Векторные представления текстов и графов

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

13 октября 2020 • ШАД Яндекс

Содержание

Дистрибутивная гипотеза и виды семантической близости слов

«Смысл слова определяется множеством его контекстов»

- Words that occur in the same contexts tend to have similar meanings [Harris, 1954].
- You shall know a word by the company it keeps [Firth, 1957].

Синтагматическая близость слов:
сочетаемость слов в одном контексте
(здание–строитель, кран–вода, функция–точка)

Парадигматическая близость слов:
взаимозаменяемость слов в одном контексте
(здание–дом, кран–смеситель, функция–отображение)

Z.Harris. Distributional structure. 1954.
J.R.Firth. A synopsis of linguistic theory 1930-1955. Oxford, 1957.
P.Turney, P.Pantel. From frequency to meaning: vector space models of semantics. 2010.

Формализация дистрибутивной гипотезы

Дано: текст $(w_1 \dots w_n)$, состоящий из слов словаря W

Найти: векторные представления слов $v_w \in \mathbb{R}^d$, так, чтобы близкие по смыслу слова имели близкие векторы

Модель CBOW (continuous bag-of-words) для вероятности слова w_i по его контексту $C_i = (w_{i-k} \dots w_{i-1} w_{i+1} \dots w_{i+k})$:

$$p(w_i = w | C_i) \;=\; \underset{w \in W}{\mathrm{SoftMax}} \langle u_w, v^{-i} \rangle \;\equiv\; \underset{w \in W}{\mathrm{norm}} \big(\exp \langle u_w, v^{-i} \rangle \big),$$

$v^{-i} = \frac{1}{2k} \sum\limits_{w \in C_i} v_w$ — средний вектор слов из контекста C_i,

v_w — векторы предсказывающих слов,

u_w — вектор предсказываемого слова, в общем случае $u_w \neq v_w$.

Критерий максимума log-правдоподобия, $U, V \in \mathbb{R}^{|W| \times d}$:

$$\sum_{i=1}^{n} \log p(w_i | C_i) \to \max_{U,V}$$

Ещё одна формализация дистрибутивной гипотезы

Дано: текст $(w_1 \dots w_n)$, состоящий из слов словаря W

Найти: векторные представления слов $v_w \in \mathbb{R}^d$, так, чтобы близкие по смыслу слова имели близкие векторы

Модель Skip-gram для предсказания вероятности слов контекста $C_i = (w_{i-k} \dots w_{i-1} w_{i+1} \dots w_{i+k})$ по слову w_i:

$$p(w|w_i) = \underset{w \in W}{\mathrm{SoftMax}} \langle u_w, v_{w_i} \rangle \equiv \underset{w \in W}{\mathrm{norm}} \big(\exp \langle u_w, v_{w_i} \rangle \big),$$

v_w — вектор предсказывающего слова,

u_w — вектор предсказываемого слова, в общем случае $u_w \neq v_w$.

Критерий максимума log-правдоподобия, $U, V \in \mathbb{R}^{|W| \times d}$:

$$\sum_{i=1}^{n} \sum_{w \in C_i} \log p(w|w_i) \to \max_{U,V}$$

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.

Сравнение моделей CBOW и Skip-gram

- Различие — в структуре оптимизационного критерия:

$$\text{CBOW:} \quad \sum_{i=1}^{n} \log \underset{w_i \in W}{\text{SoftMax}} \left(\frac{1}{2k} \sum_{c \in C_i} \langle u_{w_i}, v_c \rangle \right) \to \max_{U,V}$$

$$\text{Skip-gram:} \quad \sum_{i=1}^{n} \sum_{c \in C_i} \log \underset{c \in W}{\text{SoftMax}} \langle u_c, v_{w_i} \rangle \to \max_{U,V}$$

- Skip-gram точнее моделирует вероятности редких слов
- Обе модели можно обучать с помощью SGD
- Оба критерия трудно оптимизировать из-за SoftMax
- Что делать? Заменять либо SoftMax, либо критерий

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.

Иерархический SoftMax

Идея: заменить SoftMax на другую функцию потерь, сложность вычисления которой $O(\log|W|)$ вместо $O(|W|)$.

