Латентная семантическая модель для представления смыслов многозначных слов

Дмитрий Кондрашкин,
научный руководитель: к.ф.-м.н. Ветров Д. П.

26 февраля 2015 г.

Skip-gram model

- По слову w предсказывается слово из контекста v:

$$p(v \mid w, \boldsymbol{\theta}) = \frac{\exp\{\mathrm{In}_w^\intercal \mathrm{Out}_v\}}{\sum_{t=1}^{V} \exp\{\mathrm{In}_w^\intercal \mathrm{Out}_t\}},$$

  где $\boldsymbol{\theta} = \{\mathrm{In}_v, \mathrm{Out}_v\}_{v=1}^{V}$.
- Каждому слову v соответствуют “входные” и “выходные” представления $\mathrm{In}_v, \mathrm{Out}_v \in \mathbb{R}^D$.

(Mikolov 2013) Distributed representations of words and phrases and their compositionality.

Skip-gram model

- Входной текст $\mathbf{o} = \{o_1, \dots, o_N\}$.
- Обучающий объект $(x_i, \mathbf{y}_i)$:
  - $x_i = o_i$,
  - $\mathbf{y}_i = \{o_{i-C/2}, \dots, o_{i-1}, o_{i+1}, \dots, o_{i+C/2}\}$.
- Правдоподобие:
$$\prod_{i=1}^{N} \prod_{j=1}^{C} p(y_{ij} \mid x_i, \boldsymbol{\theta}).$$
- Максимизировать будем лог-правдоподобие
$$L(X, Y, \boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{C} \log p(y_{ij} \mid x_i, \boldsymbol{\theta}).$$

Стохастическая оптимизация

- Типичная задача в машинном обучении

$$L(X, Y, \boldsymbol{\theta}) = \sum_{i=1}^{N} l(x_i, y_i, \boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta}) \to \min_{\boldsymbol{\theta}}$$

- Стохастический градиентный спуск:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta_t g(\boldsymbol{\theta}^{(t)}),$$
$$\nabla L(X, Y, \boldsymbol{\theta}) \approx g(\boldsymbol{\theta}).$$

- Условия сходимости:

$$\mathbb{E}\, g(\boldsymbol{\theta}) = \nabla L(X, Y, \boldsymbol{\theta}),$$
$$\sum_t \eta_t = \infty, \sum_t \eta_t^2 < \infty.$$

- Как правило, $g(\boldsymbol{\theta}) = N \nabla l(x_i, y_i, \boldsymbol{\theta}) + \lambda \nabla R(\boldsymbol{\theta})$.

Стохастическая оптимизация

- В нашем случае:

$$L(X, Y, \boldsymbol{\theta}) = \sum_{i=1}^{N} \underbrace{\sum_{j=1}^{C} \log p(y_{ij} \mid x_i, \boldsymbol{\theta})}_{l(x_i, \mathbf{y}_i, \boldsymbol{\theta})} \to \max_{\boldsymbol{\theta}} .$$

- $\nabla L(X, Y, \boldsymbol{\theta}) \approx N \sum_{j=1}^{C} \nabla \log p(y_{ij} \mid x_i, \boldsymbol{\theta})$.

Hierarchical softmax

$$p(v \mid w, \boldsymbol{\theta}) = \frac{\exp\{\mathrm{In}_w^\intercal \, \mathrm{Out}_v\}}{\sum_{t=1}^{V} \exp\{\mathrm{In}_w^\intercal \, \mathrm{Out}_t\}}$$

- Размер словаря $V \approx 10^5$, линейная сложность для подсчета функции и градиента — очень долго.
- Нужна функция, такая что
  - $p(v \mid w, \boldsymbol{\theta}) > 0$ и $\sum_{v=1}^{V} p(v \mid w, \boldsymbol{\theta}) = 1$,
  - Считалась быстрее, чем за $O(V)$.

