Методы реализации расстояний с использованием разнородной обучающей информации

Выполнил студент 417 группы Сенин Александр Николаевич
Научный руководитель: к.ф.-м.н. Майсурадзе Арчил Ивериевич

31 мая 2021 г.

Введение

В задачах анализа данных принято работать с векторными представлениями объектов. Источники векторных представлений различны, в частности это могут быть входные различия на парах объектов.

Методы реализации расстояний, в частности методы многомерного шкалирования позволяют построить векторные представления объектов по входным различиям.

Общая идея этих методов — найти конфигурацию точек, наилучшим образом приближающую входные различия на парах объектов. В дальнейшем точки этой конфигурации и будут использоваться как векторные представления.

Векторные представления слов

В области анализа текстов и информационного поиска принято работать с векторными представлениями слов.

Исходными данными для построения векторных представлений может быть разнородная исходная информация: размеченные экспертами степени семантической близости слов (показатели на парах слов) и наборы аналогии (четверки слов).

Цели и задачи

Цели работы:

1. Предложить нейросетевой подход к решению задач метрического и неметрического многомерного шкалирования.
2. Предложить способ построения векторных представлений слов на основе разнородной обучающей информации методами многомерного шкалирования.

Задачи:

1. Разработать и реализовать модель МО, решающую задачи метрического и неметрического многомерного шкалирования.
2. Реализовать программную библиотеку для работы с разнородными исходными данными.
3. Разработать модель МО для построения векторных представлений слов методами многомерного шкалирования на основе разнородных данных.

Многомерное шкалирование

Пусть $\mathcal{X}$ — выборка объектов. На некоторых парах M этого множества даны попарные различия $\{d_{ij}\}_{(i,j)\in M}$.

Дано метрическое пространство $(\mathbb{R}^p, \rho)$.

Требуется найти «точечную конфигурацию» $x_1, ..., x_N \in \mathbb{R}^p$, реализующую различия d_{ij} в терминах (1):

$$\rho(x_i, x_j) \approx d_{ij}, \tag{1}$$

Метрическое vs неметрическое

- *Задача метрического многомерного шкалирования (metric multidimensional scaling, mMDS):*

$$Stress(x_1, ..., x_N) = \sum_{(i,j) \in M} (d_{ij} - \rho(x_i, x_j))^2 \longrightarrow \min_{x_1, ..., x_N \in \mathbb{R}^p} \tag{2}$$

- *Задача неметрического многомерного шкалирования (non-metric multidimensional scaling, nMDS):*

Найти $x_1, ..., x_N \in \mathbb{R}^p$:

$$\forall\, (i,j), (k,l) \in M \;\; d_{ij} < d_{kl} \Longleftrightarrow \rho(x_i, x_j) + m < \rho(x_k, x_l), \tag{3}$$

где $m > 0$ — отступ.

Решение mMDS предлагаемым подходом

Пусть отображение объектов в целевое пространство $l_w : \{1, ..., N\} \to \mathbb{R}^p$ осуществляется полносвязным слоем нейронной сети с весами w.

Используем логику работы *сиамских сетей*. Обучать сеть l_w будем обратным распространением ошибки.

Для задачи *метрического многомерного шкалирования* будем оптимизировать функцию потерь stress loss:

$$L_{mMDS}(x_i, x_j) = (d_{ij} - \rho(x_i, x_j))^2 \tag{4}$$

$$Stress = \sum_{(i,j) \in M} L_{mMDS}(x_i, x_j) \longrightarrow \min_{x_1, ..., x_N \in \mathbb{R}^p} \tag{5}$$

Решение nMDS предлагаемым подходом

Для задачи *неметрического многомерного шкалирования* будем использовать triplet loss, где в качестве anchor-positive и anchor-negative для объекта с номером i будем брать такие объекты с номерами j и k соответственно, что $d_{ij} < d_{ik}$.

Таким образом, пусть пары $(i, j), (i, k) \in M$ и $d_{ij} < d_{ik}$:

$$L_{nMDS}(x_i, x_j, x_k) = max(0, \rho(x_i, x_j) - \rho(x_i, x_k) + m), \quad (6)$$

Тогда функция потерь будет выглядеть так:

$$L_{nMDS} = \sum_{(i,j),(i,k) \in M} L_{nMDS}(x_i, x_j, x_k) \quad (7)$$

Решение nMDS предлагаемым подходом

Естественнее работать с четверками. Будем также использовать quadruplet loss.

