Линейная алгебра в задачах векторного представления слов

Алексей Зобнин

Яндекс,
Факультет компьютерных наук НИУ ВШЭ

azobnin@hse.ru

Основано на совместной работе

- A. Zobnin and E. Elistratova, Learning Word Embeddings without Context Vectors, Repl4NLP - 2019.

План

Что будет в докладе:

- вспомним реализацию word2vec sgns;
- рассмотрим классические явные SVD-эмбеддинги;
- сформулируем модель, не требующую «контекстных» векторов.

План

Чего не будет в этом докладе:

- решения прикладных задач;
- драматического улучшения метрик;
- новых глубоких архитектур обработки текстов.

Векторные модели

Обычно модель содержит два сорта векторов: для слов и для контекстов.

Считается, что один из сортов отвечает
за синонимическую близость слов, а другой —
за синтагматическую сочетаемость.

Утверждается также, что одним сортом векторов
не обойтись.

Посмотрим, так ли это в классических моделях.

word2vec sgns

Как работает модель:

- инициализируем векторы слов W и контекстов C;
- идём скользящим окном радиуса k по тексту;
- рассматриваем текущее слово w
  и его настоящий контекст c_0;
- семплируем k случайных контекстов $c_1, \ldots, c_k$;
- оптимизируем градиентным спуском функционал

$$\mathcal{L} = -\log \sigma(\mathbf{w}\mathbf{c}_0^T) - \sum_{j=1}^{k} \log \sigma(-\mathbf{w}\mathbf{c}_j^T).$$

word2vec sgns

Пусть в строках матриц W и C стоят векторы слов и контекстов.

word2vec sgns

Пусть в строках матриц W и C стоят векторы слов и контекстов.

Факт 1

$\mathcal{L}$ зависит только от элементов произведения WC^T.

word2vec sgns

Пусть в строках матриц W и C стоят векторы слов и контекстов.

Факт 1

$\mathcal{L}$ зависит только от элементов произведения WC^T.

Факт 2

$\mathcal{L}$ не меняется при замене $W \mapsto WS$, $C \mapsto C\left(S^{-1}\right)^T$.

Проблема word2vec sgns

Решение $\left(WS, C\left(S^{-1}\right)^T\right)$ будет таким же оптимальным, как и (W, C).

На практике C часто игнорируют и используют только W. Например, скалярные произведения между всеми векторами слов — WW^T — превратятся в WSS^TW^T.

Проблема word2vec sgns

Решение $\left(WS, C\left(S^{-1}\right)^T\right)$ будет таким же оптимальным, как и (W, C).

На практике C часто игнорируют и используют только W. Например, скалярные произведения между всеми векторами слов — WW^T — превратятся в WSS^TW^T.

Если S не ортогональна, то они могут сильно измениться!

Хочется, чтобы оптимальные решения могли бы отличаться только на ортогональное преобразование.

Подходы к решению проблемы: теория

Наверное, sgd оптимизирует функционал как-то по-особенному, предпочитая одни решения другим...

- S. Gunasekar et al., Implicit Regularization in Matrix Factorization, NIPS - 2017.
- B. Neyshabur et al., Geometry of Optimization and Implicit Regularization in Deep Learning, 2017.
- S. Arora et al., Implicit regularization in deep matrix factorization, NeurIPS - 2019.

Подходы к решению проблемы: практика

Давайте оптимизировать другой функционал...

- S. Li et al., PSDVec: A toolbox for incremental and scalable word embedding, 2017.
- A. Fonarev et al., Riemannian optimization for skip-gram negative sampling, ACL - 2017.
- C. Mu et al., Revisiting Skip-Gram Negative Sampling Model with Rectification, 2019.

Почему вектора слова и контекстов различаются?

Классическое объяснение:

- скалярный квадрат вектора должен быть большим;
- но слово само с собой встречается очень редко.

Давайте обратимся к классическим моделям и сделаем некоторое наблюдение...

