Информационные процессы, Том 20, № 4, 2020, стр. 342–356

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Моделирование многообразий в машинном обучении

Е.В.Бурнаев[*,**], **А.В.Бернштейн**[*]

[*] *Сколковский институт науки и технологий, Москва, Россия*
[**] *Институт проблем передачи информации, Российская академия наук*
e-mail: e.burnaev@skoltech.ru

Поступила в редколлегию 01.10.2020

Аннотация—Задачи предсказательного моделирования требуют обработки многомерных данных, и из-за "проклятия размерности" использование многих методов для их решения затруднено. В приложениях реальные данные зачастую занимают лишь малую часть пространства наблюдений, внутренняя размерность которого существенно ниже размерности самого пространства. Популярной моделью для таких данных является модель многообразия, в соответствии с которой данные лежат на неизвестном низкоразмерном многообразии (Data Manifold, DM), встроенном в окружающее высокоразмерное пространство. Задачи предсказательного моделирования, изучаемые в рамках этого предположения, называются задачами оценки многообразий, общей целью которых является обнаружение низкоразмерной структуры многомерных данных по заданной выборке точек. Если точки выборки порождаются в соответствии с неизвестной вероятностной мерой на DM, возникает потребность в моделировании многообразий при решении различных задач машинного обучения. В работе мы представим краткий обзор этих задач, и обозначим некоторые подходы к их решению.

КЛЮЧЕВЫЕ СЛОВА: снижение размерности, моделирование многообразий, предсказательное моделирование.

1. ВВЕДЕНИЕ

На протяжении многих веков анализ данных используется для обработки результатов наблюдений (измерений) реальных объектов, а также в натурных и вычислительных экспериментах. Достигнутый за последние десятилетия прогресс в информационных, вычислительных и телекоммуникационных технологиях обеспечивает возможность хранения, быстрого поиска и обработки больших массивов данных, а также быструю передачу данных по каналам связи и быстрый доступ к ним. Это привело к появлению так называемой парадигмы больших данных, в которой делается акцент на новых технических возможностях обработки данных большого объема и имеющих разные модальности. С учетом таких новых возможностей удалось сформулировать и решить ряд фундаментально новых научных и прикладных задач анализа данных, что, в свою очередь, привело к появлению новой теоретической области исследования, называемой наукой о данных (data science) [1]. Наука о данных использует многие математические и статистические инструменты для нахождения фундаментальных законов, которым подчиняются данные [2].

Феномен больших данных обычно включает в себя не только большие объемы данных, но также их высокую размерность [3]. К примеру, в задачах анализа изображений и машинного зрения изображение в градациях серого, заданное с разрешением $N \times N$ пикселей, представляется как N^2-мерный вектор, компоненты которого содержат в себе информацию об интенсивности цвета в пикселях изображения, при этом N изменяется по порядку от десяти до тысячи.

Аналогичные высокие размерности также возникают во многих прикладных областях, связанных с "интенсивным использованием данных" (распознавание речи, интеллектуальный анализ текстов, веб-поиск, и т.д.).

При больших размерностях данных многие теоретические и прикладные алгоритмы анализа данных показывают неудовлетворительные результаты из-за статистического и вычислительного "проклятья размерности" (например, проблема коллинеарности или "почти коллинеарности" многомерных данных приводит к известным трудностям в построении регрессии), "феномена пустого пространства", и т.п. [3].

К примеру, минимаксная ошибка в задаче оценки регрессии, в которой неизвестная функция, дифференцируемая по крайней мере s раз, и зависящая от d-мерного входного вектора, оценивается по n независимым наблюдениям, не может иметь скорость сходимости выше, чем $n^{-s/(2s+d)}$ (см. [4])) при использовании непараметрических оценок [5]. В задаче оценки плотности стандартные подходы (например, многомерный вариант ядерной оценки Парзена–Розенблатта) в d-мерном случае имеют среднеквадратичную ошибку порядка $O(n^{-4/(d+4)})$ [6].

К счастью, во многих приложениях, особенно в медицинских приложениях и приложениях, связанных с изображениями [7,8], "реальные" многомерные данные, полученные из "естественных" источников, могут занимать только малую часть пространства "наблюдений"; иными словами, внутренняя размерность "носителя данных" существенно ниже, чем размерность объемлющего пространства многомерных данных. Этот феномен приводит к тому, что многомерные данные могут быть трансформированы в их малоразмерные представления (признаки), которые затем используются в процедурах на основе данных со сниженной размерностью. Проблема нахождения таких представлений обычно называется "задачей снижения размерности": при заданном входном наборе данных

$$\mathbf{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathcal{X}, \tag{1}$$

сгенерированных в неизвестном пространстве данных (DS) $\mathcal{X} \subset \mathbb{R}^d$, найти "$n$-точечное" отображение вложения

$$\mathbf{C}_{(n)} \colon \mathbf{X}_n \subset \mathbb{R}^d \to \mathbf{Z}_n = \mathbf{C}_{(n)}(\mathbf{X}_n) = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\} \subset \mathbb{R}^q, \tag{2}$$

переводящее выборку $\mathbf{X}_n$ в q-мерный набор данных $\mathbf{Z}_n$ (выборку векторов признаков), $q < p$, которая "правильно представляет" выборку $\mathbf{X}_n$. Здесь размерность q или известна или оценивается по выборке (1). Термин "правильно представляет" в общем не формализуется и в разных методах снижения размерности он понимается по-разному из-за различного выбора оптимизируемой функции ошибок $L_{(n)}(\mathbf{Z}_n|\mathbf{X}_n)$, которая определяет "меру точности" для решения задачи снижения размерности и отражает требуемые свойства отображения $\mathbf{C}_{(n)}$ (2). Как отмечается в [9], общий подход к задаче снижения размерности может быть основан на "понятии функции ошибок".

Если пространство данных $\mathcal{X}$ является неизвестным q-мерным аффинным линейным подпространством L в $\mathbb{R}^d$, то метод главных компонент (Principal Component Analysis; PCA) [10,11] оценивает подпространство L q-мерным линейным подпространством L_{PCA}, которое минимизирует по L сумму норм "остатков" $\sum_{i=1}^n \|\mathbf{x}_i - \mathrm{Pr}_L(\mathbf{x}_i)\|^2$, где Pr_L – линейный проектор на L.

Подпространство L_{PCA}, которое проходит через выборочное среднее $\overline{\mathbf{x}}_n$, порождается q собственными векторами выборочной ковариационной матрицы $\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}}_n) \times (\mathbf{x}_i - \overline{\mathbf{x}})^\top$, соответствующими q наибольшим собственным значениям. Тогда координаты $\mathbf{z} \in \mathbb{R}^q$ проектора $\mathrm{Pr}_L(\mathbf{x})$ на q-мерное подпространство L_{PCA} берутся в качестве низкоразмерного представления вектора $\mathbf{x}$.

Однако если пространство данных $\mathcal{X}$ является нелинейным, то для решения задачи снижения размерности имеются лишь предложенные в XX в. эвристические нелинейные методы,

такие, как многомерное шкалирование (Multidimensional Scaling) [12], автокодировщики (разновидность нейронных сетей) [13], ядерный метод главных компонент (Kernel PCA) [14], и др. Отметим, что такие методы явно не основываются ни на какой математической модели обрабатываемых данных.

Впервые модель многомерных данных, называемая моделью многообразия [15], в которой данные занимают лишь часть пространства наблюдений $\mathbb{R}^d$, была описана в 2000 г. Практически сразу такая модель стала весьма популярной при исследовании многомерных данных. В модели предполагается, что данные лежат на неизвестном многообразии (многообразии данных, Data Manifold, DM) $\mathcal{M}$ или находятся рядом с ним, при этом размерность этого многообразия $q < d$ мала и оно вложено в исходное объемлющее пространство входных многомерных данных $\mathbb{R}^d$. Как правило, предположение о том, что данные лежат на многообразии, удовлетворяется для реальных многомерных данных, полученных из "натуральных" источников.