Hierarchical softmax

$$p(v \mid w, \boldsymbol{\theta}) = \prod_{n \in \mathrm{path}(v)} \sigma(\mathrm{ch}(n)\, \mathrm{In}_w^\intercal\, \mathrm{Out}_n).$$

- Бинарное дерево: каждому листу соответствует слово из словаря.
- Теперь "выходные" представления соответствуют не словам, а внутренним вершинам в дереве.
- $\mathrm{path}(v)$ — номера вершин на пути из корня в лист, соответствующий слову v.
- $\mathrm{ch}(n)$ — $+1$ или -1 в зависимости от того, что следующая вершина на пути — это правый или левый сын n.
- Используем, что $\sigma(x) + \sigma(-x) = 1$.

Linguistic regularities

x	$\operatorname{argmax}_w \cos(w, x)$
Berlin-Germany+Russia	Moscow
Obama-USA+Russia	Putin
king-man+woman	queen

Многозначные слова

- Проблемы:
  - Смешивание смыслов.
  - Доминирование наиболее частотного смысла.
- Ближайшие соседи по косинусному расстоянию:
  - для слова apple: macintosh, iigs, ipad, ibook;
  - для слова python: perl, php, molurus, c++, ..., monty;
- Как учесть то, что слова могут иметь больше одного смысла?

Наивная многосмысловая модель

- Введем скрытую переменную z — номер смысла, тогда:

$$p(v \mid z = k, w, \boldsymbol{\theta}) = \frac{\exp\{\mathrm{In}_{w,k}^{\intercal}\,\mathrm{Out}_v\}}{\sum_{t=1}^{V}\exp\{\mathrm{In}_{w,k}^{\intercal}\,\mathrm{Out}_t\}},$$

- Теперь каждому слову соответствуют K “входных” векторов.
- Наблюдаемые данные $(x_i, \mathbf{y}_i)$, скрытые переменные z_i (неполная разметка!).
- Полное правдоподобие:

$$p(Y, Z \mid X, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(z_i) \prod_{j=1}^{C} p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta}).$$

Обучение

- Будем максимизировать логарифм неполного правдоподобия:

$$\log p(Y \mid X, \boldsymbol{\theta}) = \log \sum_{Z} p(Y, Z \mid X, \boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}.$$

- Его можно представить в виде:

$$\log p(Y \mid X, \boldsymbol{\theta}) = \mathcal{L}(q(Z), \theta) + \mathrm{KL}(q(Z) \| p(Z \mid X, Y, \boldsymbol{\theta})),$$

  где $\mathcal{L}$ вариационная нижняя оценка:

$$\mathcal{L}(q(Z), \boldsymbol{\theta}) = \mathbb{E}_{q(Z)} \log \left[\frac{p(Y, Z \mid X, \boldsymbol{\theta})}{q(Z)} \right].$$

- Нижняя, потому что $\mathrm{KL}(q \| p) \geqslant 0$.

Обучение

- Перейдем к задаче максимизации нижней оценки:

$$\mathcal{L}(q(Z), \boldsymbol{\theta}) \to \max_{q(Z), \boldsymbol{\theta}} .$$

- Будем искать $q(Z)$ в виде $\prod_{i=1}^{N} q(z_i)$.
- Перепишем нижнюю оценку:

$$\mathcal{L}(q(Z), \boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbb{E}_{q(z_i)} \left(\log p(z_i) + \sum_{j=1}^{C} \log p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta}) \right) - \sum_{i=1}^{N} \mathbb{E}_{q(z_i)} \log q(z_i).$$

EM-алгоритм

- Е-шаг:

$$q(z_i = k) = p(z_i = k \mid \mathbf{y}_i, x_i, \boldsymbol{\theta}^{old}) \propto \exp\left\{\sum_{j=1}^{C} \log p(y_{ij} \mid k, x_i, \boldsymbol{\theta}^{old})\right\}$$

