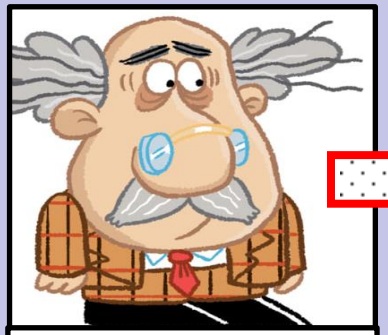
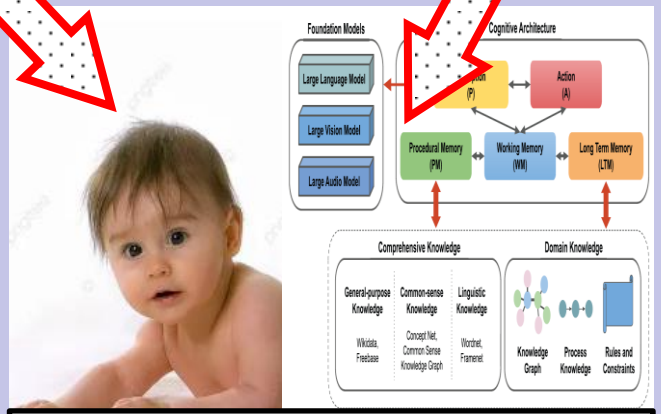


Symbolic AI

LLM-basd AI



Hybrid AI

Метамодель связей на множестве атрибутов данных и ее применения в задачах обработки больших данных

В.И. Городецкий, АО ЭВРИКА
vladim.gorodetsky@gmail,com

Содержание

1. Мотивация работы
2. Модель данных выборки и метамодель структуры статистических связей на множестве атрибутов экземпляров выборки
3. Модель обнаружения аномалий в экземплярах данных выборки с использованием ее метамодели
4. Латентные пространства и латентные переменные пространства данных
5. Оптимизационная постановка задачи поиска латентного пространства
6. Алгоритм поиска латентного пространства
7. Приложения в задачах обработки больших данных
8. Обсуждение и заключение

1. Мотивация работы

Мотивация работы (1/2)

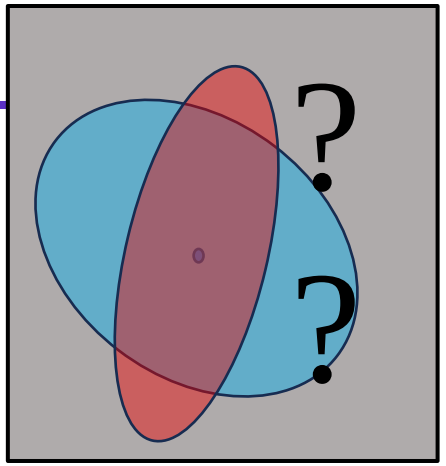
Задачи обработки **больших данных** становятся сложнее с каждым годом. Они ставят много **безответных вопросов** как к алгоритмам, так и программным инструментам.

- **Новые источники** крупномасштабных гетерогенных данных: социальные сети, СМИ, …;
- **Низкое качество данных**, извлекаемых из Интернет: ненадежные источники, смысловые искажения данных, …;
- **Непредсказуемая неполнота** данных, собираемых из Интернет, в которых **пропущенных** значений может быть больше, чем означенных;
- **Ненадежная разметка** данных, получаемая за счет средств автоматической разметки;
- **Вредоносное программное** обеспечение типа атак на данные;
- **Распространенность синтетического контента**, генерируемого средствами ИИ;
- **Заведомо ложная информация**, как синтетическая, так и созданная людьми;

Предвзятая информация, представляющая точку зрения определенных политических движений, маргинальных групп; В LLM-моделях ИИ проблема предвзятости данных имеет особую остроту. Она уже не раз была причиной ошибочных решений стратегического уровня.

//DARPA объявила несколько специальных Программ по решению **только проблемы предвзятости**//

Мотивация работы (2/2)



Ужесточаются требования к качеству данных со стороны приложений, особенно со стороны **критических приложений** («приложений с высокими ставками»).

По данным **Gartner**, время подготовки данных к машинному обучению в среднем составляет около 35 % от общего времени разработки приложений, и это очень много.

В выступлении обсуждается принципиально новый подход к решению различных проблем обработки больших данных, некоторые из которых были упомянуты здесь. Он базируется на **нетрадиционной** модели данных, ориентированной **на описание внутренней структуры** данных, т. е. на описании связей на множестве атрибутов данных, а не на топологических свойствах различных выборок В этой модели, как и в пространстве эмбеддингов LLM-моделей данных, **близость пары векторов** атрибутов данных определяется

- не в векторном пространстве атрибутов данных с эвклидовой мерой близости , как это имеет место в традиционных моделях данных,
- а в пространстве, в котором близкими оказываются пары описаний, **сильно связанных** линейной статистической связью, точно так же, как это имеет место в пространствах эмбеддингов.

2. Исходная модель данных выборки и метамодель структуры линейных статистических связей на множестве атрибутов экземпляров выборки

Принятые предположения о модели данных в традиционной форме и ключевые обозначения

Далее везде рассматривается случайная выборка данных $\boldsymbol{XX}=[X_1, \dots, X_n]^T, X_i \in R^m, i=1,\dots,n$, из генеральной совокупности многомерных случайных величин $X=[x_1, \dots, x_m]^T$ с нормальным распределением вероятностей $X \sim N_m\{\bar{X}, \boldsymbol{W}\}$, где $\bar{X}$ и $\boldsymbol{W}$ –эмпирические оценки вектора математического ожидание (м. о.) и матрицы ковариаций, вычисленные по выборке $\boldsymbol{XX}$.

Важные замечания

1. Результаты, обсуждаемые далее, будут справедливы также и для выборок данных с другими распределениями, если ограничиться классом *линейных статистических связей* [Рао, 1968].

2. Замечание технического характера. Далее параллельно используются два эквивалентных способа представления многомерных данных: в форме *множеств случайных величин* и в форме *векторов случайных величин*, как показано в таблице.

Векторы атрибутов данных	**Множества атрибутов данных**
$X=[x_1, \dots, x_m]^T$ $Y=[x_{j_1}, \dots, x_{j_r}]^T, Z=[x_{i_1}, \dots, x_{i_k}]^T$	$X=\{x_1, \dots, x_m\}$ $Y=\{x_{j_1}, x_{j_2}, \dots, x_{j_r}\}, Z=\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\},$
$\boldsymbol{W}_X, \boldsymbol{W}_{YY}, \boldsymbol{W}_{ZY}, \boldsymbol{W}_{YZ}, \boldsymbol{W}_{ZZ}$	

Формальное представление линейных статистических связей на множестве координат векторного пространства данных

Множество всех линейных функционалов вектора X образует линейное векторное пространство $\mathcal{M}(\mathrm{X})$ (традиционный вариант модели), а ранг $r \leq m$ матрицы $\boldsymbol{W}$ задает его размерность, обычно $r << m$.

Хорошо известно, что для любой пары подмножеств атрибутов экземпляра данных связи на множестве атрибутов позволяют строить оптимальные линейные оценки атрибутов любого из них при заданных атрибутах другого. Эти оценки имеют распределение вероятностей, которое совпадает с условным распределением подмножества неизвестных атрибутов при заданных значениях атрибутов другого подмножества. Это распределение вероятностей является нормальным. При этом оптимальные (линейные) оценки задаются уравнениями линейной регрессии. Множество всех таких уравнений линейной регрессии(число связей) имеет экспоненциальную мощность $2^{|2m|}$,где m – размерность вектора атрибутов.