Предварительный этап:

- По словарю частот строится *бинарное дерево Хаффмана*
- Каждая внутренняя вершина n *хранит* вектор $u_n \in \mathbb{R}^d$
- Каждый лист *вычисляет* вектор v_w для слова $w \in \mathbb{R}^d$
- Модель переходов из внутренних вершин дерева:

$$\begin{aligned} \text{направо:} \quad & p(+1|n,w) = \sigma(\langle u_n, v_w\rangle) \\ \text{налево:} \quad & p(-1|n,w) = \sigma(-\langle u_n, v_w\rangle) = 1 - p(+1|n,w) \end{aligned}$$

Обучаются векторы u_n во внутренних вершинах дерева

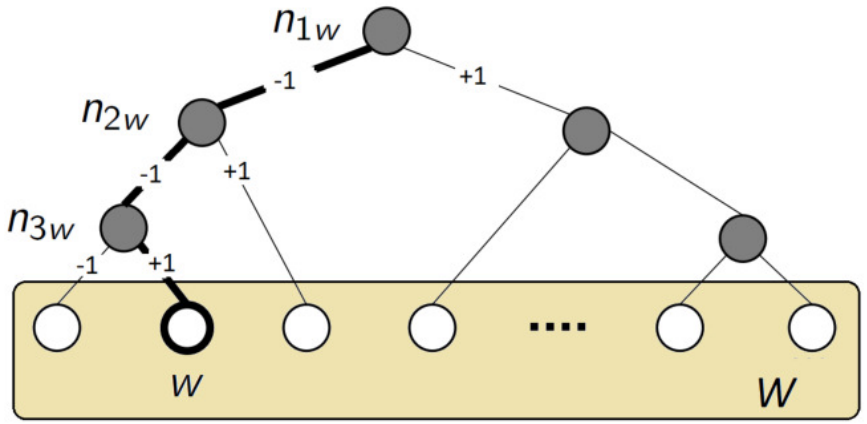
Иерархический SoftMax: обучение модели

Модель $p(w|w_i)$, гарантирующая нормировку $\sum_w p(w|w_i) = 1$:

$$p(w|w_i) = \prod_{j=1}^{\ell(w)} p\big(\beta_{jw}|n_{jw}, w_i\big) = \prod_{j=1}^{\ell(w)} \sigma\big(\beta_{jw}\langle u_{n_{jw}}, v_{w_i}\rangle\big)$$

где $\ell(w)$ — длина пути к листу w,
n_{jw} — j-я внутренняя вершина на пути к листу w,
$\beta_{jw} \in \{-1, +1\}$ — поворот из j-й вершины на пути к w.

Пример: $p(w|w_i) = p(-1|n_{1w}, w_i)\, p(-1|n_{2w}, w_i)\, p(+1|n_{3w}, w_i)$



Подмена задачи: классификация пар слов на два класса

Критерий log-loss для SGNS (Skip-gram Negative Sampling):

$$\sum_{i=1}^{n} \sum_{w \in C_i} \Big(\log p(+1|w, w_i) + \log p(-1|\bar{w}, w_i) \Big) \to \max_{U,V}$$

где $p(y|w, w_i) = \sigma\big(y\langle u_w, v_{w_i}\rangle\big)$ — модель классификации, $y = \pm 1$;
$y = +1$, если пара слов (w, w_i) находится в общем контексте;
$y = -1$, если пара слов (w, w_i) не находится в общем контексте;
$\bar{w} \sim p(w)^{3/4}$ сэмплируется из $W \setminus C_i$ в методе SG.

Эвристики и прочие замечания:

- Dynamic window: случайный выбор $k \sim [3..10]$
- Итоговые векторы слов: $\alpha v_w + (1 - \alpha) u_w$
- Приём NS применяют, когда не хватает второго класса
- Что делать со словами, которые встречаются впервые?

Модель векторных представлений FastText

Идея: векторное представление слова w определяется как сумма векторов всех его буквенных n-грамм $G(w)$:

$$u_w = \sum_{g \in G(w)} u_g$$

В Skip-gram вместо векторов слов u_w обучаются векторы u_g

Пример: $G(\text{дармолюб}) = \{<\text{да}, \text{арм}, \text{рмо}, \text{мол}, \text{олю}, \text{люб}, \text{юб}>\}$

Преимущества:

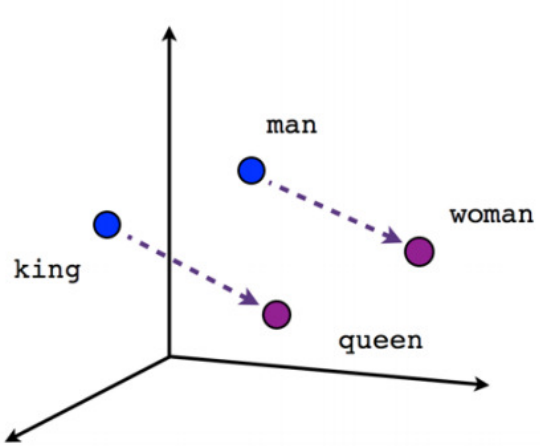
- Это решает проблемы новых слов и слов с опечатками
- Подходит для обработки текстов социальных медиа
- Словарь 2- и 3-грамм обычно меньше словаря W
- Существует много предобученных моделей

Bojanowski et al. Enriching word vectors with subword information. 2016.

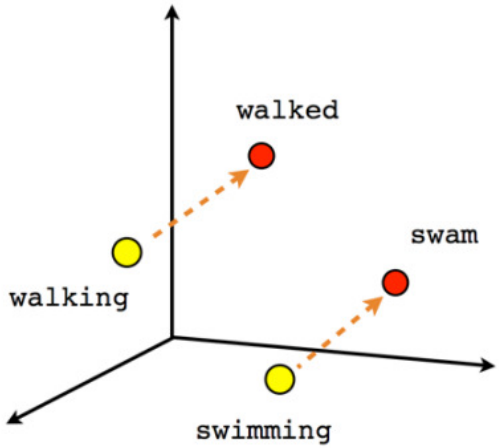
Проверка на задачах семантической близости и аналогии слов

Задача семантической близости слов:
по выборке пар слов (a, b) оценивается корреляция Спирмена между $\cos(v_a, v_b)$ и экспертными оценками близости $y(a, b)$

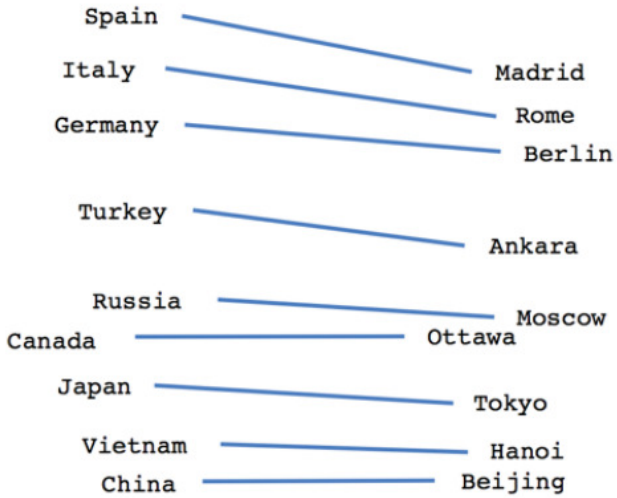
Задача семантической аналогии слов:
по трём словам угадать четвёртое



Male-Female



Verb tense



Country-Capital

Связь word2vec с матричными разложениями

d — размерность векторов слов v_w и u_w
$V = (v_w)_{W \times d}$ — матрица предсказывающих векторов слов
$U = (u_w)_{W \times d}$ — матрица предсказываемых векторов слов

SGNS строит матричное разложение $P \approx UV^{\mathrm{T}}$ матрицы Shifted PMI (Point-wise Mutual Information):

$$P_{ab} = \ln \frac{n_{ab} n}{n_a n_b} - \ln k,$$

n_{ab} — частота пары слов a, b в окне $\pm k$ слов,
n_a, n_b — число пар с участием слова a и b соответственно,
n — число всех пар слов в коллекции.

В качестве эвристики используют также Shifted Positive PMI:

$$P^+_{ab} = \left(\ln \frac{n_{ab} n}{n_a n_b} - \ln k \right)_+.$$

O.Levy, Y.Goldberg. Neural word embedding as implicit matrix factorization. 2014.

Модели векторных представлений для текстов и графов

word2vec: *эмбединги* (векторные представления) слов
T.Mikolov et al. Efficient estimation of word representations in vector space. 2013.

paragraph2vec: эмбединги фрагментов или документов
Q.Le, T.Mikolov. Distributed representations of sentences and documents. 2014.

sent2vec: эмбединги предложений
M.Pagliardini et al. Unsupervised learning of sentence embeddings using compositional n-gram features. 2017.