Различные задачи анализа данных, изучаемые при предположении о том, что данные лежат на многообразии (такие данные называются данными со значениями на многообразии), обычно известны под названием задач моделирования многообразий [16, 17], в которых общей целью является выявление и описание низкоразмерной структуры многомерных данных со значениями на многообразии, взятых из заданного набора обучающих данных $\mathbf{X}_n$ (см. (1)), сгенерированных на многообразии данных. Если точки набора данных выбираются на многообразии данных $\mathcal{M}$ независимо относительно некоторой неизвестной вероятностной меры μ, то мы приходим к статистическим задачам машинного обучения по данным со значениями на многообразии.

В работе дается краткий обзор задач машинного обучения на данных со значениями на многообразии. Статья организована следующим образом. В разделе 2 приводятся основные предположения об обрабатываемых данных, а в разделе 3 дается краткое описание типичных задач моделирования многообразий и их решений. В разделе 4 приведено заключение.

2. ПРЕДПОЛОЖЕНИЯ ОБ ОБРАБАТЫВАЕМЫХ ДАННЫХ

2.1. Предположения о многообразии данных

Пусть $\mathcal{M}$ – неизвестное "удобное для анализа" q-мерное многообразие данных, вложенное в объемлющее d-мерное пространство $\mathbb{R}^d$, $q \leqslant d$; предполагается, что внутренняя размерность q известна. Предположим, что многообразие данных является компактным римановым многообразием с положительным числом обусловленности [18]; т.е. на многообразии нет самопересечений ("коротких замыканий"). Для простоты мы предполагаем, что многообразие данных покрыто одной координатной картой $\boldsymbol{\varphi}$ и, следовательно, имеет вид

$$\mathcal{M} = \{\mathbf{x} = \boldsymbol{\varphi}(\mathbf{z}) \in \mathbb{R}^d : \mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^q\}, \tag{3}$$

где $\boldsymbol{\varphi}$ – однозначное отображение из открытой ограниченной координатной области $\mathcal{Z} \subset \mathbb{R}^q$ в многообразие $\mathcal{M} = \boldsymbol{\varphi}(\mathcal{Z})$ с обратным отображением $\boldsymbol{\psi} = \boldsymbol{\varphi}^{-1} \colon \mathcal{M} \to \mathcal{Z}$. Обратное отображение $\boldsymbol{\psi}$ определяет маломерную параметризацию многообразия данных $\mathcal{M}$ (q-мерные координаты или свойства $\boldsymbol{\psi}(\mathbf{x})$ точек многообразия $\mathbf{x}$), и отображение $\boldsymbol{\varphi}$ восстанавливает точки $\mathbf{x} = \boldsymbol{\varphi}(\mathbf{z})$ по их свойствам $\mathbf{z} = \boldsymbol{\psi}(\mathbf{x})$.

Отметим, что пара $(\boldsymbol{\varphi}, \mathcal{Z})$ в (3) определяется с точностью до произвольного однозначного отображения $\boldsymbol{\chi}$ из $\mathbb{R}^q$ в $\mathbb{R}^q$, при этом другая пара $(\boldsymbol{\varphi}^*, \mathcal{Z}^*)$, для которой $\boldsymbol{\varphi}^*(\mathbf{z}^*) = \boldsymbol{\varphi}(\boldsymbol{\chi}^{-1}(\mathbf{z}^*))$ и $\mathcal{Z}^* = \boldsymbol{\chi}(\mathcal{Z})$, задает другое представление $\mathcal{M} = \boldsymbol{\varphi}^*(\mathcal{Z}^*)$ многообразия $\mathcal{M}$ (3) и другие низкоразмерные признаки $\mathbf{z}^* = \boldsymbol{\psi}^*(\mathbf{x}) = \boldsymbol{\chi}(\boldsymbol{\psi}(\mathbf{x}))$ точек многообразия.

Если отображения $\boldsymbol{\psi}(\mathbf{x})$ и $\boldsymbol{\varphi}(\mathbf{z})$ являются дифференцируемыми (в $\boldsymbol{\psi}(\mathbf{x})$, $\mathbf{x} \in \mathcal{M}$ используется ковариантное дифференцирование), и если $\mathbf{J}_\psi(\mathbf{x})$, $\mathbf{J}_\varphi(\mathbf{z})$ – их матрицы Якоби размера $q \times d$

и $d \times q$, соответственно, то q-мерное линейное подпространство

$$\mathrm{L}(\mathbf{x}) = \mathrm{Span}(\mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x}))) \tag{4}$$

в $\mathbb{R}^d$ является касательным пространством к многообразию данных $\mathcal{M}$ в точке $\mathbf{x} \in \mathcal{M}$; здесь и далее $\mathrm{Span}(\mathbf{H})$ обозначает линейное подпространство, натянутое на столбцы произвольной матрицы $\mathbf{H}$, которые, в свою очередь, образуют базис для $\mathrm{L}(\mathbf{x})$. Такие касательные пространства рассматриваются как элементы многообразия Грассмана $\mathrm{Grass}(p, q)$, состоящего из всех q-мерных линейных подпространств в $\mathbb{R}^d$ [19].

Пусть $\mathbf{t} = \mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x})) \times \mathbf{z}$ и $\mathbf{t}' = \mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x})) \times \mathbf{z}'$ – два вектора из касательного пространства $\mathrm{L}(\mathbf{x})$ с коэффициентами $\mathbf{z} \in \mathbb{R}^q$ и $\mathbf{z}' \in \mathbb{R}^q$ их разложений по столбцам матрицы Якоби $\mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x}))$. Скалярное произведение $(\mathbf{t}, \mathbf{t}') = \mathbf{z}^{\top} \times \Delta_{\varphi}(\mathbf{x}) \times \mathbf{z}'$ индуцируется скалярным произведением объемлющего пространства $\mathbb{R}^d$, которое, в свою очередь, определяется матрицей $\Delta(\mathbf{x}) = (\mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x})))^{\top} \times \mathbf{J}_{\varphi}(\boldsymbol{\psi}(\mathbf{x}))$ размера $q \times q$ (такая матрица называется метрическим тензором на многообразии данных $\mathcal{M}$ в точке $\mathbf{x} \in \mathcal{M}$, гладко меняющимся от точки к точке); см. [20, 21]. Тензор $\Delta(\mathbf{x})$ индуцирует бесконечно малый элемент объема на касательном пространстве $\mathrm{L}(\mathbf{x})$, и, как следствие, риманову меру на многообразии

$$m(d\mathbf{x}) = \sqrt{|\det \Delta(\mathbf{x})|} \times \mathrm{mes}(d\mathbf{x}), \tag{5}$$

где $\mathrm{mes}(d\mathbf{x})$ – мера Лебега на многообразии данных $\mathcal{M}$ (см. [22]).

2.2. Предположения о вероятностной мере на многообразии данных

Пусть $\sigma(\mathcal{M})$ – борелевская σ-алгебра на $\mathcal{M}$ (минимальная σ-алгебра, содержащая все открытые подмножества $\mathcal{M}$) и пусть μ – вероятностная мера на измеримом пространстве $(\mathcal{M}, \sigma(\mathcal{M}))$, носитель которой совпадает с многообразием данных $\mathcal{M}$. Предположим, что мера μ абсолютно непрерывна относительно меры $m(dx)$ (5) на $\mathcal{M}$ и пусть

$$f(\mathbf{x}) = \mu(d\mathbf{x})/m(d\mathbf{x}) \tag{6}$$

– ее плотность, равномерно отделенная от нуля и бесконечности на $\mathcal{M}$.