В наших обозначениях уравнения линейной регрессии для конкретной пары подмножеств атрибутов $\boldsymbol{Y}=\{x_{j_1}, x_{j_2}, \ldots, x_{j_r}\}$ и $\boldsymbol{Z}=\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$, таких, что $\boldsymbol{Z} \cap \boldsymbol{Y}=\varnothing$, $\boldsymbol{Z} \cup \boldsymbol{Y} \subseteq \boldsymbol{X}$., Y, при известных значениях атрибутов из подмножества $\boldsymbol{Y}$ (они объединены в вектор $\bar{\bar{Y}}$) , имеют вид ([Андерсон 1962],[Рао 1968]):

$$\hat{Z} = \bar{Z} - \boldsymbol{W}_{ZY}(\boldsymbol{W}_{YY})^{-1}(\bar{\bar{Y}} - \bar{Y}),$$

где $\bar{Y}, \bar{Z}, \boldsymbol{W}_{YY}, \boldsymbol{W}_{ZZ}, \boldsymbol{W}_{ZY}$ – характеристики (блочного представления) исходного распределения вероятностей, и матрица условного распределения оценки $\hat{Z}$ вычисляется по формуле

$$\boldsymbol{W}_{\hat{Z}\hat{Z}} = \boldsymbol{W}(Z \mid Y) = \boldsymbol{W}_{ZZ} - \boldsymbol{W}_{ZY}(\boldsymbol{W}_{YY})^{-1}\boldsymbol{W}_{YZ},$$

причем она не зависит от реализации $\bar{\bar{Y}}$.

Метамодель множества связей

1. **Эмпирические оценки** вектора **м. о**. и м**атрицы ковариаций** множества атрибутов данных. Матрица ковариаций в неявной форме задает некоторое бинарное отношение. Она является формой представления знаний о парных связях атрибутов. Как известно, она описывает в неявной форме линейные статистические связи на множестве атрибутов выборки в целом, но не свойства отдельных ее экземпляров. т.е. она описывает общие свойства связей атрибутов конкретной выборки. По отношению к атрибутам эту характеристику можно отнести к **метауровню**.

2. **Модель линейной регрессии (МЛР)**: она явно описывает связи на подмножествах атрибутов в функциональном виде, на множестве атрибутов *в выборке в целом*. Эта модель рассматривается здесь в качестве второй компоненты **метамодели**. Ее входом является известная реализация $\overline{\overline{Y}}$ вектора Y, составленного из аргументов **МЛР,** и параметров его (нормального) распределения, а ее результатом – оптимальная оценка вектора, составленного из множества искомых атрибутов.

Эта пара компонент метамодели полностью определяет статистические свойства связей на множестве атрибутов выборки.

Соотношение между традиционной моделью данных в векторном пространстве и метамоделью структуры связей

Модель генеральной совокупности данных Ψ

1 Структура вектора атрибутов данных $X=[x_1, \dots, x_m]^T$;

2. Многомерное распределение вероятностей $P(X)$

Механизм случайного выбора

Случайная выборка данных

$XX=[X_1, \dots, X_n]^T, X_i \in R^m,$

$i=1,\dots,n,$

с эмпирическим нормальным распределением вероятностей

Вычисление оценок и распределений вероятностей неизвестных компонент экземпляра данных

Вход:

1. Экземпляр данных X_s ;
2. Два списка атрибутов $Y=\{x_{i_1}, \dots, x_{i_k}\}$ и $Z=\{x_{j_1}, \dots x_{j_r}\}$

Алгоритм вычисления модели данных выборки

Модель данных выборки

1 Вектор атрибутов данных $X=[x_1, \dots, x_m]^T$;

2. **Эмпирические оценки** $\bar{X}$ и матрицы ковариаций W

Метамодель структуры связей атрибутов=

= (1) Модель уравнений регрессии + (2) параметры модели данных, специализирующие модель

Результат

1. Оптимальные линейные оценки значений атрибутов экземпляра данных X_s из подмножества Z при заданных значениях его атрибутов из подмножества Y..

2. Параметры условного распределения оптимальной оценки.

Краткие промежуточные выводы

Предложенная метамодель структуры связей атрибутов данных полностью описывает и модель данных, и все связи между любой парой подмножеств атрибутов, а также правила трансформации модели связей в зависимости от дополнительной информации о значениях любого подмножества ее атрибутов.

1. Метамодели структур связей двух различных выборок, как правило, различны, и поэтому ее можно рассматривать как цифровой идентификатор выборки, аналог биометрического кода человека и, даже скорее, ее можно рассматривать как ДНК выборки. По этому ДНК можно восстановить «почти такую же» выборку (аналогично тому, как это делается в диффузионных моделях, только здесь – не в концепции «черного ящика», т.е. лучше).

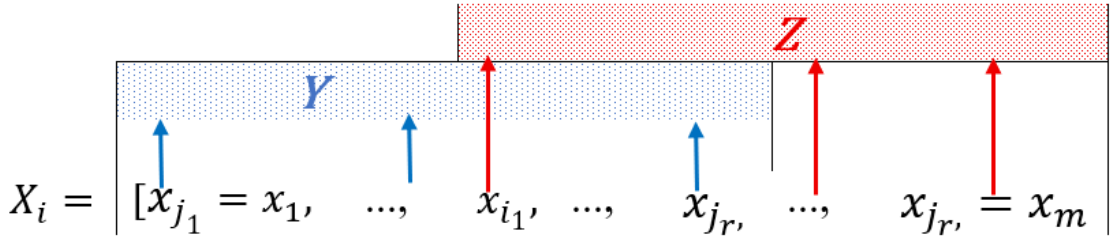
2. **Внешние воздействия** на экземпляры выборки, случайные или сознательные, любое ее **повреждение** данных теоретически можно **обнаружить** сравнением эталонной и текущей метамоделей. Вопрос –каким образом это можно выполнить?

Теоретически понятно, что на основе метамодели связей выборки можно строить различные алгоритмы обнаружения аномалий в конкретных экземплярах данных. Рассмотрим один из таких алгоритмов.

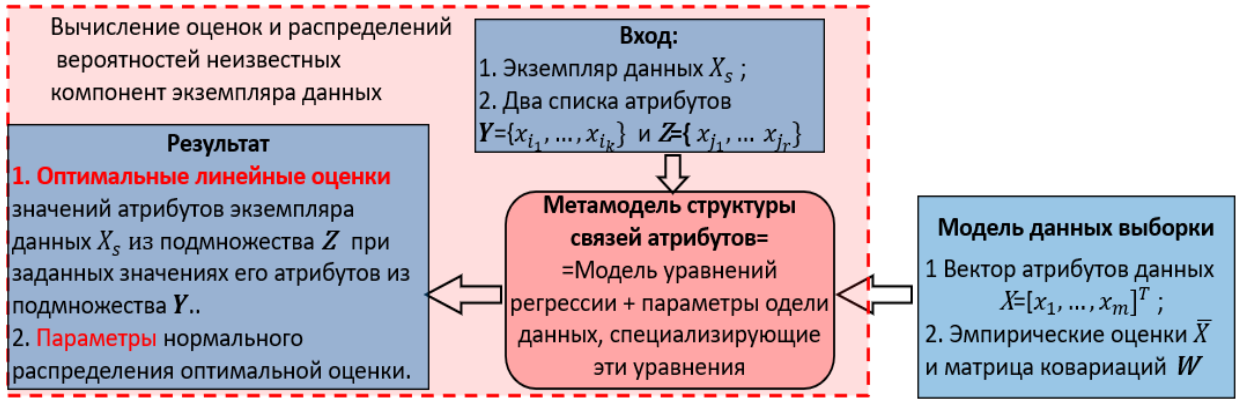
3. Подход к обнаружению аномалий в экземплярах данных выборки с использованием метамодели связей

Концепция обнаружения аномалий в экземплярах данных на основе метамодели множества связей атрибутов

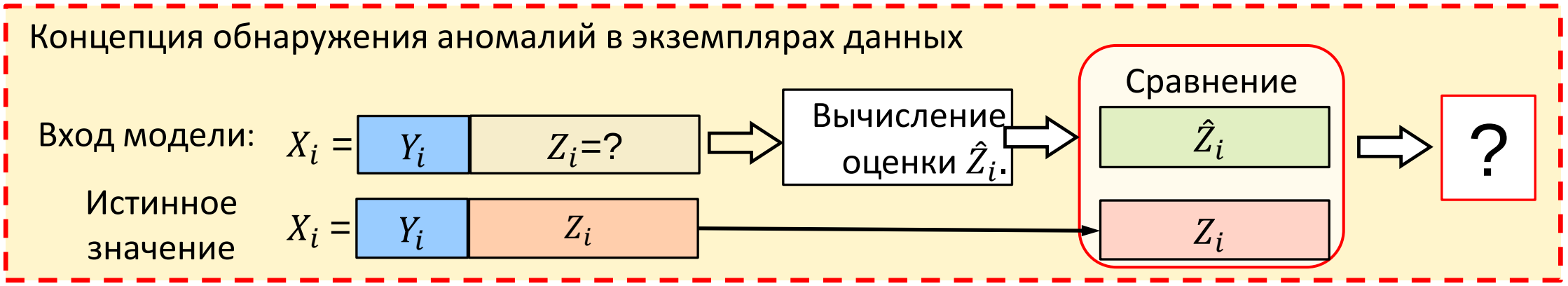
Вход модели – экземпляр $X_i \in \boldsymbol{XX}$ и произвольное разбиение подмножества его атрибутов на пару подмножеств $[\boldsymbol{Y}, Z]$, например, как показано ниже:



После перенумерации атрибутов и перестановки их в соответствии с новым порядком



Использование метамодели связей выборки ***XX*** в задаче «**Обнаружение аномалий»**



А как правильно выбрать разбиение, чтобы обеспечить требуемую точность оценивания второй части экземпляра данных?

