

Associative Knowledge

Recognizing "common sense" for natural language processing

cold front passes ⟶ begin to rain
dine with a friend ⟶ have a happy time
take medicine ⟶ recover from a cold

- Computers do not know these common sense or world knowledge.
- World knowledge is essential for everyday computing (e.g. robotics, nursery)
- Crucial also for artificial intelligence in general
  - Causal inference
  - Market basket analysis
  - Computational social science
  - Medicine, Pharmacy, · · ·
    men, 30 years old, night → beer, magazine, peanuts
    women, short sleep, anxiety → breast cancer

Problem: Granularity of Knowledge

What information should be included as a knowledge?

cold front passes yesterday →
It began to rain heavily in East Japan.
Jim had a dinner with his close friend →
He had a happy time yesterday.

- We don't know necessary knowledge in advance.

Heuristics employed so far

Hand-written rules to identify the range of information.

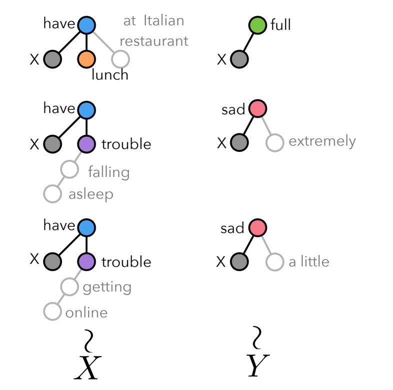
1. Subject + Verb
   - cold front passes → it begins
   - Jim had → he had
2. Verb + Object
   - passes → rain
   - had a dinner → had a time

⇓
Cannot be predicted from syntax!

Statistically: problem of generalization.

Extracting Co-Substructures

Associative knowledge should be dependent each other.



× Pearson correlation (must be in ℝ)
× Spearman's rank correlation (no natural order)
○ Mutual information
△ Canonical correlation analysis (must be linear)

Mathematically..

Given a set of item pairs

$$\mathcal{D} = \{\langle \mathbf{x}_n, \mathbf{y}_n \rangle\}_{n=1}^N \quad \mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y} \tag{1}$$

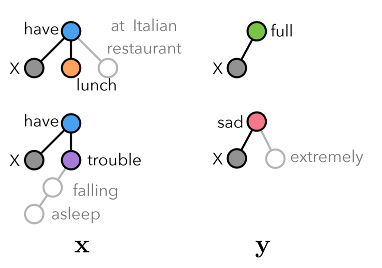
Find the pairs of substructures

$$S = \{\langle \mathbf{x}'_n, \mathbf{y}'_n \rangle\}_{n=1}^N \quad \mathbf{x}'_n \subset \mathbf{x}_n, \mathbf{y}'_n \subset \mathbf{y}_n \tag{2}$$

that maximize dependence to be defined; specifically, we assume

$$S \sim P_{XY} \tag{3}$$

and find S that maximizes $P_{XY} \,||\, P_X P_Y$.



Vanilla Mutual Information?

Assume $\mathbf{x}' = (v_1, v_2, \cdots, v_L)$, $\mathbf{y}' = (w_1, w_2, \cdots, w_M)$. Then

$$I(\mathbf{x}', \mathbf{y}') = \sum_{i=1}^{L} \sum_{j=1}^{M} p(v_i, w_j) \log \frac{p(v_i, w_j)}{p(v_i)p(w_j)} \tag{4}$$

$$= D(P_{XY} \| P_X P_Y). \tag{5}$$

However,

1. $p(v, w)$ is extremely sparse!
2. Nonlinear relationship between words? (eg. dependency)
3. Too big search space for I.

Our objective: HSIC

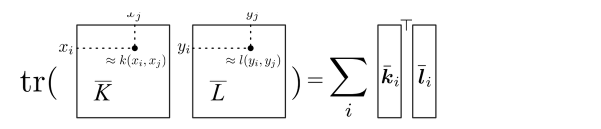
HSIC: Hilbert-Schmidt Independence Criterion (Gretton+ 2005)
Measuring independence with a kernel method.

$$\mathrm{HSIC}(S|\mathcal{D}) = \frac{1}{N^2}\mathrm{tr}(\mathbf{KHLH}) = \frac{1}{N^2}\mathrm{tr}(\bar{\mathbf{K}}\bar{\mathbf{L}}) \tag{6}$$