3. МАШИННОЕ ОБУЧЕНИЕ НА ДАННЫХ СО ЗНАЧЕНИЯМИ НА МНОГООБРАЗИИ

3.1. Общие замечания

Пусть набор данных (1) случайно выбирается относительно неизвестной вероятностной меры μ, носитель которой $\mathrm{Supp}(\mu)$ является неизвестным многообразием данных $\mathcal{M}$ (3) с неизвестной внутренней размерностью q, вложенным в объемлющее пространство $\mathbb{R}^d$, $q < d$. Целью задачи моделирования многообразий является получение статистических выводов о многообразии данных по выборке $\mathbf{X}_n$. Ниже мы представим некоторые типичные примеры задач машинного обучения на данных со значениями на многообразии:

- оценка внутренней размерности,
- малоразмерная параметризация многообразия данных,
- оценка многообразия данных,
- оценка касательных пространств к многообразию данных,
- оценка плотности на многообразии данных,
- регрессия на многообразиях,

а также другие задачи, на решении которых мы вкратце остановимся ниже.

3.2. Предварительные сведения и обозначения

В этом разделе мы напомним определения некоторых основных понятий и обозначений, используемых в большинстве современных методов моделирования многообразий. Для начальной точки $\mathbf{x} \in \mathcal{M}$, через $\mathbf{x}_k(\mathbf{x}) \in \mathbf{X}_n$ обозначим k ближайших соседей точки $\mathbf{x}$ (то есть $\|\mathbf{x}_1(\mathbf{x}) - \mathbf{x}\| \leqslant \|\mathbf{x}_2(\mathbf{x}) - \mathbf{x}\| \leqslant \ldots \leqslant \|\mathbf{x}_k(\mathbf{x}) - \mathbf{x}\|$).

Пусть $K_{E,\varepsilon}(\mathbf{x}, \mathbf{x}') = \mathrm{I}\{\mathbf{x}' \in U(\mathbf{x}, \varepsilon)\}$ – евклидово ядро (здесь I – индикаторная функция и $U(\mathbf{x}, \varepsilon) = \{\mathbf{x}' \in \mathbf{X}_n : \|\mathbf{x}' - \mathbf{x}\| \leq \varepsilon\}$). Если размер выборки n является достаточно большим, то при малом ε и не очень больших значениях k точки $\mathbf{x}' \in U(\mathbf{x}, \varepsilon)$ и ближайшие соседи $\{\mathbf{x}_k(\mathbf{x})\}$ лежат вблизи q-мерного касательного пространства $\mathrm{L}(\mathbf{x})$ (см. (4)).

Рассмотрим взвешенный неориентированный выборочный граф $\Gamma(\mathbf{X}_n)$, состоящий из точек выборки $\mathbf{x}_i$, являющихся его узлами. Для заданного ε (или k) ребра графа $\Gamma(\mathbf{X}_n)$ соединяют точки $\mathbf{x}_i$ и $\mathbf{x}_j$ только если $\|\mathbf{x}_i - \mathbf{x}_j\| \leq \varepsilon$ (или когда это точки лежат в k ближайших соседях друг друга).

3.3. Оценка внутренней размерности многообразия данных

Грубо говоря, внутренняя размерность (Internal Dimension, ID) подмножества $\mathcal{X} \subset \mathbb{R}^d$ есть минимальное число $q = \mathrm{ID}(\mathcal{X})$ параметров, требуемых для генерации его описания, при которой потеря информации минимальна (см. [23]). Имеются различные определения внутренней размерности (топологическая внутренняя размерность, хаусдорфова размерность, колмогоровская емкость, корреляционная внутренняя размерность, а также другие; см. [24]), которые, однако, дают одни и те же значения для "неэкзотических" подмножеств.

К примеру, внутренняя размерность Хаусдорфа–Безиковича определяется следующим образом. Пусть множество E_r состоит из "полуоткрытых" кубов $\{Q = [k \times r, k \times r + r)^d,\ k = 0, \pm1, \pm2, \ldots\}$ с ребром r. Положим $N(\mathcal{X}, r) = \#\{Q \in E_r : \mathcal{X} \cap Q \neq \varnothing\}$. Предположим, что существуют числа $q = q_{HB}(\mathcal{X})$ и $V = V_{HB}(\mathcal{X})$ (называемые размерностью Хаусдорфа–Безиковича и объемом множества $\mathcal{X}$, соответственно), такие, что $N(\mathcal{X}, r)/(V \times r^{-q}) \to 1$ при $r \to 0$; это в частности означает, что множество $\mathcal{X}$ измеримо по Жордану.

Здесь задачей является оценка внутренней размерности $\mathrm{ID}(\mathcal{X})$ по набору данных (1), случайно выбираемых из пространства данных $\mathcal{X}$. Отметим, что $\mathrm{ID}(\mathbf{X}_n) = 0$ для приведенных выше определений внутренней размерности.

Для решения этой задачи имеется ряд методов (см. [25–28]), для некоторых из которых [25, 26] внутренняя размерность $\mathrm{ID}(\mathcal{M})$ оценивается в предположении о том, что данные лежат на многообразии $\mathcal{X} = \mathrm{DM}\,\mathcal{M}$ (см. (3)). К примеру, оценка по методу максимального правдоподобия

$$\widehat{q}(\mathcal{M}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k-1}\sum_{j=1}^{k-1}\ln\frac{\|\mathbf{x}_k(\mathbf{x}_i) - \mathbf{x}_i\|}{\|\mathbf{x}_j(\mathbf{x}_i) - \mathbf{x}_i\|}\right)^{-1}$$

основа на использовании k ближайших соседей $\{\mathbf{x}_j(\mathbf{x}_i)\}$ точек выборки [25].

3.4. Маломерная параметризация многообразия данных

Для заданной внутренней размерности q и выборки $\mathbf{X}_n$ статистическая задача состоит в построении отображения вложения $\mathbf{C}\colon \mathcal{M} \subset \mathbb{R}^d \to \mathcal{Z} = \mathbf{C}(\mathcal{M}) \subset \mathbb{R}^q$ из многообразия данных $\mathcal{M}$ в пространство признаков (Feature Set, FS) $\mathcal{Z}$, сохраняющего локальную геометрию данных, отношения близости, геодезические расстояния, углы и т.д. на многообразии данных. Большинство решений этой задачи (см. [16, 17]) начинаются с построения "n-точечного" отображения $\mathbf{X}_n \to \mathbf{Z}_n = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ (2), минимизирующего заданную функцию ошибок

$L_{(n)}(\mathbf{Z}_n|\mathbf{X}_n)$. После этого значения $\mathbf{C}(\mathbf{x})$ для точек $\mathbf{x} \in \mathcal{M} \setminus \mathbf{X}_n$ за пределами выборки вычисляются с использованием методов интерполяции.

К примеру, отображение ISOmetric MAPing (ISOMAP) [29] сохраняет геометрию многообразия данных, фиксируя геодезические расстояния $\{D_{ij}\}$ между всеми парами $\{(\mathbf{x}_i, \mathbf{x}_j)\}$ точек выборки. Для начала геодезические расстояния D_{ij} оцениваются через длины кратчайших путей $\{d_{ij}\}$ между узлами $\mathbf{x}_i$ и $\mathbf{x}_j$ графа $\Gamma(\mathbf{X}_n)$; в [30] доказывается хорошее качество таких оценок. После этого выборка признаков $\mathbf{Z}_n$ строится с использованием техники многомерного шкалирования [12] путем минимизации функции ошибок $L_{MDS}(\mathbf{Z}_n|\mathbf{X}_n) = \sum_{i,j=1}^{n}(d_{ij}^2 - \|\mathbf{z}_i - \mathbf{z}_j\|^2)^2$.