Пусть пары $(i, j), (k, l) \in M$ и $d_{ij} < d_{kl}$:

$$L_{nMDS}(x_i, x_j, x_k, x_l) = max(0, \rho(x_i, x_j) - \rho(x_k, x_l) + m) \quad (8)$$

Тогда функция потерь будет выглядеть так:

$$L_{nMDS} = \sum_{(i,j),(k,l) \in M} L_{nMDS}(x_i, x_j, x_k, x_l) \quad (9)$$

Демонстрация предлагаемого подхода

Реализация на PyTorch.
Сгенерирован набор данных $s_1, ..., s_N \in \mathbb{R}^D$ при $D = 3$.
Функция различия $d(s_i, s_j) = \rho_E(s_i, s_j)$ — евклидово расстояние.
Размерность искомой конфигурации $p = 2, 3$.
Сравниваем величину нормированного функционала стресса и ранговую корреляцию Спирмена.

Решение задачи метрического многомерного шкалирования:

Размерность конфигурации p = 3		
Метод	Stress	Корреляция
cMDS	10^{-16}	1.0
SMACOF	10^{-3}	1.0
fc-mMDS	10^{-3}	1.0

Размерность конфигурации p = 2		
Метод	Stress	Корреляция
cMDS	0.277	0.811
SMACOF	0.218	0.849
fc-mMDS	0.222	0.845

Демонстрация предлагаемого подхода

Реализация на PyTorch.

Сгенерирован набор данных $s_1, ..., s_N \in \mathbb{R}^D$ при $D = 3$.

Функция различия $d(s_i, s_j) = \rho_E(s_i, s_j)$ — евклидово расстояние.

Размерность искомой конфигурации $p = 2$.

Сравниваем ранговую корреляцию Спирмена.

Решение задачи неметрического многомерного шкалирования:

Размерность конфигурации p = 2	
Метод	Корреляция
cMDS	0.811
sklearn nMDS	0.017
SMACOF	0.849
triplet fc-nMDS	0.848
quadruplet fc-nMDS	**0.851**

Задача построения векторных представлений слов

Будем работать со словами естественного языка. Пусть задан *словарь* V — множество слов.

Задача построения векторных представлений заключается в нахождении и сопоставлении конфигурации $x_1, ..., x_N \in \mathbb{R}^p$ всем словам из словаря V таким образом, что $x_1, ..., x_N$ отвечают некоторым критериям качества.

Используем внутреннее оценивание:

1. Сохранение семантической близости на парах слов с помощью ранговой корреляции Спирмена с размеченными экспертами «золотыми стандартами» MEN, SimLex-999, WordSim353
2. Сохранение аналогий с помощью точности (accuracy) решения задач аналогий на наборе BATS.

Библиотека

Для работы с «золотыми стандартами» — наборами семантической близости и наборами аналогий разработана и реализована библиотека.

Наборы MEN, SimLex-999, WordSim353 проанализированы, устранены дефекты (асимметричность, дубликаты и т.д.).

Реализованы классы для представления пары слов, хранилища «золотого стандарта», «золотого стандарта», процедуры оценивания.
Реализован функционал нахождения распределения частей речи, нахождения словарей наборов, оценивания представлений (ранговая корреляция Спирмена и точность решения задач аналогий).
`https://github.com/obj2vec/datasets_similarity/blob/main/EvaluateEmbeddingsLab.py`

Построение векторных представлений

Дано: размеченные экспертами наборы семантической близости на парах слов M: $\{s_{ij}\}_{(i,j)\in M}$, набор аналогий A, словари наборов пересекаются.

Задача: предложить способ построения векторных представлений слов методами многомерного шкалирования.

Набор экспертных близостей $\{s_{ij}\}$ преобразуем в различия $\{d_{ij}\}$. Используем предложенную модель с функцией потерь stress loss в случае метрического многомерного шкалирования и функцией потерь quadruplet loss в случае неметрического многомерного шкалирования. Тогда можем строить векторные представления слов на основе «золотых стандартов» семантической близости.