Результаты сравнения векторов могут быть основой для принятия решений только в том случае, если полученная оценка второй части разбиения будет обладать достаточной точностью. В принятых предположениях для нее можно вычислить распределение вероятностей $\boldsymbol{N}(\overline{\boldsymbol{Z}}|\overline{\overline{\boldsymbol{Y}}}, \boldsymbol{W}(Z\,|\,Y)$, где $\boldsymbol{W}(Z\,|\,Y) = \boldsymbol{W}_{ZZ} - \boldsymbol{W}_{ZY}(\boldsymbol{W}_{YY})^{-1}\boldsymbol{W}_{YZ}$.

Вопрос:

.

Как следует выбирать подмножество атрибутов $\boldsymbol{Y}$, которое используется в качестве аргументов уравнения регрессии, чтобы обеспечить требуемую точность оценок?

Множество возможного выбора $\boldsymbol{Y}$ в общем случае имеет экспоненциальную сложность $2^{|2m|}$ от размерности m векторов пространства данных. Заметим, что прообразом этой задачи является задача редукции пространства данных, для решения которой уже имеется ряд алгоритмов. Но, во первых, эта задача тоже имеет экспоненциальную сложность, алгоритмы ее решения достаточно трудоемки и дают только приближенные решения, во-вторых, далеко не всегда существует множество атрибутов $\boldsymbol{Y}$, гарантирующее получебние нужной точности.

Далее будет предложен алгоритм, основанный на использовании концепции латентного пространства, в котором явно представляется скрытая (латентная) структура данных выборки.

4. Латентные пространства и латентные переменные пространства данных

Что такое латентные пространства и латентные переменные (1/3)

Латентное (скрытое) пространство — это компактное, низкоразмерное пространство представления сложных данных (изображений, текста, звука), применяемое в нейросетях. В этом пространстве объекты преобразуются в векторы-точки, где схожие по смыслу элементы расположены рядом, что позволяет ИИ «понимать» и генерировать новые данные.

Латентные (скрытые) **переменные** — множество некоррелированных случайных величин (атрибутов данных), обладающих определенными свойствами, и тогда л**инейное латентное пространство –** это **множество всех линейных функционалов** от латентных переменных.

Содержательно:

Латентные (скрытые) переменные — это факторы, о значениях которых можно судить по косвенным признакам — наблюдаемым переменным, т.е. это глубинные переменные, которые зависят

- от взаимосвязей на множестве переменных пространства данных, и
- от варианта использования этих переменных в приложении. Этот вариант формулируется в виде набора ограничений и оптимизируемого функционала.

Латентные переменные являются результатом решения оптимизационной задачи, но не результатом измерений.

Примеры латентных переменных и латентных пространств (2/3)

- **В психологии:** Мы не можем измерить «интеллект» человека (это латентная переменная). Но мы можем судить о нем косвенно на основе результатов тестирования по математике, физике, по языку, литературе и искусству и т.п. (по значениям наблюдаемых переменных), измеряемых, например, с помощью тестов, экзаменов и т.п.).
- **В бизнесе:** «Лояльность клиента» нельзя увидеть физически или измерить. Но её можно вычислить через частоту покупок, средний чек и отзывы.
- **В анализе данных:** В тематическом моделировании (NLP) «темы» текста являются латентными переменными, которые проявляются через специфический набор слов или терминов (наблюдаемые переменные).
- **В медицине:** используется понятие синдрома, под которым понимается устойчивая совокупность симптомов (признаков), которые часто встречаются вместе в некотором механизме развития, возможно, различных болезней. Синдром — это промежуточное звено процесса принятия решений. Диагноз ставится по множеству синдромов, которые врач строит на основе наблюдаемых симптомов.

Примеры латентных переменных и латентных пространств (3/3)

- **В математической статистике:** статистический метод, называемый факторным анализом, в котором предполагается, что наблюдаемые переменные могут быть представлены как функции от меньшего количества ненаблюдаемых переменных (факторов) и случайной ошибки (Ч. Спирмен, 1904).
- **В производственном процессе**: факторы, вызывающие аварии. Одна из задач управления технологическим процессом состоит в том, чтобы о найти такие факторы как функции от наблюдаемых параметров процесса. Например, в бумагоделательном производстве стоит задача поиска факторов, которые обуславливают обрыв бумаги.

Латентные переменные и латентные пространства в больших языковых моделях (1/2)

Понятие латентных переменных и латентных пространств, которые являются базовыми понятиями в LLM-моделях, и содержательно, и формально ничем не отличаются от понятий с такими же названиями, описанными выше применительно к символьной модели ИИ. **Ничем**! Покажем это.

1. Латентные переменные в LLM-моделях используются для кодирования невидимой структуры данных. В LLM кодирование текста, изображения или звука выполняется не с помощью слов, пикселей или частот (как в традиционном символьном ИИ), а с помощью координат множества латентных пространств, называемых пространствами эмбеддингов.

2. Каждому такому пространству отвечает *слой* нейросети, в котором представлены смысловые категории определенного уровня агрегирования обучающих данных (аналогично онтологии, но смысловые понятия, представленные узлами LLM не существуют в ЕЯ). Эти пространства (слои нейросети) упорядочены в соответствии с уровнями агрегирования смысла данных выборки.

3. Построение каждого латентного пространства эмбеддингов выполняется как процедура оптимизации некоторого квадратичного функционала с помощью градиентного спуска, которая именуется в LLM процедурой обучения. Цель каждой задачи обучения LLM – это построение латентного пространства (пространства эмбеддингов), которые описывают **не данные, а их взаимосвязи.** При обучении нейросеть ищет в данных не совпадения, а повторяющиеся формы, пропорции, отношения, стабильные при вариации деталей.

Латентные переменные и латентные пространства в больших языковых моделях и символьном ИИ (2/2)

4. В итоге строится латентное пространство непрерывных переменных , математически аналогичное латентному пространству в области символьного ИИ.

5. Трансформер реализует функцию минимизации размерности построенного пространства с сохранением **возможности его реконструкции**. Аналогичная задача решается применительно к модели латентного пространства в символьном ИИ (см. далее).

5. Формально построение модели латентных пространств при обучении LLM-модели выполняется при точно тех же предположениях о нормальности распределения данных выборки в пространствах эмбеддингов, что и принятые в постановке, описанной выше.

Гипотезы

1. Все результаты, представленные далее в терминах символьного ИИ, формально справедливы и в задачах «обучения» LLM-моделей.

2. Предложенный далее алгоритм поиска латентного пространства минимальной размерности, возможно, является альтернативой модели трансформера.

3. Отображение модели данных из пространства эмбеддингов в графовую модель данных и ее реверс-инжиниринг по полученной графовой модели можно рассматривать в качестве мостика для синтеза сервиса объяснения решений LLM с терминах символьного ИИ.

Уместные примечания и выводы

Следует различать задачи поиска латентного пространства и поиска латентных переменных. Соответственно, можно использовать две стратегии решения:

1. Отыскивать сначала латентное пространство, а затем строить в нем ортонормированный базис с требуемыми экстремальными свойствами, как правило, базис главных компонент.

2. Находить сразу латентные переменные, которые при этом автоматически определяют и латентное пространство (как в задачах редукции пространства данных).