В методе Laplacian Eigenmaps (см. [31]) выборка признаков $\mathbf{Z}_n$ минимизирует функцию ошибок

$$L_{LE}(\mathbf{Z}_n|\mathbf{X}_n) = \sum_{i,j=1}^{n} K_{E,\varepsilon}(\mathbf{x}_i, \mathbf{x}_j) \times \|\mathbf{z}_i - \mathbf{z}_j\|^2 \tag{7}$$

при условии нормировки $\sum_{i,j=1}^{n} K_{E,\varepsilon}(\mathbf{x}_i, \mathbf{x}_j) \times (\mathbf{z}_i \times \mathbf{z}_i^\top) = \mathbf{I}_q$, которое вводится для того, чтобы избежать вырожденного решения; при этом подходе сохраняется внутренняя геометрическая структура многообразия данных.

В статистической постановке при асимптотически малом ε, функция ошибок (7) является выборочной оценкой величины

$$F(\mathbf{C}) = \int_{\mathcal{M}} |\nabla_{\mathcal{M}}\mathbf{C}(\mathbf{x})|^2 \, \mu(d\mathbf{x}) = \int_{\mathcal{M}} (\mathbf{C} \times \nabla_{\mathcal{M}}\mathbf{C})(\mathbf{x}) \, \mu(d\mathbf{x}),$$

называемой лапласианом графа $\Gamma(\mathbf{X}_n)$; здесь $\mathbf{C}(\mathbf{x})$ – некоторая компонента непрерывного интерполируемого отображения вложения $\mathbf{z} = \mathbf{C}_n(\mathbf{x}) \in \mathbb{R}^q$, определенного на многообразии данных $\mathcal{M}$, $\nabla_{\mathcal{M}}\mathbf{C}$ – ее ковариантный градиент и $\Delta_{\mathcal{M}}\mathbf{C}$ – оператор Лапласа–Бельтрами на многообразии данных. В [32] показано, что компоненты $\mathbf{C}_{1,n}(\mathbf{x}), \mathbf{C}_{2,n}(\mathbf{x}), \ldots, \mathbf{C}_{q,n}(\mathbf{x})$ отображения $\mathbf{C}_n(\mathbf{x})$ сходятся к собственным функциям $\mathbf{C}_1(\mathbf{x}), \mathbf{C}_2(\mathbf{x}), \ldots, \mathbf{C}_q(\mathbf{x})$ оператора Лапласа–Бельтрами $\Delta_{\mathcal{M}}$, соответствующего его самым малым ненулевым собственным значениям $\lambda_1 \leqslant \ldots \leqslant \lambda_q$.

Среди других примеров алгоритмов параметризации многообразий отметим следующие алгоритмы: алгоритм локально линейного вложения (Locally Linear Embedding) [33], алгоритм собственных отображений гессианов (Hessian Eigenmaps) [34], алгоритм максимального раскрытия дисперсии (Maximum Variance Unfolding) [35], алгоритм картирования многообразия (Manifold charting) [36], и др.

3.5. Сохранение информации в малоразмерной параметризации

В приложениях параметризация многообразия обычно служит первым шагом в различных задачах анализа данных, в которых редуцированные q-мерные признаки $\mathbf{z} = \mathbf{C}(\mathbf{x})$ используются в процедурах понижения размерности вместо изначально заданных d-мерных векторов $\mathbf{x}$. Если отображение вложения $\mathbf{C}$ сохраняет только специфические свойства многомерных данных, то при использовании редуцированного вектора $\mathbf{z} = \mathbf{C}(\mathbf{x})$ вместо начального вектора $\mathbf{x}$ возможны существенные потери информации. Во избежание таких потерь отображение $\mathbf{C}$ должно сохранять по возможности как можно больше доступной информации, содержащейся в многомерных данных (см. [37]).

Это означает возможность восстановления многомерных точек $\mathbf{x}$ по их малоразмерным представлениям $\mathbf{z} = \mathbf{C}(\mathbf{x})$, используя некоторое отображение восстановления $\mathbf{R}(\mathbf{z})\colon \mathcal{Z} \to \mathbb{R}^d$ с малой ошибкой восстановления $\delta_{\mathbf{C},\mathbf{R}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{R}(\mathbf{C}(\mathbf{x}))\|$.

Отображение $(\mathbf{C}, \mathbf{R})$ определяет q-мерное многообразие восстановленных данных (Recovered, Data Manifold, RDM)

$$\mathcal{M}_{\mathbf{C},\mathbf{R}} = \{\mathbf{x} = \mathbf{R}(\mathbf{z}) \in \mathbb{R}^d : \mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^q\}, \tag{8}$$

которое вложено в объемлющее пространство $\mathbb{R}^d$ и покрыто одной картой $\mathbf{R}$. Малая ошибка восстановления влечет близость $\mathcal{M}_{\mathbf{C},\mathbf{R}} \approx \mathcal{M}$ многообразий, что вследствие неравенства $d_H(\mathcal{M}_{\mathbf{C},\mathbf{R}}, \mathcal{M}) \leqslant \sup_{\mathbf{x}\in\mathcal{M}} \delta_{\mathbf{C},\mathbf{R}}(\mathbf{x})$ означает малость расстояния Хаусдорфа $d_H(\mathcal{M}_{\mathbf{C},\mathbf{R}}, \mathcal{M})$.

Пусть $\mathbf{J}_{\mathbf{R}}(\mathbf{z})$ – матрица Якоби размера $d \times q$ отображения восстановления $\mathbf{R}$. Тогда q-мерное линейное пространство $\mathrm{L}_{\mathbf{C},\mathbf{R}}(\mathbf{x}) = \mathrm{Span}(\mathbf{J}_{\mathbf{R}}(\mathbf{C}(\mathbf{x})))$ в $\mathbb{R}^d$ является касательным пространством к многообразию восстановленных данных $\mathcal{M}_{\mathbf{C},\mathbf{R}}$ в точке $\mathbf{R}(\mathbf{C}(\mathbf{x})) \in \mathcal{M}_{\mathbf{C},\mathbf{R}}$.

Для восстановления многообразия данных $\mathcal{M}$ по $\mathcal{Z} = \mathbf{C}(\mathcal{M})$ имеется (небольшое) число методов. Для линейного многообразия восстановление может быть легко получено на основе метода главных компонент [10]. Для нелинейных многообразий автокодировщики (разновидность нейронных сетей), см. [13], позволяют получить как вложение, так и отображение восстановления. В [38] был предложен общий метод, конструирующий отображение вложения таким же способом, как алгоритм локально линейного вложения (Locally Linear Embedding) [33]. Непараметрический регрессионный метод интерполяционного типа для восстановления многообразий используется в процедуре моделирования многообразий, называемой методом выравнивания касательного пространства (Tangent Space Alignment) [39]. Задача восстановления многообразия также может быть решена GSE-алгоритмом (Grassman–Stiefel Eigenmaps) [40, 41].

3.6. Оценка многообразия данных

В этой задаче требуется построить q-мерное многообразие $\mathcal{M}_n$, оценивающее (приближающее, восстанавливающее) многообразие данных $\mathcal{M}$ по выборке $\mathbf{X}_n$.