Построение векторных представлений

Для привлечения наборов аналогий усовершенствуем модель, уже работающую с четверками.

Используем analogy loss (подход PairDistance) на четверках слов $a : a' :: b : b'$

$$L_{analogy} = -\cos(b' - b, a' - a) \tag{10}$$

Тогда искомый способ построения векторных построений заключается в нахождении конфигурации $x_1, ..., x_N$ путем минимизации функции потерь:

$$L = \sum_{(i,j),(k,l) \in M} L_{nMDS} + \alpha \sum_{(i,j,k,l) \in A} L_{analogy} \tag{11}$$

Сравнительный анализ полученного решения

Реализация на PyTorch. Размерность $p = 100$. Обучаемся на: семантическая близость — MEN (набор с наиболее удачным соотношением размера словаря и числа пар слов; есть пары разных частей речи),
набор аналогий A — BATS (большое разнообразие типов связи, большое пересечение словаря с MEN, работаем со словами, которые есть в MEN)

Метод	корр. MEN	корр. WS353	корр. SimLex
fc-mMDS	0.99	0.34	0.30
fc-nMDS	0.99	0.34	0.31
fc-nMDS + analogy	0.95	0.34	0.30
GloVe (wiki2010)	0.68	0.60	0.32

Аналогии решаем методом PairDistance.

Метод	корр. MEN	точн. BATS
fc-nMDS	**0.99**	4.7%
fc-nMDS + analogy	0.95	**20.4%**

Результаты

На защиту выносятся:

- Архитектура для решения задач метрического и неметрического многомерного шкалирования с использованием специализированных функций потерь типа quadruplet loss.
- Программная библиотека для работы с наборами семантической близости и наборами аналогий, а также для оценивания векторных представлений слов.
- Новый подход к построению векторных представлений слов на основе разнородной обучающей информации: размеченная экспертами семантическая близость и наборы аналогий.

Формализация близости и различия

Различием (*dissimilarity*) будем называть функцию $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\forall x, y \in \mathcal{X}$ удовлетворяющую следующим требованиям:

(D1) $d(x, x) = 0$

(D2) $d(x, y) \geqslant 0$

Сходством (*similarity*) будем называть функцию $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\forall x \in \mathcal{X}$ удовлетворяющую единственному требованию:

(S1) $s(x, x) > 0$

Решение nMDS

Раньше задачу nMDS сводили к задаче оптимизации:

$$Stress(x_1, ..., x_N) = \sum_{(i,j) \in M} (f(d_{ij}) - \rho(x_i, x_j))^2 \longrightarrow \min_{x_1, ..., x_N \in \mathbb{R}^p}, \tag{1}$$

где f находится методами изотонической регрессии.