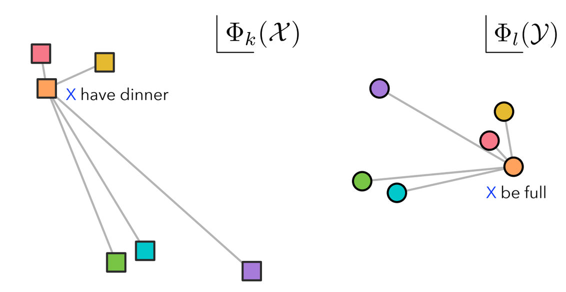
- $\mathbf{K} = (K_{ij})$: Gram matrix on $\mathbf{x}' \in S$
- $\mathbf{L} = (L_{ij})$: Gram matrix on $\mathbf{y}' \in S$
- $H_{ij} = \delta(i,j) - \frac{1}{N}$
- $\bar{\mathbf{K}} = \mathbf{HKH}$, $\bar{\mathbf{L}} = \mathbf{HLH}$

Intuitive explanation of HSIC

Empirical estimator of HSIC:



Large HSIC coincide with that "relative placements among $\mathcal{X}$ and among $\mathcal{Y}$ will correspond each other" in the projected space Φ.



Advantages of HSIC

- Nonparametric and nonlinear relationship of $\mathbf{x} \to \mathbf{y}$
  - eat in a restaurant → pay
  - eat at late hours → get fat
- Computed only through the kernels among $\mathcal{X}$ and among $\mathcal{Y}$
- Tree kernels, HMM (marginalized) kernels, string kernels, · · ·

HSIC and Mutual information

Remember that mutual information is a sum of pairwise mutual information (PMI).

$$\mathrm{PMI}(\mathbf{x}, \mathbf{y}) = \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \tag{7}$$

$$I(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \tag{8}$$

$$= \sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{PMI}(\mathbf{x}, \mathbf{y}). \tag{9}$$

HSIC and Mutual information (2)

"Kernelized PMI" is an element of HSIC.

$$f(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{N} \bar{k}(\mathbf{x}, \mathbf{x}_n) \bar{k}(\mathbf{y}, \mathbf{y}_n) \tag{10}$$

$$= \begin{pmatrix} \bar{k}(\mathbf{x}, \mathbf{x}_1) \\ \bar{k}(\mathbf{x}, \mathbf{x}_2) \\ \vdots \\ \bar{k}(\mathbf{x}, \mathbf{x}_N) \end{pmatrix} \cdot \begin{pmatrix} \bar{k}(\mathbf{y}, \mathbf{y}_1) \\ \bar{k}(\mathbf{y}, \mathbf{y}_2) \\ \vdots \\ \bar{k}(\mathbf{y}, \mathbf{y}_N) \end{pmatrix} \tag{11}$$

Then,

$$\mathrm{HSIC}(X, Y) = \sum_{\mathbf{x}} \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}). \tag{12}$$

PMI ↔ MI
|||
kPMI ↔ HSIC.

Optimization problem

Given

$$\mathcal{D} = \{\langle \mathbf{x}_n, \mathbf{y}_n \rangle\}_{n=1}^N \tag{13}$$

Find co-substructures S that maximize

$$\mathrm{HSIC}(S|\mathcal{D}) = \mathrm{tr}(\bar{\mathbf{K}}\bar{\mathbf{L}}) \tag{14}$$

where

$$\mathbf{K} = \text{Gram matrix on } \mathbf{x}' \in S \tag{15}$$

$$\mathbf{L} = \text{Gram matrix on } \mathbf{y}' \in S \tag{16}$$

- Note: this is a statistical "pruning" problem.

From a Bayesian point of view

Each word $w_i \in \mathbf{x}$ has latent binary variable z_i of inclusion (1) or exclusion (0) from knowledge:

$$p(\mathcal{D}) = \sum_{Z} p(\mathcal{D}, Z) \tag{17}$$

$$= \sum_{Z} \underbrace{p(\mathcal{D}|Z)}_{\text{HSIC}} p(Z) \tag{18}$$

We define a Gibbs distribution:

$$p(\mathcal{D}|Z) \propto \exp(\beta \cdot \mathrm{HSIC}(S|\mathcal{D})) \tag{19}$$

where $\beta \in \mathbb{R}$ is an inverse temparature.