В работах, относящихся к вычислительной геометрии, эта задача ставится следующим образом: для заданного конечного набора данных $\mathbf{X}_n$ построить некоторое множество $\mathcal{M}^* \subset \mathbb{R}^d$, аппроксимирующее $\mathcal{M}$ в подходящем смысле [42]. Подходы к решению этой задачи обычно основываются на разложении многообразия данных $\mathcal{M}$ на малые области (например, используя диаграмму Вороного или триангуляцию Делоне многообразия $\mathcal{M}$), после чего каждая область кусочно аппроксимируется некоторой геометрической структурой (например, симплициальным комплексом [42], касательным комплексом Делоне [43], конечным числом аффинных подпространств (“плоскостями”) [44], k-средними и k-плоскостями, см. [45], и т.д.). Однако, такие методы обладают общим недостатком – они не позволяют найти малоразмерную параметризацию многообразия данных, а такая параметризация обычно требуется в задачах машинного обучения, в которых исходные данные имеют большую размерность.

В статистической постановке, когда оценивается многообразие данных $\mathcal{M}$ (см. (3)), покрытое одной картой, решение $\mathcal{M}_n$ должно быть также q-мерным многообразием, покрытым одной картой и, следовательно, иметь вид (8).

Ошибка восстановления $\delta_{\mathbf{C},\mathbf{R}}(\mathbf{x})$ может быть напрямую вычислена в точках выборки $\mathbf{x} \in \mathbf{X}_n$; для точек вне выборки $\mathbf{x}$ она описывает обобщающую способность решения $(\mathbf{C}, \mathbf{R})$ в конкретной точке $\mathbf{x}$. В [41] получены локальные оценки снизу и сверху для максимальной ошибки восстановления в малой окрестности произвольной точки $\mathbf{x} \in \mathcal{M}$. Такие оценки определяются в терминах расстояния между касательными пространствами $\mathrm{L}(\mathbf{x})$ и $\mathrm{L}_{\mathbf{C},\mathbf{R}}(\mathbf{x})$ к многообразию данных $\mathcal{M}$ и многообразию восстановленных данных $\mathcal{M}_{\mathbf{C},\mathbf{R}}$ в точках $\mathbf{x}$ и $\mathbf{R}(\mathbf{C}(\mathbf{x}))$, соответственно, относительно выбранной метрики на многообразии Грассмана $\mathrm{Grass}(d, q)$.

Из этих оценок следует, что чем больше расстояние между этими касательными пространствами, тем ниже обобщающая способность решения $(\mathbf{C}, \mathbf{R})$. Таким образом, естественным

является требование, что для решения $(\mathbf{C}, \mathbf{R})$ должна обеспечиваться не только близость многообразий $\mathcal{M}_{\mathbf{C},\mathbf{R}} \approx \mathcal{M}$, но также и близость касательных пространств $\mathrm{L}_{\mathbf{C},\mathbf{R}}(\mathbf{x}) \approx \mathrm{L}(\mathbf{x})$ для всех точек $\mathbf{x} \in \mathcal{M}$. В теории многообразий [20, 21] объединение касательных пространств к многообразию в разных точках называется касательным расслоением многообразия. Таким образом, задача восстановления многообразия, которая включает в себя и задачу восстановления касательных пространств, называется задачей моделирования касательного расслоения многообразий. GSE-алгоритм (Grassman/Stiefel Eigenmaps) [40], [41] позволяет получить решение этой задачи и имеет скорость сходимости

$$\|\mathbf{x} - \mathbf{R}(\mathbf{C}(\mathbf{x}))\| = O(n^{-2/(q+2)}), \quad d_{P,2}(\mathrm{L}(\mathbf{x}), \mathrm{L}_{\mathbf{C},\mathbf{R}}(\mathbf{x})) = O(n^{-1/(q+2)}) \tag{9}$$

с высокой вероятностью равномерно по точкам $\mathbf{x} \in \mathcal{M}$ (см. [46]) асимптотически при $n \to \infty$ для подходящего выбора параметров алгоритма (например, когда радиус шара $\varepsilon = \varepsilon_n$ в ядре $K_{E,\varepsilon}(\mathbf{x}, \mathbf{x}')$ имеет порядок $O(n^{-1/(q+2)})$); здесь $d_{P,2}$ – проекция L^2-нормы на многообразие Грассмана [47]; в статистике такая метрика называется мин-корреляционной метрикой (Min Correlation Metric), см. [48]. Фраза "событие случается с высокой вероятностью" означает, что вероятность события превышает значение $(1 - C_\alpha/n^\alpha)$ при любых n и $\alpha > 0$, при этом константа C_α зависит только от α. Первая скорость сходимости в (9) совпадает с асимптотически минимаксной нижней оценкой расстояния Хаусдорфа между многообразием данных и многообразием восстановленных данных (см. [49]).

3.7. Оценка касательных пространств к многообразию данных

Простейшей оценкой для касательного пространства $\mathrm{L}(\mathbf{x})$ к многообразию данных $\mathcal{M}$ является линейное подпространство $\mathrm{L}_{PCA}(\mathbf{x})$, которое получается применением метода главных компонент [10] к локальному набору данных $U(\mathbf{x}, \varepsilon)$ при достаточно малом пороговом значении ε. Асимптотические свойства оценок изучались в работах [50–52]. Неасимптотический анализ подпространства $\mathrm{L}_{PCA}(\mathbf{x})$ касательных пространств проведен в [53].

Собственные векторы локальной выборочной ковариационной матрицы

$$\mathbf{S}(\mathbf{x}) = \sum_{i=1}^{n} K_{E,\varepsilon}(\mathbf{x}, \mathbf{x}_i) \times (\mathbf{x}_i - \mathbf{x}) \times (\mathbf{x}_i - \mathbf{x})^\top,$$

соответствующие q наибольшим собственным значениям, образуют базис подпространства $\mathrm{L}_{PCA}(\mathbf{x})$. Но эти базисы не согласованы друг с другом и могут быть существенно различными даже в близких точках. В ряде работ [40, 41] в подпространстве $\mathrm{L}_{PCA}(\mathbf{x})$ были построены "выровненные" базисы $\{\mathbf{H}_1(\mathbf{x}), \mathbf{H}_2(\mathbf{x}), \ldots, \mathbf{H}_q(\mathbf{x})\}$. В [54] ортогональные "выровненные" базисы были построены для обеспечения свойств локальной изометрии и конформности параметризации многообразий.

3.8. Оценка плотности на многообразии данных

Задача оценки неизвестной плотности $f(\mathbf{x})$ (см. (6)) на многообразии данных $\mathcal{M}$ изучалась в работах [55–57]. В [58] предложена новая геометрически мотивированная нестационарная оценка ядерной плотности для неизвестной плотности, основанная на GSE-алгоритме [40, 41].

3.9. Построение генеративных моделей на многообразии

Предположим, что многообразие $\mathcal{M}$ представляется в виде модели типа (3), при этом отображение $\boldsymbol{\varphi}(\cdot)$ задается параметрическим образом с помощью глубокой нейросети $\boldsymbol{\varphi}_{\boldsymbol{\theta}} : \mathcal{Z} \to \mathcal{X}$,

где $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ – её параметры. Пусть $\mathbf{z} \sim p(\mathbf{z})$ – некоторое "простое" распределение в маломерном латентном пространстве, например, гауссовское. В таком случае, данные, "живущие" на многообразии, будут иметь распределение

$$\mathbf{x} \sim \int_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{x}|\boldsymbol{\varphi}_{\boldsymbol{\theta}}(\mathbf{z}))p(\mathbf{z})d\mathbf{z}$$

Соответственно, сгенерировав $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\} \sim p(\mathbf{z})$, можно получить выборку $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \sim p(\mathbf{x}|\boldsymbol{\varphi}_{\boldsymbol{\theta}}(\mathbf{z}_i))$ со "сложным" распределением.

