

カーネル法に基づく共起尺度

※ 学生優秀プレゼンテーション賞対象

D2-46

横井 祥^{1,2} 福水 健次³ 小林 颯介⁴ 乾 健太郎^{1,2} ¹東北大 ²理研AIP ³統数研 ⁴PFN  yokoi@ecei.tohoku.ac.jp

入力	$\mathcal{Z} = \{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$	ペアで観測されるデータ
学習したい尺度	$f: (x, y) \mapsto D[p(x, y) \ p(x)p(y)]$	データ (x, y) の 関連の強さ を測りたい

$p(x, y)$ が $p(x)p(y)$ に対してどの程度大きいかを適当な尺度で測りたい

疎な対象（文や句）でも高速に計算できる**共起・関連の強さの尺度**の提案

1 自己相互情報量（PMI）

NLPで典型的に用いられる尺度

相互情報量 $I(X, Y) = \text{KL}[P_{XY} \| P_X P_Y]$

$$= \mathbf{E}_{x, y} \left[\log \frac{p(x, y)}{p(x)p(y)} \right]$$

自己相互情報量 $\text{PMI}(x, y; X, Y) = \log \frac{p(x, y)}{p(x)p(y)}$

2 PMIの推定

単語間の共起の強さを測る
コロケーション抽出 [Church&Hanks'90]
強く共起する単語ペア(e.g., “New York”)を抽出したい.
 $p(x, y)$ 上位には “is the” などが出てきてしまう.

$$\widehat{\text{PMI}}(x, y; \mathcal{Z}) = \log \frac{n \cdot c(x, y)}{c(x, \cdot) c(\cdot, y)}$$

各 p を頻度で推定

😊 **学習が軽い**
カウントするだけ

😞 **データの疎性に弱い**
文や句には使えない

文間の関連の強さを測る

対話の応答文生成 [Li+’16]
 $p(y|x)$ 上位には “I don’t know.” のような
「ありきたりでつまらない」応答が出てきてしまう.

$$\widehat{\text{PMI}}(x, y; \mathcal{Z}) = \log \frac{\hat{p}(y|x)}{\hat{p}(y)}$$

各言語モデルをRNNで学習

😞 **学習が重い**
 y を生成して誤差逆伝搬, ハイパーパラメータ調整

😊 **データの疎性に強い**
連続空間への埋込

4 Pointwise HSIC

提案手法

k, ℓ : 正定値カーネル; ϕ, ψ : 特徴関数
 $\mathcal{F}$: RKHS with $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}; \phi(x) := k(x, \cdot)$
 $\mathcal{G}$: RKHS with $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}; \psi(y) := \ell(y, \cdot)$

カーネル平均埋め込み
 $m_X := \mathbf{E}[\phi(X)] \in \mathcal{F}, m_Y := \mathbf{E}[\psi(Y)] \in \mathcal{G}$
 $m_{XY} := \mathbf{E}[\phi(X)\psi(Y)^\top] \in \mathcal{F} \otimes \mathcal{G}$

HSIC [Gretton+’05]

$$\text{HSIC}(X, Y; k, \ell) = \|m_{XY} - m_X m_Y^\top\|_{\text{HS}}^2$$
$$= \mathbf{E}_{x, y} [(\phi(x) - m_X)^\top C_{XY} (\psi(y) - m_Y)]$$
$$= \mathbf{E}_{x, y} [\mathbf{E}_{x', y'} [\tilde{k}(x, x') \tilde{\ell}(y, y')]]$$

$\tilde{k}: X$ で中心化した正定値カーネル
 $\tilde{k}(x, x') = k(x, x') - \mathbf{E}_{x'}[k(x, x')] - \mathbf{E}_x[k(x, x')] + \mathbf{E}_{x, x'}[k(x, x')]$

Pointwise HSIC

データ (x, y) の**相互共分散**への寄与

$$\text{PHSIC}(x, y; X, Y, k, \ell) = (\phi(x) - m_X)^\top C_{XY} (\psi(y) - m_Y)$$
$$= \mathbf{E}_{x', y'} [\tilde{k}(x, x') \tilde{\ell}(y, y')]$$

3 Pointwise HSIC の推定

$$\widehat{\text{PHSIC}}(x, y; \mathcal{Z}, k, \ell)$$

- $(x - \bar{x})^\top \hat{C}_{XY} (y - \bar{y})$
- $\frac{1}{n} (a - \frac{1}{n} A^\top \mathbf{1})^\top ((HA)^\top B) (b - \frac{1}{n} B^\top \mathbf{1})$

😊 **学習が軽い**

- 線形カーネルの場合は特徴空間で計算
- 非線形カーネルの場合はデータ空間で近似計算

いずれの場合も
学習 $\mathcal{O}(nd^2)$ モデルサイズ $\mathcal{O}(d^2)$ 予測 $\mathcal{O}(d^2)$
 n : サンプルサイズ, d : 次元数