Классические (явные) векторные представления

Общая схема:

- определяется, что такое «контекст» слова;
- выбирается функция $f(w, c)$, как-то описывающая связь слова и контеста;
- рассматривается матрица $M = (f(w, c))$;
- строится её SVD: $M = U\Sigma V^T$;
- ранг понижается до требуемого размера d: $M \approx U_d \Sigma_d V_d^T$;
- представления слов и контекстов получаются как $W = U_d\sqrt{\Sigma_d}$ и $C = V_d\sqrt{\Sigma_d}$, откуда, $M \approx WC^T$.

Явные представления: примеры функций f

- суммарная частота появления слова w в контексте c;
- логарифм этой частоты;
- $\mathrm{PMI}(w, c) = \log \frac{P(w,c)}{P(w)P(c)}$;
- shifted PMI: $\mathrm{PMI}(w, c) - \log k$;
- shifted PPMI: $\max(\mathrm{PMI}(w, c) - \log k, 0)$.

Неявные векторные представления

В чём отличие неявных представлений от явных?

- матрицы W и C получаются не разложением большой матрицы M, а иначе.
- смысл произведения WC^T заранее не ясен.

Примеры: word2vec, fastText, GloVe, StarSpace, ...

Неявные векторные представления

В чём отличие неявных представлений от явных?

- матрицы W и C получаются не разложением большой матрицы M, а иначе.
- смысл произведения WC^T заранее не ясен.

Примеры: word2vec, fastText, GloVe, StarSpace, ...

Утверждение (Levy & Goldberg, 2014)

Для модели word2vec sgns WC^T сходится к shifted PMI-матрице, где k — количество отрицательных примеров.

Явные представления: примеры контекстов

- документ (латентно-семантический анализ);
- слово;
- лемма;
- да что угодно (см. StarSpace).

Предположения о симметричности

Мы будем рассматривать классический случай, когда контексты — это те же слова.

В этом случае функция f обычно бывает симметричной:

$$f(w, c) = f(c, w).$$

Вспомним линейную алгебру!

Задача

Дана симметричная действительная матрица M.
Что можно сказать о её SVD-разложении $M = U\Sigma V^T$?

Вспомним линейную алгебру!

Задача

Дана симметричная действительная матрица M.
Что можно сказать о её SVD-разложении $M = U\Sigma V^T$?

Неверный ответ

$U = V$.

Вспомним линейную алгебру!

Задача

Дана симметричная действительная матрица M.
Что можно сказать о её SVD-разложении $M = U\Sigma V^T$?

Неверный ответ

$U = V$.

Правильный ответ

Существует такое согласованное SVD-разложение M, в котором столбцы матриц U и V либо совпадают, либо отличаются знаком.

Существование согласованного разложения

M симметрична $\Longrightarrow$ M диагонализируется: $M = V\Lambda V^{-1}$, причём V можно выбрать ортогональной: $V^{-1} = V^T$.

Собственные значения, стоящие в Λ на диагонали, вещественны.

Перенесём знаки от отрицательных лямбд в столбцы левого множителя, получим SVD-разложение.

Согласованные разложения

Если матрица M положительно определена, то $U = V$, а сингулярные числа совпадают с собственными значениями M:

$$M = U\Sigma U^T.$$

Согласованные разложения

Если матрица M положительно определена, то $U = V$, а сингулярные числа совпадают с собственными значениями M:

$$M = U\Sigma U^T.$$

В противном случае, вообще говоря, это неверно:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \neq UU^T.$$

Согласованные разложения

Если матрица M положительно определена, то $U = V$, а сингулярные числа совпадают с собственными значениями M:

$$M = U\Sigma U^T.$$

В противном случае, вообще говоря, это неверно:

$$M = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \neq UU^T.$$

Эта же матрица показывает, что не всякое SVD-разложение согласовано:

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} \cos x & \sin x \\ \sin x & -\cos x \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix}^T.$$

Согласованные разложения

Столбцы U и V в согласованном SVD-разложении симметричной матрицы отличаются знаком в точности тогда, когда они соответствуют её отрицательным собственным значениям.