Такого рода вероятностную (генеративную) модель можно использовать для восстановления пропущенных значений, повышения разрешения изображений, и т.п. [59,60]. Действительно, допустим, что некоторые координаты заданного вектора $\mathbf{x}$ неизвестны. В таком случае, по известным координатам вектора $\mathbf{x}_I = \{x_i, i \in I\}$ с индексами I мы можем получить оценку максимума правдоподобия для латентного представления точки данных

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} \left[\log p(\mathbf{x}_I|\boldsymbol{\varphi}_{\boldsymbol{\theta}}(\mathbf{z})) + \log p(\mathbf{z})\right].$$

Оценку пропусков можно будет сделать в таком случае согласно формуле $\hat{\mathbf{x}} = \boldsymbol{\varphi}_{\boldsymbol{\theta}}(\hat{\mathbf{z}})$.

Для построения вероятностной модели зачастую используется подход на основе вариационного автокодировщика [61]. В этом случае для моделирования распределения $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ используется представление вида (отображение восстановления) $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mu_{\boldsymbol{\theta}}(\mathbf{x}), \sigma^2_{\boldsymbol{\theta}}(\mathbf{x}))$, $\mathbf{z} \in \mathcal{Z}$, где функции $\mu_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathbb{R}^d$ и $\sigma_{\boldsymbol{\theta}}(\mathbf{z}) = \mathrm{diag}\{\sigma_{\boldsymbol{\theta},1}(\mathbf{z}), \ldots, \sigma_{\boldsymbol{\theta},d}(\mathbf{z})\} \in \mathbb{R}^{d\times d}$ – представляются с помощью глубоких нейросетей. В данном случае $\mu_{\boldsymbol{\theta}}(\mathbf{z})$ играет роль координатной карты $\boldsymbol{\varphi}_{\boldsymbol{\theta}}(\mathbf{z})$, а $\sigma_{\boldsymbol{\theta}}(\mathbf{z})$ — моделирует отклонение от многообразия наблюдаемых данных.

Основная сложность получения оценки параметров $\boldsymbol{\theta}$ заключается в том, что для правдоподобия выборки данных $\mathbf{X}_n$ нельзя получить явного выражения, так как интеграл $p_{\boldsymbol{\theta}}(\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})d\mathbf{z}$ не берется в явном виде. Как следствие, невозможно и вычисление апостериорного распределения $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z})/p_{\boldsymbol{\theta}}(\mathbf{x})$, которое можно было бы использовать для оценки латентного представления $\mathbf{z}$ объекта $\mathbf{x}$. Для решения данной проблемы на практике обычно используют вариационный вывод, а именно, аппроксимируют апостериорное распределение $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ распределением $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ из вариационного семейства с амортизированным выводом (amortized inference), имеющим вид $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu_{\boldsymbol{\phi}}(\mathbf{x}), \sigma^2_{\boldsymbol{\phi}}(\mathbf{x}))$, $\mathbf{x} \in \mathcal{X}$, где функции $\mu_{\boldsymbol{\phi}}(\mathbf{x}) \in \mathbb{R}^q$ и $\sigma_{\boldsymbol{\phi}}(\mathbf{x}) = \mathrm{diag}\{\sigma_{\boldsymbol{\phi},1}(\mathbf{x}), \ldots, \sigma_{\boldsymbol{\phi},q}(\mathbf{x})\} \in \mathbb{R}^{q\times q}$ – представляются с помощью глубоких нейросетей. В данном случае $\mu_{\boldsymbol{\phi}}(\mathbf{x})$ играет роль обратного отображения $\boldsymbol{\psi}$ (отображение сжатия), которое определяет маломерную параметризацию многообразия данных $\mathcal{M}$.

На практике, для оценки параметров $(\boldsymbol{\theta}, \boldsymbol{\phi})$ используют максимизацию нижней границы правдоподобия, а именно, выполнено, что

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{KL}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})] + L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) \geq L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}),$$

так так дивергенция Кульбака-Лейблера $\mathrm{KL}[\cdot\|\cdot] \geq 0$. Здесь

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}) = -\mathrm{KL}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] + \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})]. \tag{10}$$

Второе слагаемое в правой части уравнения (10) по сути является ошибкой восстановления значения $\mathbf{x}$ по значению $\mathbf{z}$, усредненному с распределением $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, которое аппроксимирует апостериорное распределение $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. Первое слагаемое в (10) играет роль регуляризации. В итоге, решается задача оптимизации $\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \geq \sum_{i=1}^n L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}_i) \to \max_{\boldsymbol{\theta},\boldsymbol{\phi}}$. Для оценки $L(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x})$ на практике используют Монте-Карло, то есть генерируют $\mathbf{z}_{i,l} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_i)$,

$l = 1, \ldots, L$, и, предполагая гауссовость распределений $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ и $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}_q)$, оценивают выражение (10) как

$$\hat{L}(\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{x}_i) = \frac{1}{2}\sum_{j=1}^{d}[1 + \log\sigma^2_{j,\boldsymbol{\phi}}(\mathbf{x}_i) - \mu^2_{j,\boldsymbol{\phi}}(\mathbf{x}_i) - \sigma^2_{j,\boldsymbol{\phi}}(\mathbf{x}_i)] + \frac{1}{L}\sum_{l=1}^{L}\log_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_{i,l}).$$

Описанная вероятностная модель на многообразии позволяет порождать новые объекты $\mathbf{x} \in \mathcal{X}$ из $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$, предварительно генерируя $\mathbf{z} \sim p(\mathbf{z})$. Отметим, что в латентном пространстве $\mathcal{Z}$ также имеет смысл рассматривать не только "простые" модели распределений типа гауссовского $p(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}_q)$, но и использовать параметризации многообразий с более сложной структурой, что иногда позволяет достичь более эффективных практических результатов при решении задач восстановления, повышения разрешения данных, и т.п., см. [59, 62].

3.10. Регрессия на многообразиях данных

Пусть $\mathbf{y} = \mathbf{f}(\mathbf{x})$ – неизвестное гладкое отображение из своей области определения $\mathcal{M}$ ($\mathcal{M}$ лежит в пространстве входных данных $\mathbb{R}^d$) в k-мерное выходное пространство $\mathbb{R}^k$; область определения $\mathcal{M}$ также предполагается неизвестной. При заданной входной-выходной выборке $\{(\mathbf{x}_i, \mathbf{y}_i = \mathbf{f}(\mathbf{x}_i),\ i = 1, 2, \ldots, n\}$, типичной проблемой задачи регрессии является оценка неизвестного отображения $\mathbf{f}$. Если область определения $\mathcal{M}$ является q-мерным входным многообразием, $q < d$, то мы говорим о задаче оценки регрессии, определенной на многообразии.

Если входное многообразие $\mathcal{M}$ известно, т.е. если известна малоразмерная параметризация $\boldsymbol{\psi}$ в (3), то регрессия в задачах оценки многообразий может быть сведена к классической многомерной задаче регрессии (многовыходной, если $k > 1$) (см. [63]). При неизвестном входном многообразии в [64–67] были получены различные частные решения этой задачи. В [68–71] рассмотрена типичная задача регрессии, состоящая в оценке неизвестного отображения $\mathbf{f}$, ее матрицы Якоби и неизвестного входного многообразия $\mathcal{M}$.

3.11. Перенос опыта в задачах прогнозирования на основе моделей многообразий

В приложениях классификации и сегментации изображений, особенно в случае обработки медицинских снимков, зачастую не имеется достаточного количества размеченных данных, поскольку разметка — требует значительных ресурсов. Более того, некоторые метки могут быть "слабыми" [72], в данных присутствовать аномалии, и сама выборка — может быть несбалансированной [73, 74]. В этом случае такой инструмент машинного обучения, как методы переноса опыта (transfer learning — повышение скорости и эффективности обучения за счет учета уже имеющихся результатов обучения на данных, собранных при сходных условиях), может помочь получить модели с более высокой точностью.