5. Оптимизационная постановка задачи поиска латентного пространства

Пространство данных с W-метрикой – это пространство скрытых связей на множестве атрибутов экземпляров выборки данных

Введем в векторном пространстве данных скалярное произведение произвольной пары случайных величин $h = H^T X$ и $g=G^T X$, такие, что , $H,G \in R^m$ по формуле:

$$(h, g)=M[(h-\bar{h}, g-\bar{g})]=H^T \boldsymbol{W} G$$

где M [*] – м.о. случайной величины в квадратных скобках, равное ковариации случайных величин h и g [Рао, 1968], а норму случайной величины h как ее СКО:

$$\|h\|=M[(h,h)]^{1/2}=(H^T \boldsymbol{W} H)^{1/2} = \sigma(h).$$

Это пространство является эвклидовым и называется пространством с W-метрикой [Гантмахер]. Далее будем обозначать его символом $\boldsymbol{M}_W(X)$.

Пусть $[\boldsymbol{Y},\boldsymbol{Z}]$ – произвольное разбиение **всего** множества $\boldsymbol{X}=\{x_1, \dots, x_m\}$ пространства данных $\boldsymbol{M}_W(X)$, в котором первый элемент – это подмножество известных координат множества $\boldsymbol{X}$, а второй – подмножество остальных его координат, линейные оценки которых могут/должны быть вычислены как функции значений координат из множества $\boldsymbol{Y}$.

На множестве всех разбиений введем функционал F , $F([Y, Z|Y]) \geq 0$, который каждому разбиению $[\boldsymbol{Y},\boldsymbol{Z}]$ ставит в соответствие неотрицательное число, характеризующее *качество оценивания* координат вектора Z при заданной реализации вектора Y.

Функционал качества разбиения $([\boldsymbol{Y}, \boldsymbol{Z}])$

$$F\ ([\boldsymbol{Y}, \boldsymbol{Z}]) = F\ ([Y, Z|Y]) = M[(Z|\mathrm{Y} - \bar{Z}|Y)^{\boldsymbol{T}} A_{ZZ}(Z|\mathrm{Y} - \bar{Z}|Y)], \qquad (1)$$

где $Z|\mathrm{Y}$ – вектор с. в. с условным распределением вероятностей вектора Z при известной реализации вектора Y; $\bar{Z}|Y$ – вектор его эмпирического среднего, $M[*]$ – эмпирическое среднее квадратичной функции в квадратных скобках и $\boldsymbol{A}_{ZZ}$ – неотрицательная симметричная матрица весов. Ее выбор зависит от варианта использования оценки $\hat{Z}|\mathrm{Y}$ в приложении.

Если $\boldsymbol{Y} = \emptyset$, то

$$F^{(0)} = F\ ([\emptyset, \boldsymbol{X}])\ = M[(X - \bar{X})^T \boldsymbol{A}_{XX}(X - \bar{X})] = \sum_{i,j=1}^{m} a_{ij} \times w_{ij}. \qquad (2)$$

Далее $\bar{X} = 0$, т.к. значения функционалов в (1) и (2) не зависят от м. о $\bar{Z}|Y$ и $\bar{X}$ [Рао].

Оптимизационная постановка задачи поиска латентного подпространства в пространстве данных $\widetilde{\boldsymbol{M}}(X)$

Определение 1. Разбиение $[\boldsymbol{U}, \boldsymbol{Z}]$ является *допустимым*, если для него выполнено

$$\boldsymbol{U} \subseteq \boldsymbol{V} \subseteq \boldsymbol{X}, \boldsymbol{U} \cup \boldsymbol{Z} = \boldsymbol{X}, \boldsymbol{U} \cap \boldsymbol{Z} = \emptyset, F([\boldsymbol{U}, \boldsymbol{Z}]) \leq, \Delta . \geq 0. \qquad (3)$$

Определение 2. Допустимое разбиение $([\boldsymbol{Y}, \boldsymbol{Z}])$ называется *оптимальным*, если в нем множество $\boldsymbol{Y}$ имеет минимальную мощность среди всех допустимых разбиений:

$$[\boldsymbol{Y}, \boldsymbol{Z}] = argmin_{\boldsymbol{U}} \{|U|: [\boldsymbol{U}, \boldsymbol{Z}] - \text{допустимое разбиение множества } \boldsymbol{V}, \qquad (4)$$

Определение 3. Подпространство $\boldsymbol{M}_W(\mathrm{Y}) \subseteq \boldsymbol{M}_W(X)$, порождаемое множеством всех линейных функционалов от координат подмножества данных $\boldsymbol{Y}$, полученного решением задачи (4), назовем *латентным* (под)пространством пространства $\boldsymbol{M}_W(X)$ при ограничениях (3).

Эти определения совместно вводят понятие латентного пространства в рамках решаемой задачи, а выражения (3) и (4) представляет формальную постановку задачи его поиска.

6. Алгоритм поиска латентного пространства

Метод решения: Обозначения

Пусть, как и ранее, $[\boldsymbol{Y}, \boldsymbol{Z}]$ –некоторое разбиение координат множества $\boldsymbol{X}$, а Y и Z – векторы, такие, что Y включает координаты из множества $\boldsymbol{Y}$, а Z – из множества $\boldsymbol{Z}$, причем порядок следования координат в векторах Y и Z тот же, что и в исходном векторе X.

Пусть $X^{(1)}$ - блочное представление вектора X с учетом разбиения $[\boldsymbol{Y}, \boldsymbol{Z}]$, и тогда

$$X^{(1)} = [(Y^{(1)})^T (Z^{(1)})^T]^T,$$

и тогда блочные представления векторов м.о. и матриц ковариаций и весов будут иметь вид

$$\bar{X}^{(1)} = \begin{bmatrix} \bar{Y}^{(1)} \\ \bar{Z}^{(1)} \end{bmatrix}, \boldsymbol{W}^{(1\text{и})} = \begin{bmatrix} \mathbf{W}_{YY}^{(1)} & \mathbf{W}_{YZ}^{(1)} \\ \mathbf{W}_{ZY}^{(1)} & \mathbf{W}_{ZZ}^{(1)} \end{bmatrix}, \boldsymbol{A}^{(1)} = \begin{bmatrix} \boldsymbol{A}_{YY}^{(1)} & \boldsymbol{A}_{YZ}^{(1)} \\ \boldsymbol{A}_{ZY}^{(1)} & \boldsymbol{A}_{ZZ}^{(1)} \end{bmatrix}. \quad (5)$$

Важное примечание

В принятых предположениях о модели данных, представленной в евклидовом пространстве с W-_метрикой линейное подпространство $\boldsymbol{M}_W(Z|Y)$, всегда ортогонально подпространству $\boldsymbol{M}_W(Y)$, т. е . $F([Y, Z|Y]) = F(Z|Y])$ и его величина статистически не зависит от множества случайных величин пространства $\boldsymbol{M}_W(Y)$. Это свойство Л.П. автоматически учитывает скрытые структуры связей на множестве атрибутов данных выборки (через вклад значение функционала).

Характеристики условного распределения вероятностей случайного вектора $Z^{(1)}|\bar{\bar{Y}}^{(1)}$ при $Y=\bar{\bar{Y}}$:

Напомним, что характеристики условного распределения вероятностей вектора $Z^{(1)}|\bar{\bar{Y}}^{(1)}$ при $Y=\bar{\bar{Y}}$:

$$Z^{(1)}|\bar{\bar{Y}}^{(1)} = Z^{(1)} - \boldsymbol{W}_{\mathrm{ZY}}\mathbf{W}_{YY}^{-1}\left(\bar{\bar{Y}}^{(1)} - Y^{(1)}\right), \tag{6}$$

$$M[Z^{(1)}|\bar{\bar{Y}}^{(1)}] = \bar{Z}^{(1)} - \mathbf{W}_{ZY}^{(1)}(\mathbf{W}_{YY}^{(1)})^{-1}(\bar{\bar{Y}}^{(1)} - \bar{Y}^{(1)}); \tag{7}$$

$$\mathbf{W}_{ZZ}^{(1)}(Z^{(1)}|\bar{\bar{Y}}^{(1)}) = \mathbf{W}_{ZZ}^{(1)} - \mathbf{W}_{ZY}^{(1)}(\mathbf{W}_{YY}^{(1)})^{-1}\mathbf{W}_{YZ}^{(1)}\ . \tag{8}$$

Теорема об аддитивном представлении квадратичного функционала в пространстве *с* ***W***-метрикой

Теорема. Пусть $[\boldsymbol{Y}, \boldsymbol{Z}]$ – разбиение вектора с. в. $X \sim \boldsymbol{N_m}$ с нулевым вектором м. о., заданного в евклидовом пространстве $\boldsymbol{M}_W(X)$ с ***W***-метрикой, и вектор с. в. Y размерности $k \leq m$ имеет *неособенную* матрицу ковариаций $\boldsymbol{W}_{\mathrm{YY}}$.