Согласованные разложения

Столбцы U и V в согласованном SVD-разложении симметричной матрицы отличаются знаком в точности тогда, когда они соответствуют её отрицательным собственным значениям.

Несогласованные разложения существуют в случае, когда у матрицы есть кратные сингулярные числа, соответствующие собственным значениям разных знаков.

В частности, если все сингулярные числа различны, то любое SVD-разложение согласовано.

А что на практике?

Для построения SVD-представлений выбираются d максимальных сингулярных значений.

Попадают ли туда те, что соответствуют смене знака?
Если да, то за что отвечают эти компоненты?

Количество «отрицательных» компонент

Проведём эксперимент для явных матриц shifted PPMI:

	$k = 1$:			$k = 5$:		
	100	**200**	**300**	**100**	**200**	**300**
en wiki	15	27	39	9	16	24
fr wiki	10	22	34	7	15	22
ru wiki	11	22	31	7	14	20

Количество «отрицательных» компонент

Проведём эксперимент для явных матриц shifted PPMI:

	$k = 1$:			$k = 5$:		
	100	**200**	**300**	**100**	**200**	**300**
en wiki	15	27	39	9	16	24
fr wiki	10	22	34	7	15	22
ru wiki	11	22	31	7	14	20

Доля «отрицательных компонент» почти не меняется при увеличении размерности и при переходе к другому языку.

Векторы слов и контекстов

Для явных SVD-моделей верны следующие следствия:

- векторы контекстов так же «хороши» для прикладных задач, как векторы слов;
- векторы контекстов получаются из векторов слов отражением по некоторым размерностям.

Векторы слов и контекстов

Для явных SVD-моделей верны следующие следствия:

- векторы контекстов так же «хороши» для прикладных задач, как векторы слов;
- векторы контекстов получаются из векторов слов отражением по некоторым размерностям.

Похожий результат (но с утверждением про половину размерностей) был получен в работе

- Zh. Assylbekov, R. Takhanov, Context vectors are reflections of word vectors in half the dimensions, 2019.

Смысл «отрицательных» компонент

Построим SVD-модель по ruwiki и посмотрим на слова с min/max значениями «отрицательной» компоненты:

top positive	
решил	0.386144
мог	0.371851
пытался	0.362951
смог	0.353824
пытается	0.347533
хотел	0.344098
вынужден	0.337988
сам	0.332514
должен	0.332267

bottom negative	
биться	-0.387683
соблазнить	-0.380923
слугой	-0.370942
мечтать	-0.361095
сойти	-0.360904
смириться	-0.360609
простить	-0.356762
признаться	-0.357335
мириться	-0.356923

Смысл «отрицательных» компонент

Ещё пример:

top positive	
белецкий	0.387469
артамонов	0.381354
превратить	0.377891
карасёв	0.374371
нефёдов	0.370402
прояснить	0.366967
пиотровский	0.36689
азаров	0.365763
перечислить	0.36311
анохин	0.362307

bottom negative	
необходимость	-0.35352
намерена	-0.340987
планировала	-0.340456
намереваясь	-0.318498
николай	-0.317861
позволив	-0.317704
владимир	-0.316368
михаил	-0.312346
намеревались	-0.310583
успела	-0.309929

Для сравнения: «обычные» компоненты

top positive	
механик	0.510821
конструктор	0.473496
инженер	0.473464
помощник	0.431185
хирург	0.430888
адвокат	0.421821
машинист	0.420466
стрелок	0.420015
водитель	0.419628

bottom negative	
веках	-0.440756
средневековье	-0.417509
столетий	-0.411927
племён	-0.411305
западных	-0.406677
веков	-0.405829
племена	-0.405003
веками	-0.403318
христианские	-0.400342

Вернёмся к модели word2vec sgns

У неё есть преимущества перед явной SVD-моделью:

- требует меньше времени и памяти для обучения;
- слегка выигрывает на практике.

Нельзя ли сделать вектора слов и контекстов «одинаковыми»?