  для $k = 1, \ldots, K$ и $i = 1, \ldots, N$.
- М-шаг:

$$\mathcal{L}(q(Z), \boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}$$

  или

$$\sum_{i=1}^{N} \sum_{j=1}^{C} \mathbb{E}_{q(z_i)} \log p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}.$$

EM-алгоритм

- Как найти $\mathbb{E}_{q(z_i)} \log p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta})$?
- Стандартный прием:

$$p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta}) = \prod_{k=1}^{K} p(y_{ij} \mid k, x_i, \boldsymbol{\theta})^{\mathbb{I}[z_i = k]}.$$

- В результате:

$$\mathbb{E}_{q(z_i)} \log p(y_{ij} \mid z_i, x_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \mathbb{E}_{q(z_i)} \mathbb{I}[z_i = k] \log p(y_{ij} \mid k, x_i, \boldsymbol{\theta}) =$$

$$\sum_{k=1}^{K} q(z_i = k) \log p(y_{ij} \mid k, x_i, \boldsymbol{\theta}).$$

Стохастический EM-алгоритм

- Для i-го обучающего объекта $(x_i, \mathbf{y}_i)$ выполняем E-шаг для нахождения распределения $q(z_i)$.
- При выполнении M-шага делаем шаг по стохастическому градиенту нижней оценки:

$$\nabla \mathcal{L}(q(Z), \boldsymbol{\theta}) \approx N \sum_{j=1}^{C} \sum_{k=1}^{K} q(z_i = k) \nabla \log p(y_{ij} \mid k, x_i, \boldsymbol{\theta}).$$

- Сравним с градиентом Skip-gram:

$$\nabla L(X, Y, \boldsymbol{\theta}) \approx N \sum_{j=1}^{C} \nabla \log p(y_{ij} \mid x_i, \boldsymbol{\theta}).$$

Проблемы наивного подхода

- Предположение о том, что все слова имеют K смыслов нереалистично.
- Как автоматически подбирать число смыслов для каждого слова?
- Можно использовать процесс Дирихле.
- См. нашу работу:

  Breaking Sticks and Ambiguities with Adaptive Skip-gram.
  Bartunov S., Kondrashkin D., Osokin A., Vetrov D.
  `http://arxiv.org/abs/1502.07257`

Примеры найденных смыслов

Смысл 1	Смысл 2
almond	macintosh
cherry	iifx
plum	iigs
apricot	computers
orange	kaypro

Для слова apple.

Смысл 1	Смысл 2	Смысл 3
monty	perl	molurus
spamalot	php	pythons
cantsin	java	peafowl
zirkus	c++	tortoise
circus	objective-c	snake

Для слова python.

Разрешение лексической многозначности

Определение смысла слова по контексту:

$$\operatorname{argmax}_k p(z = k \mid \mathbf{y}, x, \boldsymbol{\theta}).$$

Примеры для слова apple

- $p(z \mid (\text{tasty, sweet}), \text{apple}) = [\mathbf{0.99998}, 0.00002]$,
- $p(z \mid (\text{announce, today}), \text{apple}) = [0.121476, \mathbf{0.878524}]$.

Примеры для слова python

- $p(z \mid (\text{interesting, show}), \text{python}) = [\mathbf{0.950317}, 0.0327315, 0.0169517]$,
- $p(z \mid (\text{code}), \text{python}) = [0.0219413, \mathbf{0.976705}, 0.00135406]$,
- $p(z \mid (\text{dangerous, animal}), \text{python}) = [0.262747, 0.00012, \mathbf{0.737133}]$.

Основные идеи

- Модель Skig-gram и иерархический soft-max.
- Стохастическая оптимизация.
- Введение скрытых переменных.
- Непараметрический байес.

Библиография

- (Mikolov 2013) Distributed representations of words and phrases and their compositionality. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Advances in Neural Information Processing Systems.
- (Hoffman 2013) Stochastic variational inference. Hoffman M., Blei D., Wang C., Paisley J. The Journal of Machine Learning Research.