Как было показано в [75, 76], задачу переноса опыта при обучении новой модели, можно сформулировать на языке моделирования многообразий. Пусть $p(\mathbf{y}|\mathbf{x}, \mathbf{w})$ — вероятностный прогноз модели, параметры $\mathbf{w} \in \mathbb{R}^p$ которой были обучены на выборке данных $S^1_{n_1} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_1}$. Между компонентами вектора параметров $\mathbf{w}$ имеются зависимости (особенно это явно проявляется в случае обработки изображений, когда $\mathbf{w}$ — это коэффициенты сверточных фильтров [77]). Соответственно, для новых задач, сходных с задачей, которая представлена выборкой $S^1_{n_1}$, параметры моделей $\mathbf{w}$, обученных на соответствующих выборках из новых задач, также должны иметь аналогичные зависимости. Такого рода ограничения можно формализовать предположив, что параметры модели $\mathbf{w}$ лежат на неизвестном q-мерном многообразии $\mathcal{M} \subset \mathbb{R}^p$, описывающем особенности исходной задачи.

Многообразие $\mathcal{M}$ можно моделировать с помощью вариационного автокодировщика, задаваемого распределением $p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}_p)$, играющим роль отображения восстановления, и вариационным распределением $r(\mathbf{z}|\mathbf{w}, \boldsymbol{\phi}_r)$, играющим роль отображения сжатия, см. раздел 3.9 (здесь, $(\boldsymbol{\phi}_p, \boldsymbol{\phi}_r)$ — параметры соответствующих нейросетей). При этом мы определяем не только само многообразие, но и некоторую меру μ на нем неявным образом. Для обучения вариационного автокодировщика $\{r(\mathbf{z}|\mathbf{w}, \boldsymbol{\phi}_r), p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}_p)\}$ в [75] предложена процедура построения выборки значений параметров $\{\mathbf{w}_i\}_{i=1}^n$ на основе бутстрепа по исходной выборке данных $S_{n_1}^1$.

Итак, при максимизации логарифма правдоподобия

$$\mathcal{L}(\mathbf{w}|S_{n_2}^2) = \sum_{(\mathbf{x},\mathbf{y})\in S_n^2} \log p(\mathbf{y}|\mathbf{x}, \mathbf{w})$$

для обучения модели на выборке $S_{n_2}^2$, представляющей новую задачу, имеет смысл накладывать ограничение, что $\mathbf{w} \in \mathcal{M}$. Так как многообразие $\mathcal{M}$ и распределение μ на нем заданы неявно с помощью вариационного автокодировщика $\{r(\mathbf{z}|\mathbf{w}, \boldsymbol{\phi}_r), p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}_p)\}$, то явная оптимизация $\mathcal{L}(\mathbf{w}|S_{n_2}^2)$ по $\mathbf{w}$ с ограничением $\mathbf{w} \in \mathcal{M}$ затруднена. Использование байесовского вариационного вывода позволяет получить нижнюю границу для правдоподобия [75]

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{x}, \mathbf{w}) \geq & \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{w})} \log p(\mathbf{y}|\mathbf{x}, \mathbf{w}) + \mathbb{H}(q_{\boldsymbol{\theta}}(\mathbf{w})) - \\ & - \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{w})} \left\{ \mathrm{KL}[r(\mathbf{z}|\mathbf{w}, \boldsymbol{\phi}_r) \| p(\mathbf{z})] - \mathbb{E}_{r(\mathbf{z}|\mathbf{w}, \boldsymbol{\phi}_r)} \log p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi}_p) \right\},\end{aligned}$$

где $\mathbb{H}(q_{\boldsymbol{\theta}}(\mathbf{w}))$ — энтропия вариационного распределения $q_{\boldsymbol{\theta}}(\mathbf{w})$, используемого для аппроксимации апостериорного распределения $p(\mathbf{w}|S_{n_2}^2)$. Оптимизируя эту нижнюю границу по параметрам $\boldsymbol{\theta}$, можно оценить $\mathbf{w}$. Эксперименты, проведенные в [76], показали, что данный подход позволяет получить модели с высокой точностью даже если размер выборки $n_2 \ll n_1$, что особенно ценно при обработке медицинских данных.

4. ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрены следующие различные задачи машинного обучения на многомерных данных со значениями на многообразии: задача оценки многообразия данных (включая оценку его внутренней размерности и касательных пространств, а также построение малоразмерной параметризации многообразия), задача оценки плотности многообразия данных, задача регрессии на многообразии, и ряд других. Дается небольшой обзор решений, полученных для данных задач.

СПИСОК ЛИТЕРАТУРЫ

1. William S. Cleveland. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1):21–26, 2001.

2. L.M. Chen, Z. Su, and B. Jiang. *Mathematical Problems in Data Science: Theoretical and Practical Methods*. Springer, 2015.

3. David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS conf. on Math Challenges of the 21st Century*, 2000.

4. Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

5. Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2006.

6. T. Cacoullos. Estimation of a multivariate density. *Ann Inst Stat Math*, 18(18):179–189, 1966.

7. A. Kuleshov, A. Bernstein, E. Burnaev, and Y. Yanovich. Machine learning in appearance-based robot self-localization. In *16th IEEE Int. Conf. ICMLA*, pages 106–112, 2017.
8. A. Kuleshov, A. Bernstein, and E. Burnaev. Mobile robot localization via machine learning. In *13th Int. Conf. MLDM*, volume 10358 of *Lecture Notes in Computer Science*, pages 276–290. Springer, 2017.
9. K Bunte, M. Biehl, and B. Hammer. Dimensionality reduction mappings. In *Proc. of the IEEE Symp. on Comp. Intel. and Data Mining, CIDM 2011*, pages 349–356. IEEE, 2011.
10. T. Jollie. *Principal Component Analysis*. Springer, 2002.
11. E. Burnaev and S. Chernova. On an iterative algorithm for calculating weighted principal components. *Journal of Communications Technology and Electronics*, 60(6):619–624, Jun 2015.
12. T.F. Cox and M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall/CRC, 2001.
13. G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
14. B. Scholkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.
15. H.S. Seung and D.D. Lee. The manifold ways of perception. *Science*, 290(5500):2268–2269, 2000.
16. Xiaoming Huo, Xuelei (Sherry) Ni, and Andrew K. Smith. *A Survey of Manifold-Based Learning Methods*, pages 691–745. World Scientific, 2007.
17. Y. Ma and Y. Fu. *Manifold Learning Theory and Applications*. CRC Press, 2011.
18. Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1–3):419–441, 2008.
19. R.P. Woods. Differential geometry of grassmann manifolds. *Proc. Nqt. Acad. Sci. USA*, 57(3):589–594, 1967.
20. J. Jost. *Riemannian Geometry and Geometric analysis*. Springer, 2011.
21. J.M. Lee. *Manifolds and Differential Geometry*. AMS, 2009.
22. X. Pennec. Probabilities and statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.
23. R.S. Bennett. The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory*, 15(5):517–525, 1969.
24. M. Katetov and P. Simon. *Origins of dimension theory*, pages 113–134. Kluwer, 1997.
25. Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *NIPS 17*, pages 777–784. 2005.
26. M. Fan, H. Qiao, and B. Zhang. Intrinsic dimension estimation of manifolds by incising balls. *Pattern Recognition*, 42(5):780–787, 2009.
27. P. Campadelli, E. Casiraghi, C. Ceruti, and A. Rozza. Intrinsic dimension estimation: Relevant techniques and a benchmark framework. *Mathematical Problems in Engineering*, pages 1–21, 2015.
28. F. Camastra and A. Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.
29. J.B. Tehenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.
30. M. Bernstein, V. de Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. *Technical report*, 2000.
31. M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

32. Yu. Yanovich. Asymptotic properties of eigenvalues and eigenfunctions estimates of linear operators on manifolds. *Lobachevskii Journal of Mathematics*, 38(6):1–12, 2017.