Вернёмся к модели word2vec sgns

У неё есть преимущества перед явной SVD-моделью:

- требует меньше времени и памяти для обучения;
- слегка выигрывает на практике.

Нельзя ли сделать вектора слов и контекстов «одинаковыми»?

- Просто отождествить W и C не выйдет: M не является положительно определенной!

Вернёмся к модели word2vec sgns

У неё есть преимущества перед явной SVD-моделью:

- требует меньше времени и памяти для обучения;
- слегка выигрывает на практике.

Нельзя ли сделать вектора слов и контекстов «одинаковыми»?

- Просто отождествить W и C не выйдет: M не является положительно определенной!
- Заманчиво выйти в $\mathbb{C}$ и положить $C = \overline{W}$! Но тогда будут проблемы с вычислением $\log \sigma(\mathbf{w}\mathbf{c}^T)$.

Вернёмся к модели word2vec sgns

У неё есть преимущества перед явной SVD-моделью:

- требует меньше времени и памяти для обучения;
- слегка выигрывает на практике.

Нельзя ли сделать вектора слов и контекстов «одинаковыми»?

- Просто отождествить W и C не выйдет: M не является положительно определенной!
- Заманчиво выйти в $\mathbb{C}$ и положить $C = \overline{W}$! Но тогда будут проблемы с вычислением $\log \sigma(\mathbf{w}\mathbf{c}^T)$.
- Мы выберем другой способ.

Свяжем матрицы векторов слов и контекстов

Зафиксируем параметр q. Пусть

$$D_q = \operatorname{diag}(\underbrace{1, 1, \ldots, 1}_{p=d-q}, \underbrace{-1, \ldots, -1}_{q}).$$

Рассмотрим «скалярное произведение», заданное матрицей D_q:

$$\langle \mathbf{w_i}, \mathbf{w_j} \rangle_q = \mathbf{w_i} D_q \mathbf{w_j}^T.$$

Такое «скалярное произведение» индефинитно: квадрат вектора может оказаться неположительным.

Теперь вместо WC^T будет неявная матрица WD_qW^T.

Context-free sgns

Заменим C на WD_q в целевой функции word2vec sgns:

$$WC^T \quad \longrightarrow \quad WD_qW^T.$$

$$\mathcal{L} = -\log\sigma(\mathbf{wc_0}^T) - \sum_{j=1}^{k}\log\sigma(-\mathbf{wc_j}^T)$$

$$\downarrow$$

$$\mathcal{L}_q = -\log\sigma(\mathbf{w}D_q\mathbf{w_0}^T) - \sum_{j=1}^{k}\log\sigma(-\mathbf{w}D_q\mathbf{w_j}^T).$$

Назовём эту модель context-free sgns.

Эксперименты

- Мы обучили sgns и context-free sgns на английской Википедии.
- За baseline мы взяли fastText без n-грамм.
- Мы провели две серии экспериментов в размерности 100.

Эксперимент 1

Фиксируем $d = 100$ и варьируем q:

Dataset	**sgns**	**0**	**5**	**10**	**15**	**20**	**25**
MEN-TR-3k	**.731**	.679	.723	.719	.717	.712	.709
MTurk-287	.667	.633	.657	.660	.662	.661	**.673**
RW-Stanford	.380	.286	.409	.420	**.422**	.415	.409
SIMLEX-999	**.332**	.252	.314	.320	.313	.310	.323
SimVerb-3500	**.202**	.132	.190	.195	.198	.194	.197
VERB-143	.300	.300	.332	.330	.350	**.383**	.360
WS-353-REL	**.665**	.614	.637	.660	.639	.631	.630
WS-353-SIM	**.761**	.708	.751	.754	.753	.751	.733
average	.438	.370	.435	**.439**	.438	.435	.434

Эксперимент 2

Строим векторы для $d = 100 + q$ при разных q, затем берем только «положительную» часть векторов.