😊 **データの疎性に強い**
充実した類似尺度を介したスムージング

Pointwise HSIC の特徴

お気持ち：類似度で**スムージングしたPMI**

- PMI**は**完全一致**で既知のデータと照合
- PHSIC**は**類似度関数**（カーネル）を用いて既知のデータと照合

	PMI	PHSIC
	$\log \frac{n \cdot \sum_i \mathbb{I}[x = x_i \wedge y = y_i]}{\sum_i \mathbb{I}[x = x_i] \sum_i \mathbb{I}[y = y_i]}$	$\frac{1}{n} \sum_i \tilde{k}(x, x_i) \tilde{\ell}(y, y_i)$
↑	$\mathcal{Z} = \{\dots, \underset{\parallel}{(\underset{\parallel}{x}, \underset{\parallel}{y})}, \dots\}$	$\{\dots, (\underset{\parallel}{x}, \underset{\parallel}{y_i}), \dots, (\underset{\parallel}{x_j}, \underset{\parallel}{y_j}), \dots\}$
↓	$\{\dots, (\underset{\parallel}{x_i}, \underset{\parallel}{y_i}), \dots, (\underset{\parallel}{x_j}, \underset{\parallel}{y_j}), \dots\}$	$\{\dots, (\underset{\parallel}{x_i}, \underset{\parallel}{y_i}), \dots, (\underset{\parallel}{x_j}, \underset{\parallel}{y_j}), \dots\}$

既存の**類似度関数**をそのまま活用できる

- 分布仮説に基づく**単語・文の分散表現**を再学習抜きで利用可能
 - 単語ベクトルの和 + cos で高精度な類似度計算 [Tian+’17]
 - 文の表現学習法も多数 [Kiros+’15][Hill+’16]
- SVM時代に作られた多くの**構造カーネル**がそのまま利用可能
 - 文字列カーネル [Kate&Mooney’06], 木カーネル [Collins&Duffy’02]
- 複数の類似尺度を**組合せる**ことができる
 - 正定値カーネルの和や積も正定値カーネル

まとめ・議論

類似性（正定値カーネル）を用いた**関連・共起の尺度**の提案

- 高速**な推定（線形カーネル, 不完全コレスキー分解）
- データの疎性に強い**（類似度による緩和）

類似性（正定値カーネル）→ **関連・共起**

- 分布仮説に基づく**単語ベクトル学習**（共起 → 類似性）との接続？
- コールドスタート & ユーザアイテム行列が疎の場合の**内容ベースフィルタリング**に使える？

$\phi(x) = x \in \mathbb{R}^d$ $k(x, x') = x^\top x'$
※ 既存の分散表現間の cos（**線形カーネル**）で高精度の類似度計算が可能

	$\widehat{\text{HSIC}}(\mathcal{Z}; k, \ell)$	$\widehat{\text{PHSIC}}(x, y; \mathcal{Z}, k, \ell)$
RKHS	$\ \hat{C}_{XY}\ _{\text{HS}}^2$ $= \ \frac{1}{n} \sum_i \phi(x_i) \psi(y_i)^\top - \overline{\phi(x)} \overline{\psi(y)}^\top\ _{\text{HS}}^2$	$(\phi(x) - \overline{\phi(x)})^\top \hat{C}_{XY} (\psi(y) - \overline{\psi(y)})$
RKHS 線形カーネル	$\ \hat{C}_{XY}\ _{\text{F}}^2 = \ \frac{1}{n} \sum_i x_i y_i^\top - \bar{x} \bar{y}^\top\ _{\text{F}}^2$	① $(x - \bar{x})^\top \hat{C}_{XY} (y - \bar{y})$
カーネル	$\frac{1}{n^2} \text{tr}(\tilde{K} \tilde{L})$ $K = (k(x_i, x_j)) \in \mathbb{R}^{n \times n}$ $H = I - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \in \mathbb{R}^{n \times n}$ $\tilde{K} = HKH$	$\frac{1}{n} \sum_i \tilde{k}(x, x_i) \tilde{\ell}(y, y_i)$ $= \frac{1}{n} (k - \frac{1}{n} K \mathbf{1})^\top H (\ell - \frac{1}{n} L \mathbf{1})$
カーネル 非線形カーネル	$\frac{1}{n^2} \ (HA)^\top B\ _{\text{F}}^2$	② $\frac{1}{n} (a - \frac{1}{n} A^\top \mathbf{1})^\top ((HA)^\top B) (b - \frac{1}{n} B^\top \mathbf{1})$

$K \approx AA^\top, A \in \mathbb{R}^{n \times d}, d < n$ $k(x, x') \approx a^\top a'$
※ **非線形カーネル**の場合も
不完全コレスキー分解を介して近似計算が可能

$a_i = \frac{k(x, x^{(i)}) - \sum_{m=1}^{i-1} a_m A_{im}}{A_{ii}}$

予備実験

- タスク：Narrative Event Chain [Schank&Abelson’77, Chambers&Jurafsky’08]
2文が同一文書内に連続して登場するかを予測
- $\mathcal{Z}$ ：文対の集合 例：“Tom killed Nancy.”, “The police arrested Tom immediately.”
- 先行研究：**PMI(X kill, arrest X)**  登場人物の位置と述語動詞
- 提案手法：**PHSIC** (“Tom killed Nancy.”, “The police arrested Tom immediately.”)
カーネル： $k(x, x') = \cos(\sum \text{wordvecs}(x), \sum \text{wordvecs}(x'))$  文をまるまる利用できる
- コーパス：(a) Gigawordから1万文 (b) Fairy Taleから1000文
- 結果：**AUC-ROCが大幅上昇** →

	PMI	PHSIC
Gigaword	0.569	0.711
Fairy Tale	0.596	0.666