FastText: эмбединги символьных *n*-грамм
`https://github.com/facebookresearch/fastText`

node2vec: эмбединги вершин графа
A.Grover, J.Leskovec. Node2vec: scalable feature learning for networks. 2016.

graph2vec: более общие эмбединги на графах
A.Narayanan et al. Graph2vec: learning distributed representations of graphs. 2017.

StarSpace: эмбединги чего угодно от Facebook AI Research
L.Wu, A.Fisch, S.Chopra, K.Adams, A.B.J.Weston. StarSpace: embed all the things! 2018.

BERT: эмбединги фраз и предложений от Google AI Language
J.Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

GPT-3: эмбединги, предобученные по 570Gb текстов от OpenAI
T.B.Brown et al. Language Models are Few-Shot Learners. 2020.

Многомерное шкалирование (multidimensional scaling, MDS)

Дано: $(i,j) \in E$ — выборка рёбер графа $\langle V, E \rangle$,
R_{ij} — расстояния между вершинами ребра (i,j).
Например, в IsoMAP R_{ij} — длина кратчайшего пути по графу.

Найти: векторные представления вершин $z_i \in \mathbb{R}^d$, так, чтобы близкие (по графу) вершины имели близкие векторы.

Критерий стресса (stress):

$$\sum_{(i,j)\in E} w(R_{ij})\big(\rho(z_i,z_j) - R_{ij}\big)^2 \to \min_Z, \quad Z \in \mathbb{R}^{V\times d},$$

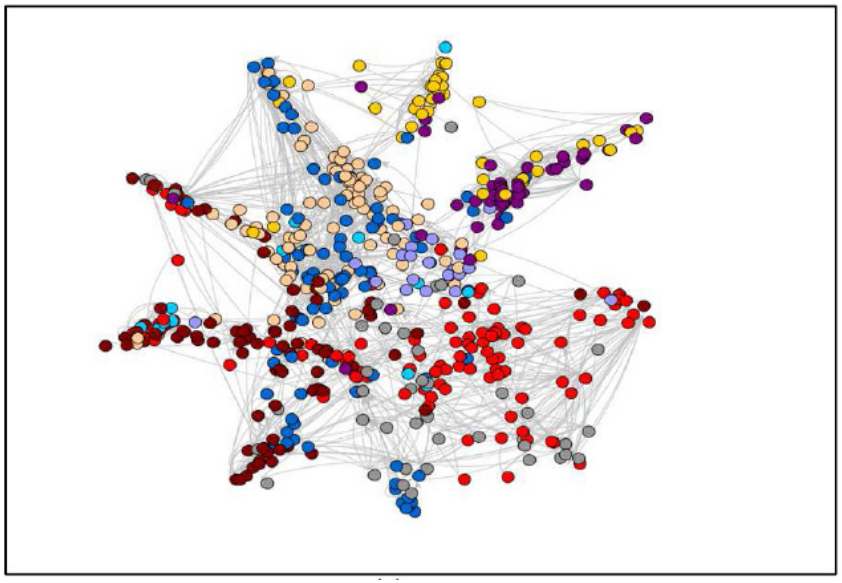
где $\rho(z_i,z_j) = \|z_i - z_j\|$ — обычно евклидово расстояние,
$w(R_{ij})$ — веса (какие расстояния важнее, большие или малые).

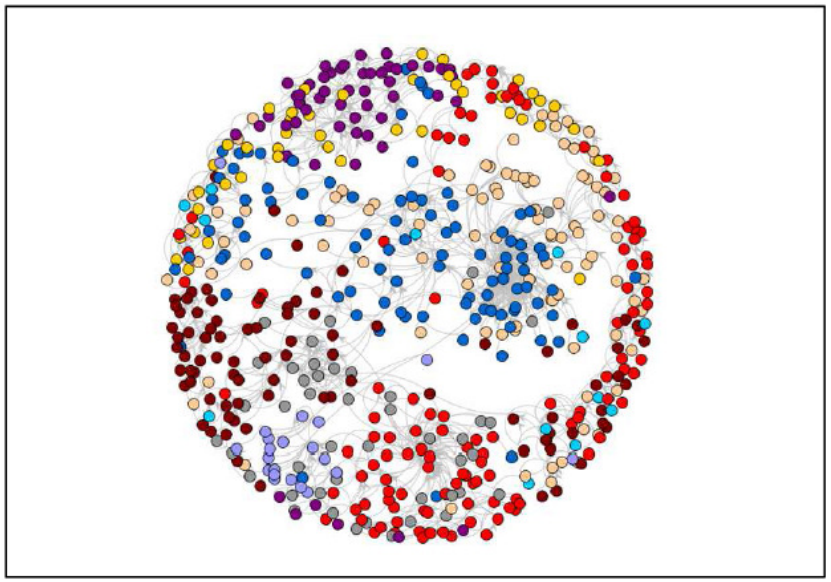
Обычно решается методом стохастического градиента (SG).