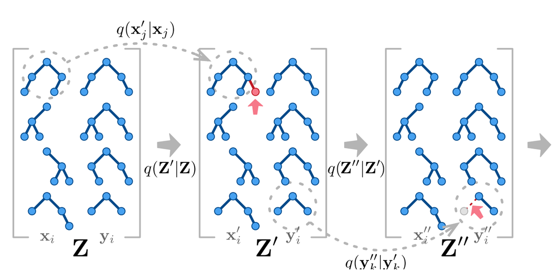
MCMC Inference algorithm

Until (convergence) {
For randomly visit $n \in 1 \cdots N$, do
— Draw a new candidate $S' \sim q(S'|S)$
— MH: accept S' with probability $\min(1, r)$ where

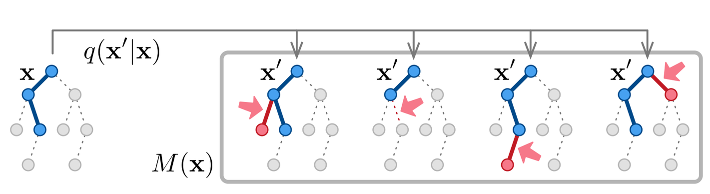
$$r = \frac{p(S'|\mathcal{D})}{p(S|\mathcal{D})} \cdot \frac{q(S|S')}{q(S'|S)}$$

$$= \exp(\beta(\mathrm{HSIC}(S'|\mathcal{D}) - \mathrm{HSIC}(S|\mathcal{D}))) \cdot \frac{q(\mathbf{x}_n|\mathbf{x}'_n)}{q(\mathbf{x}'_n|\mathbf{x}_n)}$$

}



Generating a MH candidate



- Given a parse tree of a sentence,
- Randomly select a word to expand / shrink a subtree from the original tree
- Assume that substructure is connected.

Fast computation

- Re-compute gram matrix $\bar{\mathbf{K}}$ and $\bar{\mathbf{L}}$ for MH step
  ⇒ Incremental re-computation of $\bar{\mathbf{K}}$ and $\bar{\mathbf{L}}$
- Rank-κ incomplete Cholesky decomposition and its update online

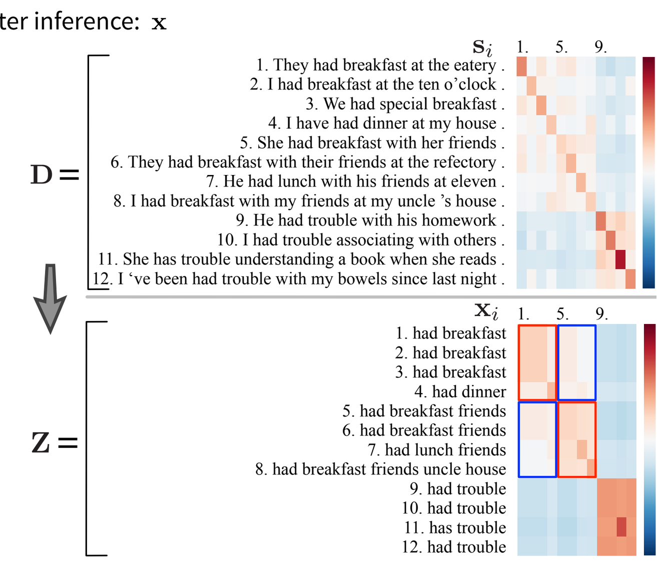
Experiments: Synthetic data

x	y
They had breakfast at the eatery	They are full now
I had breakfast at the ten o'clock	I'm full already
She had breakfast with her friends	She felt very happy
They had breakfast with their friends at the refectory	They felt happy
He had trouble with his homework	He cried in despair
· · ·	· · ·

We want to extract meaningful part from each sentence automatically.

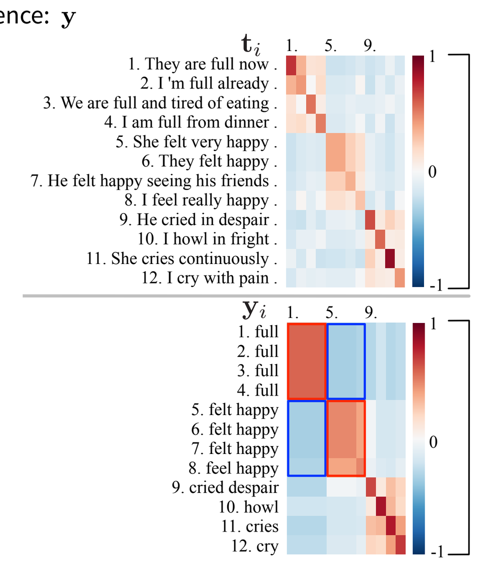
Synthetic data (2)

After inference: x



Synthetic data (3)

After inference: y



Synthetic data (4)

Our HSIC inference could extract important parts (non-gray) statistically!