Тогда квадратичный функционал $F^{(0)}$, введенный в формуле (2) для разбиения $[\boldsymbol{Y}, \boldsymbol{Z}]$, представим в *аддитивной* форме

$$F^{(0)} = F([Y, Z|Y]) + \sum_{i=1}^{k} \lambda_i \,, \qquad (9)$$

где $F^{(0)}$ вычисляется согласно (2), а $\lambda_1 \geq \ldots \geq \lambda_k \geq 0$ –собственные числа регулярного пучка квадратичных форм, представленного матрицей Ω:

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda} - \lambda \boldsymbol{W}_{\mathrm{YY}}, \lambda \in R^{(1)}, \qquad (10)$$

$$\boldsymbol{\Lambda} = \boldsymbol{W}_{\mathrm{YY}} \boldsymbol{A}_{\mathrm{YY}} \boldsymbol{W}_{\mathrm{YY}} + \boldsymbol{W}_{\mathrm{YY}} A_{YZ} \boldsymbol{W}_{\mathrm{ZY}} + \boldsymbol{W}_{YZ} A_{ZY} \boldsymbol{W}_{\mathrm{YY}} + \boldsymbol{W}_{YZ} A_{\mathrm{ZZ}} \boldsymbol{W}_{ZY}. \qquad (11)$$

$$F^{(0)} = F([\emptyset, \boldsymbol{X}]) = M[(X - \bar{X})^T \boldsymbol{A}_{XX} (X - \bar{X})] = \sum_{i,j=1}^{m} a_{ij} \times w_{ij}.$$

Точный алгоритм поиска латентного пространства полиномиальной сложности: Итерация1

Пусть подмножество $\boldsymbol{Y \subseteq V \subseteq X}$ в разбиении $[\boldsymbol{Y}, \boldsymbol{Z}]$ является *одноэлементным*. Тогда по ***Теореме*** вклад любой координаты $x_i \in \boldsymbol{V}$, в величину функционала $F^{(0)}$ будет равен собственному числу $\lambda(x_i)$, которое в этом случае вычисляется просто:

$$\boldsymbol{\lambda}(x_i) = \lambda_{x_i} / w_{x_i x_i}, \text{ где } \lambda_{x_i} = \boldsymbol{W}_{x_i}^T \mathrm{A}\, \boldsymbol{W}_{x_i}, \qquad (12)$$

где $w_{x_i x_i}$ – диагональный элемент матрицы $\boldsymbol{W}$ на позиции x_i, а $\boldsymbol{W}_{x_i}$ – столбец этой же матрицы на позиции координаты x_i. Из (12) следует, что разбиение $[\boldsymbol{Y} = \{x_{s_1}\}, \boldsymbol{Z} = \boldsymbol{X} \backslash x_{s_1}]$, минимизирующее слагаемое $F\ ([Y, Z|Y])$ в правой части (9), находится решением задачи (4):

$$x_{s_1} = argmax_{x_{i_l}: x_{i_l} \in \boldsymbol{V}, \boldsymbol{w}_{x_{i_l} x_{i_l}} \neq 0} \{\boldsymbol{\lambda}(x_{i_1}), \ldots, \lambda(x_{i_p})\}, \qquad (13)$$

где p – число ненулевых элементов на диагонали матрицы $\boldsymbol{W}$, отвечающих подмножеству $\boldsymbol{V}$.

Пусть решение уравнения (13) дается координатой x_{s_1}, а разбиение $[\{x_{s_1}\}, \boldsymbol{X} \backslash x_{s_1}]$ характеризуется парой $<x_{s_1}, \lambda(x_{s_1})>$, т.е. координатой x_{s_1}, отвечающей оптимальному выбору, и величиной ее вклада $\lambda(x_{s_1})$ в значение функционала (9). Расчеты по формулам (12), (13) для всех $x_i \boldsymbol{V}$ имеют *полиномиальную сложность* $O(Cr^3)$ от мощности r множества $\boldsymbol{V}$, $r \leq m$.

Подготовка данных к следующей итерации

Результаты вычислений на первой итерации по формулам 12), (13) – пара $< x^{(1)}, \lambda^{(1)} >$, где $x^{(1)} = x_{s_1}$ и $\lambda^{(1)}=\lambda(x_{s_1})$, и

$$F^{(0)} = F\left([Y^{(1)}, Z^{(1)}|Y^{(1)}]\right) + \lambda^{(1)} = F(x^{(1)}, Z^{(1)}|x^{(1)}) + \lambda^{(1)} = F^{(1)} + \lambda^{(1)}, \qquad (14)$$

где $F^{(1)} = F^{(0)} - \lambda^{(1)}$ точность оценки координат вектора $Z^{(1)}$ при известном значении $x^{(1)}$, а слагаемое $F^{(1)}$ в (14) определяется условным распределением вероятностей вектора $Z^{(1)}|x^{(1)}$. Координата $x^{(1)}$ переводится в состав координат латентного подпространства $\boldsymbol{Y}^{(1)} = \{x^{(1)}\}$.

После первой итерации вводим обозначения $X^{(2)} = Z^{(1)}|x^{(1)}$, $\bar{X}^{(2)} = \bar{Z}^{(1)}|x^{(1)}$, и $\boldsymbol{W}^{(2)}$ – матрица ковариаций условного распределения вероятностей вектора $Z^{(1)}|x^{(1)}$ и $X^{(2)}$ содержит все координаты исходного множества $\boldsymbol{X}^{(1)}$ без координаты $\{x^{(1)}$, и

$$\bar{X}^{(2)} = M[Z^{(1)}|\bar{\bar{Y}}^{(1)}] = \bar{Z}^{(1)} - \mathbf{W}_{ZY}^{(1)}(\mathbf{W}_{YY}^{(1)})^{-1}(\bar{\bar{Y}}^{(1)} - \bar{Y}^{(1)}); \qquad (7)$$

$$\mathbf{W}_{ZZ}^{(1)}(Z^{(1)}|\bar{\bar{Y}}^{(1)}) = \mathbf{W}_{ZZ}^{(1)} - \mathbf{W}_{ZY}^{(1)}(\mathbf{W}_{YY}^{(1)})^{-1}\mathbf{W}_{YZ}^{(1)}\ . \qquad (8)$$

Ситуация полностью идентична ситуации в начале первой итерации, есть все данные для писка очередной координаты латентного пространства!

Условия окончания итераций и итоговые результаты

На некотором шаге с номером k будет получена пара $< x^{(k)}, \lambda^{(k)} >$, а значение функционала $F^{(0)}$ в аддитивной форме будет иметь вид :

$$F^{(0)} = F\left([Y^{(k)}, Z^{(k)}]\right) + \sum_{i=1}^{k} \lambda^{(i)}, = F^{(k-1)} + \sum_{i=1}^{k} \lambda^{(i)},$$

где величина $F^{(k)} = F^{0} - \sum_{i=1}^{k} \lambda^{(i)} = F^{(k-1)} - \lambda^{(k)}$ определяет точность оценивания атрибутов вектора $Z^{(k)}$ при известных значениях множества атрибутов $\boldsymbol{Y}^{(k)} = \{ x^{(1)}, \ldots, x^{(k)} \}$, которые после выполнения k итераций будут включены в число координат латентного подпространства, а значение $\lambda^{(k)}$ равно снижению ошибки искомой оценки на шаге с номером k.