I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

Многомерное шкалирование для визуализации данных

При $d = 2$ осуществляется проекция выборки на плоскость





- Используется для визуализации кластерных структур
- Форму облака точек можно настраивать весами и метрикой
- Недостаток — искажения неизбежны
- Наиболее популярный метод для визуализации — t-SNE

Laurens van der Maaten, Geoffrey Hinton. Visualizing data using t-SNE. 2008

Матричные разложения (graph factorization)

Дано: $(i,j) \in E$ — выборка рёбер графа $\langle V, E \rangle$,
S_{ij} — близость между вершинами ребра (i,j).
Например, $S_{ij} = [(i,j) \in E]$ — матрица смежности вершин.

Найти: векторные представления вершин $z_i \in \mathbb{R}^d$, так, чтобы близкие (по графу) вершины имели близкие векторы.

Критерий для неориентированного графа (S симметрична):

$$\sum_{(i,j)\in E} \big(\langle z_i, z_j \rangle - S_{ij}\big)^2 \to \min_Z, \quad Z \in \mathbb{R}^{V\times d}$$

Критерий для ориентированного графа (S несимметрична):

$$\sum_{(i,j)\in E} \big(\langle \varphi_i, \theta_j \rangle - S_{ij}\big)^2 \to \min_{\Phi,\Theta}, \quad \Phi, \Theta \in \mathbb{R}^{V\times d}$$

Обычно решается методом стохастического градиента (SG).

I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

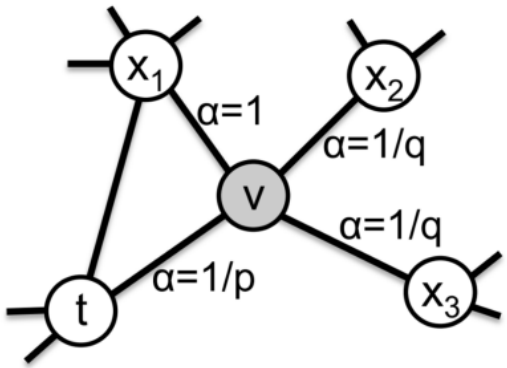
Модель случайных блужданий

Аналог модели Skip-gram (текст ведь тоже граф, но линейный):

$$\sum_{i\in V}\left(\sum_{j\in C_i}\log\sigma\big(\langle\varphi_i,\theta_j\rangle\big)+\sum_{j\in\bar{C}_i}\log\sigma\big(-\langle\varphi_i,\theta_j\rangle\big)\right)\to\max_{\Phi,\Theta},$$

C_i — окрестность («контекст») вершины i, сэмплируемая случайным блужданием длины k (DeepWalk, node2vec),
$\bar{C}_i$ — вершины, далёкие от i, сэмплируемые $j\sim p(j)^{3/4}$