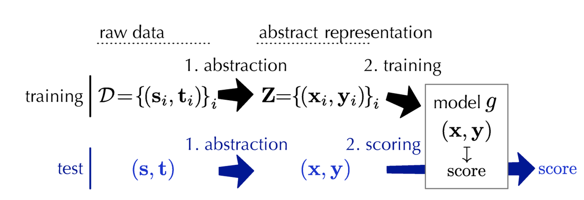
x	y
They had breakfast at the eatery	They are full now
I had breakfast at the ten o'clock	I'm full already
She had breakfast with her friends	She felt very happy
They had breakfast with their friends at the refectory	They felt happy
He had trouble with his homework	He cried in despair
· · ·	· · ·

Experiments: Actual corpora (1)

We extracted pairs of sentences that share co-referring arguments (like "she", "it") from Gigaword corpus (LDC2003T05): 17,781 documents from New York Times

- Create dependency trees to be pruned
- Training: 10,000 pairs for Gigaword, 1,000 pairs for Fairly Tale
- Testing: 500 pairs for Gigaword, 100 pairs for Fairly Tale

Prediction task:
discriminate correct sentence pair from randomly generated incorrect sentence pair.



Method

1. Learn associative substructures S from the training sentence pairs.
2. Based on these substructures, see if it correctly discriminates associative sentence pair (test data):

$$\frac{1}{|T_P|}\frac{1}{|T_N|} \sum_{\langle \mathbf{x}, \mathbf{y} \rangle \in T_P} \sum_{\langle \mathbf{x}', \mathbf{y}' \rangle \in T_N} \mathbb{I}[f(\mathbf{x}, \mathbf{y}) > f(\mathbf{x}', \mathbf{y}')] \tag{20}$$

where T_P is a set of positive pairs (= test data), and T_N is a set of negative pairs (= randomly created from training data).

$f(\mathbf{x}, \mathbf{y})$ is a measure for association (next).

Measure of association

For sentences $\mathbf{x}$ and $\mathbf{y}$, we measure association between them as

Baseline Pairwise Mutual Information (Chambers& Jurafsky 2008):

$$f(\mathbf{x}, \mathbf{y}) = \log \frac{N \cdot c(\mathbf{x}, \mathbf{y})}{c(\mathbf{x})c(\mathbf{y})} \tag{21}$$

where $c(\mathbf{x}, \mathbf{y})$ and $c(\mathbf{x})$ is a simple frequency.

Kernelized PMI Kernel estimate of PMI, where

$$f(\mathbf{x}, \mathbf{y}) = \sum_{n=1}^{N} \bar{k}(\mathbf{x}, \mathbf{x}_n) \bar{k}(\mathbf{y}, \mathbf{y}_n) \tag{22}$$

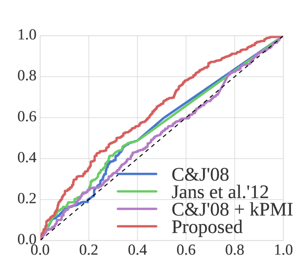
$\bar{k}$ is a centered kernel:

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \frac{1}{N}\sum_n k(\mathbf{x}, \mathbf{x}_n) - \frac{1}{N}\sum_n k(\mathbf{x}_n, \mathbf{x}') + \frac{1}{N^2}\sum_n \sum_m k(\mathbf{x}_n, \mathbf{x}_m). \tag{23}$$

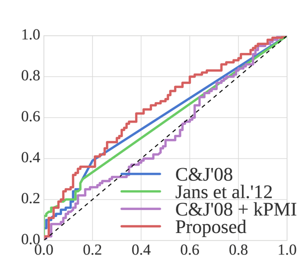
ROC curve

Precision/Recall curve: area under the curve (AUC) is a measure of performance.

- Gigaword corpus
- Fairly Tale corpus (Jans+ 2012): small collection of stories for children, 437 stories



Gigaword



Fairly Tale

Conclusion

Unsupervised learning of related substructures from paired data.
Beneficial for natural language processing, causal inference, medical diagnosis or digital marketing.

- Optimizes HSIC (Gretton+ 2005) of extracted substructures
- Combinatorial optimization: currently with MCMC
- Future work: scalarbility and more complicated kernels.

References:
"*Learning Co-Substructures by Kernel Dependence Maximization*". Sho Yokoi, Daichi Mochihashi, Ryo Takahashi, Naoaki Okazaki, Kentaro Inui. IJCAI 2017, to appear.

発表論文:

"Learning Co-Substructures by Kernel Dependence Maximization". Sho Yokoi, Daichi Mochihashi, Ryo Takahashi, Naoaki Okazaki, Kentaro Inui. IJCAI 2017, *to appear.*