Итерации выполняются, пока значение $F^{(k)}$ остается недопустимым. Если после *r*-ой итерации оно станет допустимым, то вычисления заканчиваются, и текущее множество $\boldsymbol{Y}^{(r)} = \{ x^{(1)}, \ldots, x^{(r)} \}$ определяет латентное подпространство с *ортогональным* базисом вида:

$$\begin{aligned} x_1^{(1)} &= a_{11} x_{s_1}, \\ x_2^{(2)} &= a_{21} x_{s_1} + a_{22} x_{s_2} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_r^{(r)} &= a_{r1} x_{s_1} + a_{r2} x_{s_2} + \cdots + a_{rr} x_{s_r}, \end{aligned} \qquad (15)$$

Итоговые замечания о полученных результатах

О полученных результатах

1. Переменные $x_1^{(1)}$, …, $x_r^{(r)}$ задают ортонормированный базис латентного пространства, но не являются латентными переменными. Они представлены как линейные комбинации координат исходного пространства $\widetilde{\boldsymbol{M}}^{(1)}$ данных ***c W***-метрикой.

2. Переход к базису латентных переменных выполняется с помощью ортогонального преобразования построенного базиса $\{x_1^{(1)}, \ldots, x_r^{(r)}\}$, т.е. преобразования подобия, и потому сумма «вкладов» найденных координат из множества $\{x_1^{(1)}, \ldots, x_r^{(r)}\}$ не изменится в базисе латентных переменных, но перераспределится между латентными переменными иначе.

3. Если допустимое решение не существует, то на очередной итерации с номером k множество допустимого выбора в задаче (13)

$$x_{s_k} = argmax_{x_{i_l}: x_{i_l} \boldsymbol{V}^{(k(}, \boldsymbol{w}_{x_{i_l} x_{i_l}} \neq 0} \{\boldsymbol{\lambda}(x_{i_1}), \ldots, \lambda(x_{i_p})\}$$

оказывается пустым, и при этом решение $\boldsymbol{Y}^{(r)} = \{ x^{(1)}, \ldots, x^{(r)} \}$ остается недопустимым. Тогда текущее значение функционала $F^{(k)}$ определяет предельно достижимую точность оценивания координат вектора $Z^{(k)}|Y^{(k)}$. В этом случае описанный процесс вычислений заканчивается тоже за *полиномиальное* время.

Итоговые замечания о построенном алгоритме

1. Если в задаче требуется еще найти латентные переменные, то в этом случае в латентном подпространстве нужно найти ортонормированный базис, в котором обе матрицы $\boldsymbol{\Lambda}$ и $\boldsymbol{W}_{\text{YY}}$ в (10)

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda} - \boldsymbol{\lambda} \boldsymbol{W}_{\text{YY}}, \boldsymbol{\lambda} \in R^{(1)}, \qquad (10)$$

одновременно приводятся к диагональному виду, в котором матрица Λ приводится к главным осям [Гантмахер]. Это типовая задача линейной алгебры сложности $O(r^3)$, где r – размерность латентного подпространства.

2. Описанный итерационный алгоритм, использующий формулы (7), (8), аналогичен алгоритму Грама-Шмидта ортогонализации базиса евклидова пространства, но с одним важным отличием. Это отличие состоит в том, что если в классическом варианте выбор очередной координаты определяется только порядком следования координат в исходном векторе X, то в предложенном алгоритме этот порядок определяется динамически условием (13).

Иллюстративный пример (1/2)

Пусть $M_W(X)$ – эвклидово пространство данных с W-метрикой с координатами из множества $X = \{ x_1 \ \ldots, x_m\}$, m=6 и с заданными матрицами ковариаций и весов

$$W^{(1)} = \begin{bmatrix} 0,0596 & -0,1015 & -0,0085 & 7,0010 & -10,114 & -0,4516 \\ -0,1015 & 0,2018 & 0,0229 & -11,708 & 22,080 & 1,7967 \\ -0,0085 & 0,0229 & 0,2473 & -0,6567 & 2,0759 & 28,219 \\ 7,0010 & -11,708 & -0,6567 & 870,50 & -1329,1 & -24,889 \\ -10,114 & 22,080 & 2,0759 & -1329,1 & 2502,8 & 171,27 \\ -0,4516 & 1,7967 & 28,219 & -24,889 & 171,27 & 3417,2 \end{bmatrix}$$

$Diag[A]$= [2,0; 1,5; 1.0; 0,00001; 0,0000025; 0,0000001].

с порогом относительной ошибки Δ=0,05.

Табл. 1. Результаты решения поиска лптентного пространства

	Номер итерации	$L(x_i^k)$						
		$\lambda(x_1^{(k)})$	$\lambda(x_2^{(k)})$	$\lambda(x_3^{(k)})$	$\lambda(x_4^{(k)})$	$\lambda(x_5^{(k)})$	$\lambda(x_6^{(k)})$	$\Delta^{(k)}$
1	k=1	0,3885,	**0,4201**	0,2514	0,3631	0,3890	0,2349	0,3809
2	k=2	0,0200	---------	**0,2448**	0,0172	0,0252	0,2312	0,0202

Иллюстративный пример (2/2)

Начальные условия второй итерации при $x^{(1)} \rightarrow x_2$ и $\lambda^{(1)} = 0{,}4201$, $\Delta^{(1)} =),3809$:

$$W^{(2)} = \begin{bmatrix} 0{,}0085 & -0{,}0030 & 1{,}11219 & 0{,}9916 & 0{,}4521 \\ -0{,}0030 & 0{,}2447 & 0{,}6719 & -0{,}4297 & 28{,}0151 \\ 1{,}1122 & 0{,}6719 & 191{,}227 & -48{,}0665 & 79{,}3516 \\ 0{,}9916 & -0{,}4297 & -48{,}0665 & 86{,}9117 & -25{,}3163 \\ 0{,}4521 & 28{,}0151 & 79{,}3516 & -25{,}3163 & 3\ 401{,}20 \end{bmatrix}$$

$Diag[A^{(2)}] = [\ 2{,}0;\quad 1.0;\quad 0{,}00001;\quad 0{,}0000025;\quad 0{,}0000001\ ].$

Начальные условия третьей итерации при $x^{(2)} \rightarrow x_3$ и $\lambda^{(2)} = 0{,}2428$, $\Delta^{(2)} = 0{,}0202$:

$$W^{(3)} = \begin{bmatrix} 0{,}008463 & 1{,}103953 & 0{,}996868 & 0{,}108637 \\ 1{,}103953 & 189{,}38209 & -46{,}88712 & 2{,}427425 \\ 0{,}996868 & -46{,}88712 & 86{,}15714 & 23{,}87899 \\ 0{,}108637 & 2{,}427425 & 23{,}87899 & 193{,}82024 \end{bmatrix}$$

$Diag[A^{(3)}] = [2{,}0;\quad 0{,}00001;\quad 0{,}0000025;\quad 0{,}0000001\ ].$

Окончание итераций, т.к. $\Delta^{(2)} = 0{,}0202 < \Delta = 0{,}0500.$

Результаты: 1. координаты латентного пространства $\{x^{(1)}, x^{(2)}\} \rightarrow \{x_2, x_3\}$;
2. Случайная оценка координат множества $\{\{x_1, x_4, x_5, x_6\}$ имеет нормальное распределение вероятностей с матрицей ковариаций $W^{(3)}$

7. Приложения в задачах обработки больших данных

Редукция данных: Для чего она нужна? (1/3)

Если размерность пространства атрибутов m превышает определенный порог, [Горбань А.Н., 2021], то задача редукции атрибутов данных обязательна по ряду причин: .

- Объем выборки недостаточен для надёжного оценивания эмпирических вектора средних и матрицы ковариаций, а их ошибки могут повлечь сильное искажение реальных зависимостей в данных, представленных этими оценками, вводя потребителя в заблуждение.
- Если не выполнять редукцию, то машинное обучения в полном пространстве атрибутов будут требовать нереально больших ресурсов процессора и памяти.
- При большой размерности пространства атрибутов даже точно вычисленная матрица ковариаций будет особенной или плохо обусловленной. И вычислительные проблемы в итоге будут столь серьёзными, что случайный выбор редуцированного пространства может оказаться намного удачнее его оптимизации с помощью любых методов.
- Иногда используют предварительно «низко ранговое» представление матрицы ковариаций (ее факторизация), а потом – редукцию пространства.

Технологии работы с большими данными всегда принимают все меры для того, чтобы снизить размерность данных, вовлекаемых в обработку.