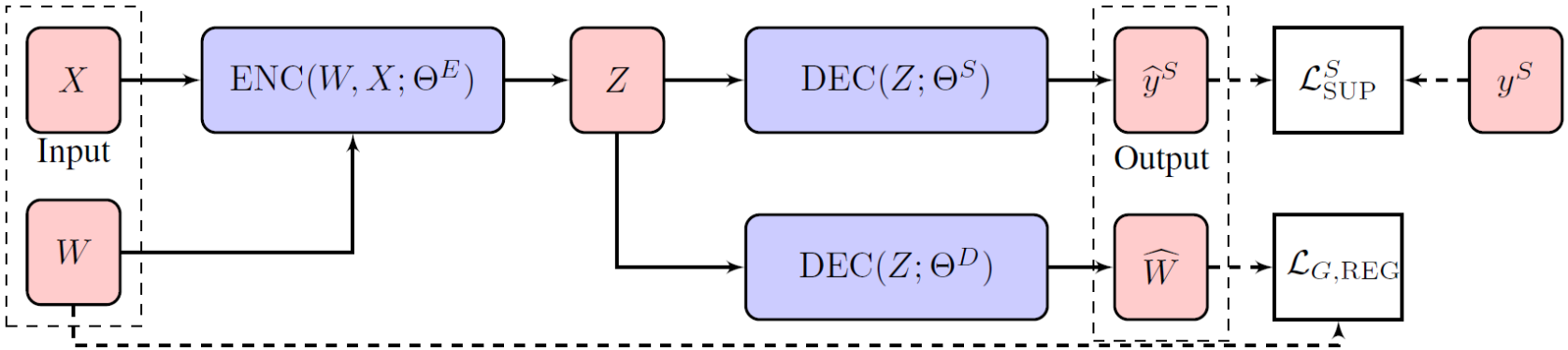
Параметризация случайных блужданий:
вероятность $p(v\to w)$ после перехода $t\to v$
$p\downarrow\ q\uparrow$ — ближе к поиску в ширину (BFS)
$p\uparrow\ q\downarrow$ — ближе к поиску в глубину (DFS)



B.Perozzi et al. DeepWalk: online learning of social representations. SIGKDD-2014.
A.Grover, J.Leskovec. node2vec: scalable feature learning for networks. SIGKDD-2016.

GraphEDM: обобщённый автокодировщик на графах

Graph Encoder Decoder Model — обобщает более 30 моделей:



$W \in \mathbb{R}^{V \times V}$ — входные данные о рёбрах
$X \in \mathbb{R}^{V \times n}$ — входные данные о вершинах, признаковые описания
$Z \in \mathbb{R}^{V \times d}$ — векторные представления вершин графа
$\mathrm{DEC}(Z; \Theta^D)$ — декодер, реконструирующий данные о рёбрах
$\mathrm{DEC}(Z; \Theta^S)$ — декодер, решающий supervised-задачу
y^S — (semi-)supervised данные о вершинах или рёбрах
$\mathcal{L}$ — функции потерь

I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

Примеры транзакционных данных

Коллекция текстов — двудольный граф, выборка пар (d, w)
Транзакционные данные — n-ки термов разных модальностей

Примеры:

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

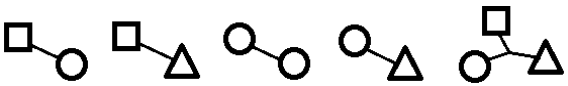
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

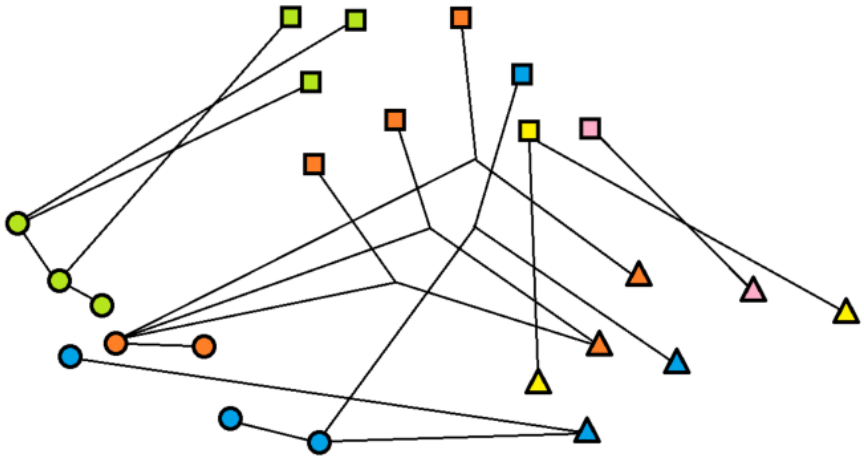
$\Gamma = \langle V, E \rangle$ — неориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



Исходные данные:

E_k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{kdx} — число вхождений ребра (d, x) в выборку E_k

Тематическая модель гиперграфа: основные предположения

- в ребре (d, x) подмножество $x \subset V$ может быть любым, независимо от типа ребра k
- первая *гипотеза условной независимости*: тематика контейнера $p(t|d)$ не зависит от типа ребра k
- вторая *гипотеза условной независимости*: распределение $p(v|t)$ термов v модальности V^m в теме t не зависит ни от контейнера d, ни от типа ребра k
- третья *гипотеза условной независимости*: термы $v \in x$ в ребре (d, x) не зависят друг от друга
- *гипотеза «мешка транзакций»*: выборка транзакций типа k порождается случайно и независимо из

$$p_k(d, x) = p(d) \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t)$$

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k

$$p_k(x|d) = \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt}$$

Задача максимизации взвешенной суммы log-правдоподобий по всем типам рёбер:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt} \to \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки:

$$\varphi_{vt} \geqslant 0, \quad \sum_{v \in V^m} \varphi_{vt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in E_k} n_{kdx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vt} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