Dataset	**sgns**	**0**	**5**	**10**	**15**	**20**	**25**
MEN-TR-3k	.731	.679	.733	.735	.735	**.738**	.737
MTurk-287	.667	.634	.661	.670	.672	**.675**	.665
RW-Stanford	.380	.286	.402	**.409**	.407	.402	.400
SIMLEX-999	**.332**	.252	.314	.313	.320	.317	.324
SimVerb-3500	**.202**	.132	.193	.192	.189	.191	.191
VERB-143	.300	.300	**.336**	.315	.314	.329	.332
WS-353-REL	.665	.614	.667	.671	**.677**	.672	.674
WS-353-SIM	.761	.708	**.763**	.748	.758	.755	.748
average	.438	.370	.438	**.439**	**.439**	**.439**	**.439**

Неоднозначность решения

В классическом word2vec можно было получить эквивалентное решение заменой $W \mapsto WS$, $C \mapsto C\left(S^{-1}\right)^T$ для произвольной обратимой матрицы S.

А что в модели context-free?

Неоднозначность решения

В классическом word2vec можно было получить эквивалентное решение заменой $W \mapsto WS$, $C \mapsto C\left(S^{-1}\right)^T$ для произвольной обратимой матрицы S.
А что в модели context-free?

Неоднозначность всё равно остаётся для матриц S, сохраняющих D_q:

$$SD_qS^T = D_q.$$

Такие матрицы не обязательно ортогональны: например, для

$$D = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

подойдёт

$$S = \begin{pmatrix} \operatorname{ch} x & \operatorname{sh} x \\ \operatorname{sh} x & \operatorname{ch} x \end{pmatrix}.$$

В поисках канонического решения

Пусть W — решение задачи context-free sgns.
Можно ли среди эквивалентных решений WS, где $SD_qS^T = D_q$, выбрать какое-нибудь каноническое?

В поисках канонического решения

Пусть W — решение задачи context-free sgns.
Можно ли среди эквивалентных решений WS, где $SD_qS^T = D_q$, выбрать какое-нибудь каноническое?

Заметим, что в явных SVD-моделях $W = U\sqrt{\Sigma}$. В этом случае $W^TW = \Sigma$ — диагональная матрица.

В поисках канонического решения

Пусть W — решение задачи context-free sgns.
Можно ли среди эквивалентных решений WS, где $SD_qS^T = D_q$, выбрать какое-нибудь каноническое?

Заметим, что в явных SVD-моделях $W = U\sqrt{\Sigma}$. В этом случае $W^TW = \Sigma$ — диагональная матрица.

Можно ли взять такую S, чтобы $(WS)^T(WS) = \text{diag}$?

В поисках канонического решения

Пусть W — решение задачи context-free sgns.
Можно ли среди эквивалентных решений WS, где $SD_qS^T = D_q$, выбрать какое-нибудь каноническое?

Заметим, что в явных SVD-моделях $W = U\sqrt{\Sigma}$. В этом случае $W^TW = \Sigma$ — диагональная матрица.

Можно ли взять такую S, чтобы $(WS)^T(WS) = \text{diag}$?

Ответ: да. Это следует из теоремы о приведении пары форм D_q и W^TW к диагональному виду.

В поисках канонического решения

Такую матрицу S можно найти из SVD-разложений для W и $W^T W$.

Увы, на практике такие канонические решения оказались хуже обычных.

В поисках канонического решения

Такую матрицу S можно найти из SVD-разложений для W и $W^T W$.

Увы, на практике такие канонические решения оказались хуже обычных.

Похожие идеи можно найти в работе

- Carrington et al., Invariance and identifiability issues for word embeddings, NeurIPS - 2019.

Открытые вопросы и задачи

- Как ведет себя доля «отрицательных» компонент PPMI-матриц для других языков и при больших размерностях?
- Какова лингвистическая роль «отрицательных» компонент?
- Внедрить context-free в GloVe.
- Добавить регуляризатор к context-free-модели.
- Найти связь с гиперболическими эмбеддингами.
- Описать неявную регуляризацию в word2vec.
- Рассматривать слова как собственные контексты (это приблизит матрицу M к положительно определенной).