Редукция данных: наиболее известные традиционные подходы (2/3)

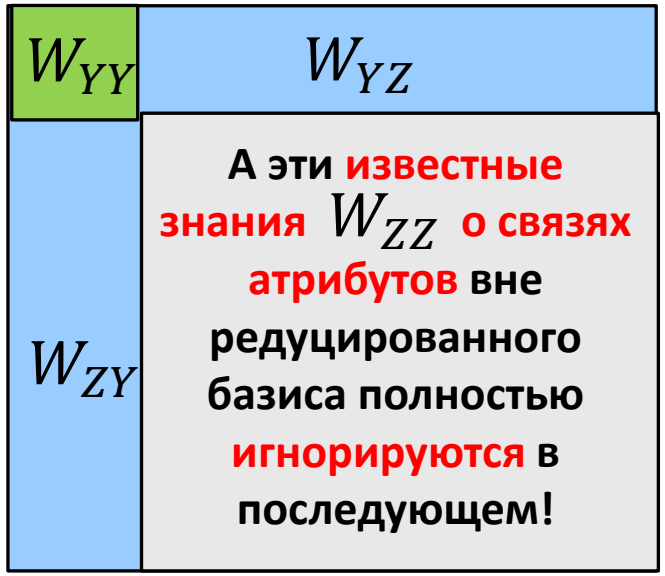
Базовый принцип всех подходов -отказ от оптимальности в пользу снижения вычислительной сложности, и выбор компромисса между ними (принцип ограниченной рациональности).

1. Использование методов компонентного анализа. Они имеют сложность $O(m^2 \times N + m^3)$, но в них используются линейные комбинации всех атрибутов вектора данных, и их нужно знать. Проблемы вычислительной сложности и неустойчивости здесь частично остаются.
2. Методы, ориентированные на сохранение структуры связей в редуцированной модели. Пример – метод случайных проекций. В нем решается задача поиска k-мерного подпространства атрибутов, в котором эвклидово расстояние между всеми разными парами точек пространства не изменяется. Приближенно это достигается тогда, когда столбцы матрицы преобразования имеют единичную эвклидову норму (она используется вместо матрицы проектирования). Гипотеза авторов (частично подтвержденная экспериментами) состоит в том, что любое случайно выбранное подмножество векторов с единичной нормой с большой вероятностью будет множеством ортогональных векторов.
3. Большая группа методов класса LASSO (Least Absolute Shrinkage and Selection Operator). В нем минимизируется квадратичная невязка плюс дополнительный член, который зависит от суммы модулей искомых коэффициентов линейной регрессии с некоторым коэффициентом, регуляризации. В процессе минимизации некоторые коэффициенты линейной регрессии обращаются в нуль, остаются в ней наиболее информативные.

Редукция данных: Решение на основе алгоритма поиска латентных пространств (3/3)

Алгоритм поиска латентного пространства (ЛП-алгоритм), по существу, решает классическую задачу редукции данных. Он гарантирует получение точного решения, если оно существует, и обладает полиномиальной сложностью, хотя проблема редукции теоретически является ***NP*-трудной**. В этой роли построенный алгоритм может найти широкое применение, т.к. имеется много разных приложений, где центральная задача – поиск оптимального базиса. То же самое касается и поиска латентных переменных: задачи причинного анализа и построение объяснений, факторный анализ при численном оценивании значений неформальных понятий, например, понятия «интеллект».

Основные преимущества редуцированного представления данных



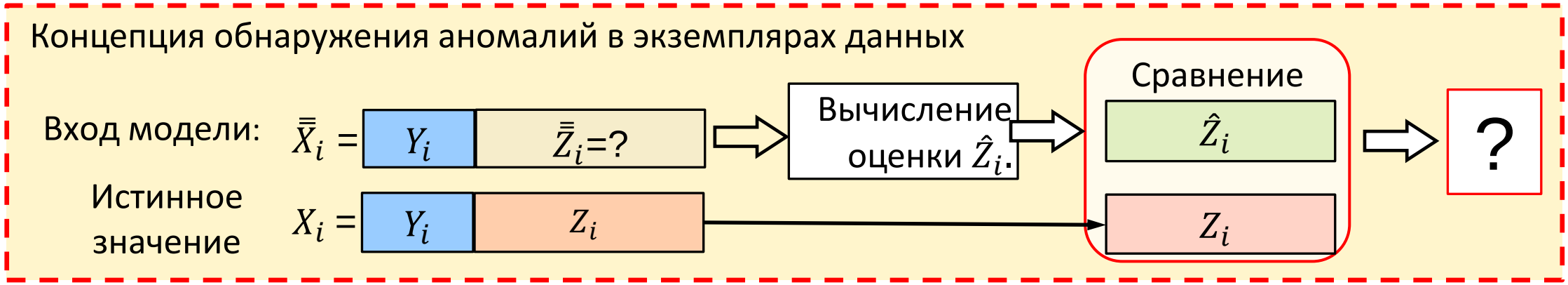
- снижение вычислительной сложности и
- релаксация проблем вычислительной неустойчивости

А все-таки, являются ли задачи редукции неизбежными??
Можно ли решать эффективно задачи обработки больших данных в полном пространстве, если это может оказаться намного лучше?
И если можно, то за счет чего?

За счет дополнительных знаний.

Дополнительные знания – это дополнительные возможности.

Алгоритм обнаружения аномалий в экземплярах данных на основе знаний о структуре связей на множестве атрибутов для улучшения качества данных (1/2)



1. ЛП-алгоритм позволяет находить оптимальное разбиение множества атрибутов данных на известную и оцениваемую (тестовую) части анализируемого экземпляра данных.
2. Для этого используется модель линейной регрессии – модель линейных статистических зависимостей на множестве атрибутов данных конкретных экземпляров выборки (**метамодель**). Метамодель вводит **новый** источник знаний, привлекаемых для анализа аномалий.
3. ЛП-алгоритм позволяет строить **все варианты** разбиений атрибутов данных, что дает возможность строить систематические поисковые алгоритмы локализации аномалий.
4. Концептуальная идея обнаружения аномалий на основе разбиения множества атрибутов данных применима к данным с пропущенными значениями.

Обнаружение аномалий в экземплярах данных на основе знаний о структуре связей на множестве атрибутов для улучшения качества данных (2/2)

1. Анализ качества данных и принятие мер к его повышению – это основная цель предварительной их обработки, и эта задача всегда требует больших затрат ресурсов. Среди различных аспектов качества данных ключевую роль играет их безошибочность.

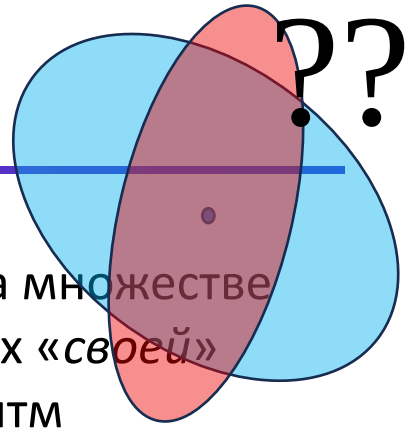
Использование метамодели линейных статистических связей на множестве атрибутов данных выборки фактически вводит новый класс алгоритмов анализа и повышения качества данных, а именно класс алгоритмов на основе знаний.

2. Алгоритм обладает высокой чувствительностью к нарушениям связей на множестве атрибутов и потенциально обладает возможностями обнаруживать ошибочные данные, которые не идентифицируются современными программными инструментами.

3. Алгоритм позволяет идентифицировать неверную разметку данных, а также обнаруживать неверные данные, которые появились как следствие атак на данные .

4. Потенциально алгоритм может быть частью алгоритма идентификации семантических ошибок.

Робастные алгоритмы классификации – особенности и достоинства (1/2)



Алгоритм обнаружения аномалий по своей сути является алгоритмом бинарной классификации. Данные разных выборок будут иметь различные модели связей на множестве атрибутов экземпляров данных. Поэтому он способен отличать экземпляры данных «*своей*» выборки от экземпляров всех других классов (иллюстрация). Обсуждаемый алгоритм классификации относиться к типу одно-классовых алгоритмов. Они выдают бинарные решения из множества <*свой, чужой*>.

Такой алгоритм классификации работает по принципу анализа совместимости частей любого экземпляра данных с линейными статистическими связями, присущими атрибутам конкретной выборки. В этом он является аналогом механизмов анализа совместимости в иммунной системе живого организма, и потому его можно называть алгоритмом **цифрового иммунитета**.