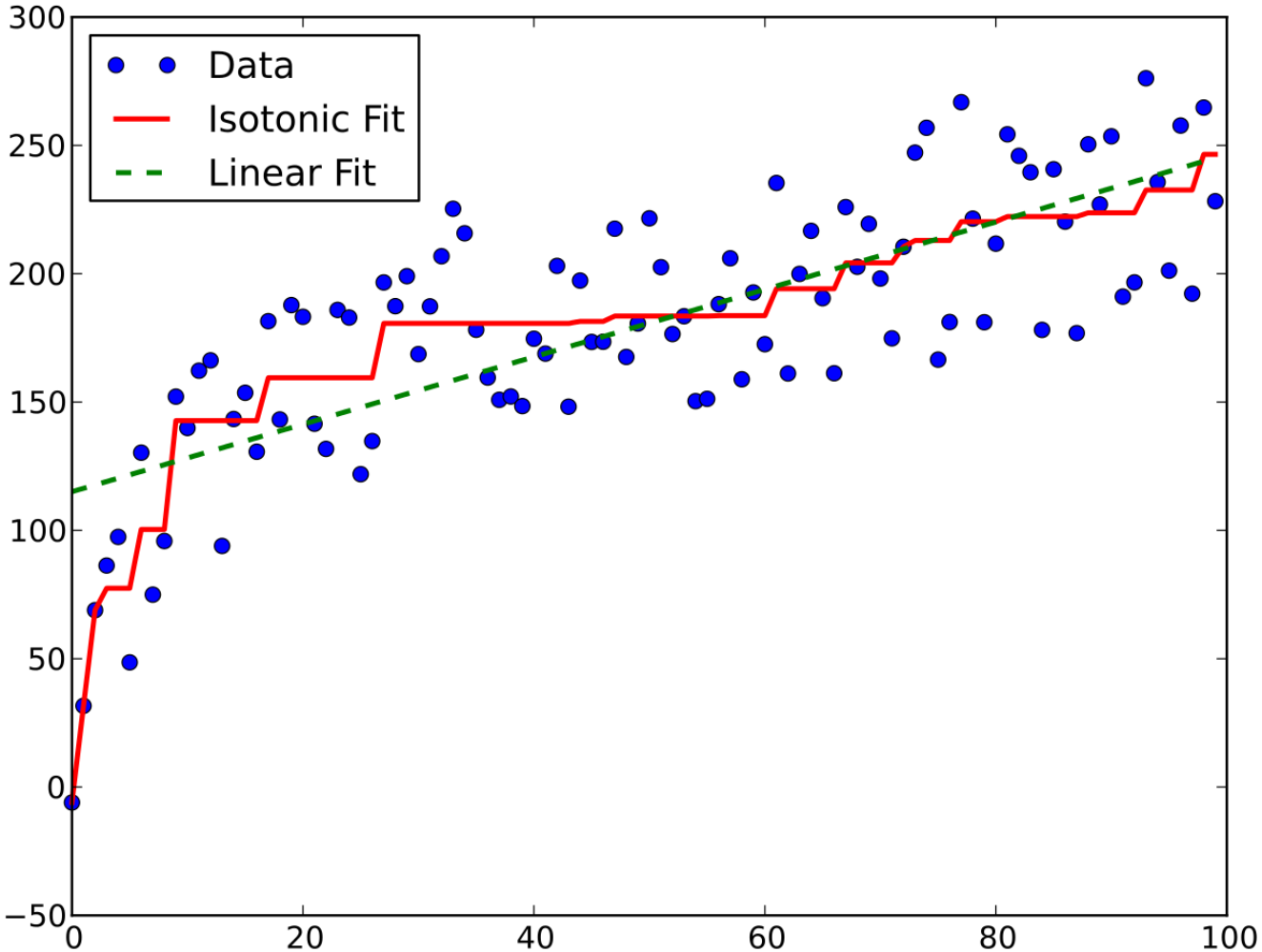
Предлагаемая формулировка

$$\forall\, (i,j), (k,l) \in M \;\; d_{ij} < d_{kl} \Longleftrightarrow \rho(x_i, x_j) + m < \rho(x_k, x_l), \tag{2}$$

плодотворнее.

Изотоническая регрессия

Иначе: монотонная регрессия.



Определения

Сиамская сеть — нейронная сеть из двух идентичных подсетей, разделяющих общие веса.

Triplet loss: $L = max(0, d_{ap} - d_{an} + m)$

Нормированный Stress:

$$Stress(x_1, ..., x_N) = \sqrt{\frac{\sum_{(i,j)\in M}(d_{ij} - \rho(x_i, x_j))^2}{\sum_{(i,j)\in M} d_{ij}^2}} \quad (3)$$

Подробнее про библиотеку

`https://github.com/obj2vec/datasets_similarity/blob/main/EvaluateEmbeddingsLab.py`
Выбраны популярные наборы данных (SimLex-999, WordSim353, MEN), проанализированы способы их построения (эксперты оценивают все пары, после усреднение или эксперт оценивает одну пару через сравнение с другой парой).

Обнаружены недостатки: нечувствительность к разнице similar и related, асимметричность, дубликаты, нет пар разных частей речи.

Подробнее про библиотеку

Недостатки, не связанные с неудачным способом построения, устранены.
Реализованы классы для представления пары слов, хранилища «золотого стандарта», «золотого стандарта», процедуры оценивания.
Реализован функционал нахождения распределения частей речи, нахождения словарей наборов, оценивания представлений (ранговая корреляция Спирмена и точность решения задач аналогий), разбиения на обучение и контроль (с сохранением распределения частей речи и с уникальными словами в контроле).