E-шаг:
$$p_{tdx} = \operatorname*{norm}_{t\in T}\Big(\theta_{td} \prod_{v\in x} \varphi_{vt}\Big)$$

M-шаг:
$$\begin{cases} \varphi_{vt} = \operatorname*{norm}_{v\in V^m}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} \big[v\in x\big] n_{kdx} p_{tdx} + \varphi_{vt}\dfrac{\partial R}{\partial \varphi_{vt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx} p_{tdx} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t \in T} p(t|d) \prod_{w \in s} p(w|t)^{n_{sw}} = \sum_{t \in T} \theta_{td} \prod_{w \in s} \varphi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d \in D} \sum_{s \in S_d} \ln \sum_{t \in T} \theta_{td} \prod_{w \in s} \varphi_{wt}^{n_{sw}} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Гиперграфовые тематические модели языка

Что ещё может быть ребром гиперграфа?

Любое подмножество термов, связанных друг с другом по смыслу, и порождаемых одной общей темой.

- предложение
- синтагма, ветка синтаксического дерева
- именная группа
- факт «объект, субъект, действие»
- пары термов в соседних предложениях:
  два синонима, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст сообщения и его автор
- финансовая транзакция с текстом платёжного поручения

Анализ транзакций корпоративных клиентов банка

Данные:
лесная отрасль, 2016 г., 10.7M транзакций, 1M компаний.
Транзакция — это тройка $\langle$покупатель, продавец, текст$\rangle$.

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Семь модальностей:

- компании: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели
- ОКВЭДы данной компании
- ОКВЭДы контрагентов: поставщики / покупатели

Примеры тем — видов деятельности компаний

покупка	продажа
0.11: услуга	0.12: лдсп
0.07: классик	0.08: дсп
0.05: дрова	0.03: мдф
0.05: пиловочник	0.03: поставка
0.05: материал	0.02: услуга
0.03: порода	0.02: охранный
0.03: лесоматериал	0.02: ламинировать
0.03: сертум	0.02: хдф
0.02: хвойный	0.02: материал
0.01: дерево	0.01: накл
0.01: транспортный	0.01: товар

покупка	продажа
0.19: право	0.16: арендный
0.17: сбис	0.10: часть
0.16: использование	0.08: плата
0.03: аккаунт	0.04: минимальный
0.02: электронный	0.04: участок
0.02: лицевой	0.04: использование
0.02: устный	0.02: земля
0.01: устройство	0.02: лесничество
0.01: генерация	0.02: земельный
0.01: хранение	0.01: фонд
0.01: ключевой	0.01: федеральный

Примеры тем — видов деятельности компаний

покупка	продажа
0.09: ткань	0.16: мебель
0.09: поставка	0.05: плёнка
0.02: мебельный	0.04: стул
0.02: деревянный	0.03: кресло
0.02: транспортный	0.03: изделие
0.02: фанера	0.02: краска
0.02: поролон	0.02: фанера
0.01: механизм	0.01: лкм
0.01: плата	0.01: лакокрасочный
0.01: частичный	0.01: лак
	0.01: материал
	0.01: клеить

покупка	продажа
0.06: лдсп	0.37: товар
0.05: фурнитура	0.15: мебель
0.02: плёнка	0.04: поставка
0.02: материал	0.04: накладный
0.02: мебельный	0.03: накл
0.02: стекло	0.03: рубль
0.02: мдф	
0.02: кромка	
0.01: транспортный	
0.01: клеить	
0.01: профиль	
0.01: пвх	

Резюме

- Векторные представления (эмбединги) — мощнейший инструмент анализа сложно структурированных данных
- Основное применение — векторизация объектов для дальнейшей обработки методами машинного обучения
- Эмбединги на графах обобщают задачи векторного представления текстов, дискретных сигналов, изображений
- Тематическое моделирование — тоже эмбединги, но вероятностные (на симплексе) и интерпретируемые