Достоинства алгоритма

1. Практически независимее обучение классификаторов.
2. Удобен в задачах с произвольным числом классов, поскольку любое изменение состава классификаторов системы не потребует переобучения оставшихся.
3. Позволяет строить *робастные* алгоритмы классификации за счет использования информационной и алгоритмической избыточности, поэтому может работать с неполными данными.

Робастные алгоритмы классификации – источники избыточности (2/2)

1. В больших данных число координат, не относящихся к редуцированному пространству, много больше мощности редуцированного пространства. Поэтому обычно существует много разбиений $[\boldsymbol{Y}, \boldsymbol{Z}]$, допустимых по значению функционала $F([\boldsymbol{Y}, \boldsymbol{Z}])$. Поэтому можно построить достаточно большой ансамбль классификаторов, решающих проблему совместимости.
2. Матричное уравнение линейной регрессии для разбиения $[\boldsymbol{Y}, \boldsymbol{Z}]$ при классификации некоторого экземпляра данных $X_i^{(1)} = [x_1^{(i)}, \ldots, x_m^{(i)}]^T$ для его блочного представление, отвечающего конкретному разбиению, в векторно-матричной

$$\hat{Z} = \bar{Z} - \boldsymbol{W}_{ZY}(\boldsymbol{W}_{YY})^{-1}(\bar{\bar{Y}} - \bar{Y}),$$

фактически задает множество правил вида

$$\text{Если } Y = \bar{\bar{Y}}, \text{то для } \forall x_j^{(i)} \in \boldsymbol{Z} \;\; \hat{x}_j^{(i)} \approx \bar{z}_{x_j} - W_{x_jY} \; (\boldsymbol{W}_{YY})^{-1}(\bar{\bar{Y}} - \bar{Y}),$$

где W_{x_jY} есть строка матрицы $\boldsymbol{W}_{ZY}$, отвечающая атрибуту x_j.Число таких правил равно$|\mathbf{Z}|$.

3. Общее количество таких правил может исчисляться сотнями, и при выбранном пороге различия справа каждое из них будет иметь контролируемую точность (вероятность, сходство), Это обеспечивает большую алгоритмическую и информационную избыточность и , значит, высокую точность мета классификации.

Обнаружение аномалий, классификация и предсказание гауссовских процессов

Многомерные временные ряды –специальная и сложная область исследований в ИТ и ИИ.

Дополнительные обозначения:

$X(t_i)$ – гауссовский m-мерный процесс, заданный на дискретной шкале времени, и $\bar{X}(t_i)$ и $\boldsymbol{W}(t_i, t_j)$ – характеристики его распределения вероятностей; $\boldsymbol{X}(t_i)$ – *множество* координат этого процесса; t_i, t_j, $t_i < t_j$, $X(t_i)$ и $X(t_j)$ – два момента времени и векторы случайных процессов в эти моменты; $U(t_i, t_j)$ – вектор с. в. размерности $2m$, составленный из координат векторов $X(t_i)$ и $X(t_j)$; $\bar{U}(t_i, t_j)$ и $\boldsymbol{W}_{UU}(t_i, t_j)$ – *эмпирические оценки* вектора м. о. и матрицы ковариаций вектора $U(t_i, t_j)$; Пусть $[\boldsymbol{Y}(t_i), \boldsymbol{Z}(t_i, t_j)]$ – разбиение вектора $U(t_i, t_j)$, такое, что подмножество $\boldsymbol{Y}(t_i)$ включает подмножество атрибутов, относящихся к моменту t_i , а $\boldsymbol{Z}(t_i, t_j)$ – все его остальные атрибуты. Как и ранее, обозначим через $Y(t_i)$ и $Z(t_i, t_j)$) векторы, составленные из атрибутов множеств $\boldsymbol{Y}(t_i)$ и $\boldsymbol{Z}(t_i, t_j)$. Тогда матрица ковариаций $U(t_i, t_j)$ в блочном виде будет состоять из матриц $W_{YY}(t_i, t_i)$, $W_{ZY}(t_i, t_j), W_{YZ}(t_i, t_i)$ и $W_{ZZ}(t_i, t_j)$. Пусть $F([\boldsymbol{Y}(t_i), \boldsymbol{Z}(t_i, t_j)])$ – функционал, оценивающий разбиения.

Тогда постановка задачи поиска оптимального разбиения $[\boldsymbol{Y}(t_i), \boldsymbol{Z}(t_i, t_j)]$ и последующего обнаружения аномалий с точностью до обозначений совпадает с постановкой задачи для статического случая, рассмотренной ранее.

Обзор других приложений ЛП-алгоритма и алгоритма обнаружения аномалий (1/2)

1. Объяснимый искусственный интеллект.

ЛП- алгоритм предлагает прозрачный механизм принятия решений и выявления наиболее значимых аргументов (объяснений) в пользу предлагаемого решения. Кроме того, он позволяет выполнять поиск наиболее значимых факторов пространства состояний объекта, которые являются опасными для объекта или, наоборот, желательным состояниями.

2. Принятие решений по существенно неполной информации

В ряде критических приложений вход решателя является непредсказуемым по составу означенных атрибутов. Предположим, что эмпирическая оценка матрицы ковариаций вычислена для полного вектора данных. Тогда ЛП-алгоритм в совместно с алгоритмом обнаружения аномалий позволяет в онлайн-режиме найти в *текущем входе* подмножества атрибутов, ведущих к получению наиболее точных оценок значений других атрибутов, означенных во входе. В задаче с большим числом пропущенных значений эти оценки могут отличаться низкой точностью, Но, с другой стороны, в этом случае, как было описано ранее, можно построить много различных разбиений $[Y, Z]$ на множестве означенных атрибутов входа, которые позволят построить ансамбль алгоритмов принятия решений. Слияние их решений приведет к существенному улучшению итогового качества решений, полученных на основе «плохи» входных данных.

Обзор других приложений ЛП-алгоритма и алгоритма обнаружения аномалий (1/2)

3. Другие приложения

Область применения алгоритма обнаружения аномалий достаточно широка. Некоторые классы задач были указаны в начале презентации. Другие примеры приводятся ниже, хотя и ими тоже область возможных приложений не исчерпывается:

- задач восстановления пропущенных значений (на основ знаний), которая, как уже отмечалось, в недавнее время приобрела особую остроту.
- автоматическая разметка новых данных при наличии размеченных данных выборок всех классов проблемы (фактически как предварительная классификация);
- построение новых алгоритмов кластеризации, которые не опираются на гипотезу компактности данных кластера;
- задачи класса «сжатый сенсинг» (англ. *compressive sensing*); есть собственный опыт решения такой задачи.

Для всех этих и ранее названных приложений необходимо провести дополнительные исследования, в основном, экспериментального характера. Представляет также интерес исследование конкурентоспособности ЛП-алгоритма для работы с латентными пространствами LLM, в частности, его конкурентоспособности с моделью трансформера.

Заключение

Результаты работы можно оценивать с двух точек зрения – теоретической и прикладной.

1. Основной результат теоретического плана – это *полиномиальный* алгоритм решения *NP-* трудной задачи поиска латентных подпространств в линейных пространствах данных, для квадратичных функционалов невязки. Алгоритм построен на основе *Теоремы*, утверждения которой позволяют на каждой итерации строить аддитивное представление функционала.

2. С практической точки зрения построенный алгоритм обнаружения аномалий – это новый подход к решению ряда актуальных задач стека обработки больших данных, который имеет эффективную и масштабируемую алгоритмическую реализацию. Принципиальное отличие этого подхода от известных состоит в том, что в нем решатели вовлекают в процессы обработки и принятия решений *новый источник знаний*, а именно линейные статистические связи на множестве атрибутов данных выборки с общей меткой. Эти связи представляют специфические внутренние свойства, **структуры связей** всех экземпляров выборки.

История ИИ показывает, что качественные скачки в его развитии обычно связаны с привлечением новых источников знаний и инструментов доступа к ним. Именно это, например, объясняет фундаментальную роль больших данных и нейросетевых технологий в успехе прикладного ИИ в последнее десятилетие. Это же дает определенные основания автору надеяться, что новый источник знаний, который используется в предложенном подходе, а также разработанный алгоритм доступа к ним, будут полезными для дальнейшего развития технологий больших данных и ИИ в целом.