33. L.K. Saul and S.T. Roweis. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

34. D.L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Arts and Sciences*, 100(10):5591–5596, 2003.

35. K.Q. Weinberger and L.K. Saul. Maximum variance unfolding: Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

36. Matthew Brand. Charting a manifold. In *5th Int. Conf. NIPS*, pages 985–992. MIT Press, 2002.

37. J.A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009.

38. L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

39. Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2005.

40. A. V. Bernstein and A. Kuleshov. Tangent bundle manifold learning via grassmann and stiefel eigenmaps. *ArXiv*, abs/1212.6031, 2012.

41. A.V. Bernstein and A.P. Kuleshov. Manifold learning: generalizing ability and tangent proximity. *International Journal of Software and Informatics*, 7(3):359–390, 2013.

42. D. Freedman. Efficient simplicial reconstructions of manifold from their samples. *IEEE TPAMI*, 24(10):1349–1357, 2002.

43. J.-D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential delaunay complexes. *Discrete and Computational Geometry*, 51(1):221–267, 2014.

44. S. Karygianni and P. Frossard. Tangent-based manifold approximation with locally linear models. *Signal Processing*, 104:232–247, 2014.

45. Guillermo Canas, Tomaso Poggio, and Lorenzo Rosasco. Learning manifolds with k-means and k-flats. In *Advances in NIPS 25*, pages 2465–2473. 2012.

46. A. Kuleshov, A. Bernstein, and Yu. Yanovich. Asymptotically optimal method in Manifold estimation. *Abstracts of the XXIX-th European Meeting of Statisticians*, pages 325, 2013.

47. J. Hamm and Daniel D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. *Proc. of the 25th ICML*, pages 376–383, 2008.

48. H. Hotelling. Relations between two sets of variables. *Biometrika*, 28(3/4):321–377, 1936.

49. C.R. Genovese, M. Perone-Pacico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *Journal Machine Learning Research*, 13:1263–1291, 2012.

50. H. Tyagi, E. Vural, and P. Frossard. Tangent space estimation for smooth embeddings of riemannian manifolds. *Information and Inference: A Journal of the IMA*, 2:69–114, 2012.

51. R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, F. Warner, and S. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, 2005.

52. Yu. Yanovich. Asymptotic properties of local sampling on manifold. *Journal of Mathematics and Statistics*, 12(3):157–175, 2016.

53. D.N. Kaslovsky and F.G. Meyer. Non-asymptotic analysis of tangent space perturbation. *Inf. J. IMA.*, 3(2):134–187, 2014.

54. A. Bernstein, A. Kuleshov, and Y. Yanovich. Information preserving and locally isometric conformal embedding via tangent manifold learning. In *IEEE Int. Conf. DSAA*, pages 1–9, 2015.

55. G. Henry, A. Munoz, and D. Rodriguez. Locally adaptive density estimation on riemannian manifolds. *Statistics and Operations Research Transactions*, 37(2):111–130, 2013.

56. G. Henry and D. Rodriguez. Kernel density estimation on riemannian manifolds: asymptotic results. *Journal of Math. Imaging and Vision*, 34(3):235–639, 2009.

57. Y.T. Kim and H.S. Park. Geometric structures arising from kernel density estimation on riemannian manifolds. *J. Multivariate Anal.*, 114:112–126, 2013.

58. A. Kuleshov, A. Bernstein, and Y. Yanovich. High-dimensional density estimation for data mining tasks. In *IEEE Int. Conf. ICDM Workshops*, pages 523–530, 2017.

59. S. Athar, E. Burnaev, and V. Lempitsky. Latent convolutional models. In *ICLR*, 2019.

60. O. Voynov, A. Artemov, V. Egiazarian, A. Notchenko, G. Bobrovskikh, E. Burnaev, and D. Zorin. Perceptual deep depth super-resolution. In *ICCV*, pages 5652–5662, 2019.

61. Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.

62. Anna Kuzina, Evgenii Egorov, and Evgeny Burnaev. Boovae: A scalable framework for continual vae learning under boosting approach. *arXiv 1908.11853*, 2019.

63. F. Steinke, M. Hein, and B. Scholkopf. Nonparametric regression between general riemannian manifolds. *SIAM J. Imaging Sci.*, 3(3):527–563, 2010.

64. X. Shi, M. Styner, J. Lieberman, J.G. Ibrahim, W. Lin, and H. Zhu. Intrinsic regression models for manifold-valued data. *J. Amer. Stat. Assoc.*, 5762(1):192–199, 2009.

65. A. Aswani, P. Bickel, and C. Tomlin. Regression on manifolds: Estimation of the exterior derivative. *The Annals of Statistics*, 39(1):48–81, 2011.

66. Jochen Einbeck and Ludger Evers. Localized regression on principal manifolds. *25th International Workshop on Statistical Modelling (IWSM 2010)*, 2010.

67. P.T. Fletcher. Geodesic regression and the theory of least squares on riemannian manifolds. *International Journal of Computer Vision*, 105(2):171–185, 2013.

68. A. Kuleshov and A. Bernstein. Nonlinear multi-output regression on unknown input manifold. *Annals of Mathematics and Artificial Intelligence*, 81(1–2):209–240, 2017.

69. A. Kuleshov, A. Bernstein, and E. Burnaev. Kernel regression on manifold valued data. In *Proc. of IEEE 5th Int. Conf. DSAA-2018*, pages 120–129, 2018.

70. A. Kuleshov, A. Bernstein, and E. Burnaev. Manifold learning regression with non-stationary kernels. In *ANNPR*, pages 152–164. Springer, 2018.

71. A. Kuleshov, A. Bernstein, and E. Burnaev. Conformal prediction in manifold learning. In *Proc. of 7th COPA Workshop*, volume 91, pages 234–253. PMLR, 2018.

72. S. Pavlov, A. Artemov, M. Sharaev, A. Bernstein, and E. Burnaev. Weakly supervised fine tuning approach for brain tumor segmentation problem. In *18th Int. Conf. ICMLA*, pages 1600–1605, 2019.

73. E. Burnaev, P. Erofeev, and A. Papanov. Influence of resampling on accuracy of imbalanced classification. In *Proc. SPIE 9875, 8th Int. Conf. ICMV 2014*, volume 9875, 2015.

74. D. Smolyakov, A. Korotin, P. Erofeev, A. Papanov, and E. Burnaev. Meta-learning for resampling recommendation systems. In *Proc. SPIE 11041, 11th Int. Conf. ICMV 2018*, volume 11041, 2019.

75. Andrei Atanov, Arsenii Ashukha, Kirill Struminsky, Dmitriy Vetrov, and Max Welling. The deep weight prior. In *ICLR*, 2019.

76. Anna Kuzina, Evgenii Egorov, and Evgeny Burnaev. Bayesian generative models for knowledge transfer in mri semantic segmentation problems. *Frontiers in neuroscience*, 13:844, 2019.

77. M. Pominova, A. Artemov, M. Sharaev, E. Kondrateva, A., E. Burnaev, and A. Bernstein. Voxelwise 3d convolutional and recurrent neural networks for epilepsy and depression diagnostics from structural and functional mri data. In *Proc. of IEEE Int. Conf. ICDM Workshops*, pages 299–307, 2018.

Manifold Modeling in Statistical Problems

Burnaev E.V., **Bernstein A.V.**

Predictive Modeling tasks deal with high-dimensional data, and curse of dimensionality is an obstacle to the use of many methods for their solutions. In many applications, real-world data occupy only a very small part of high-dimensional observation space whose intrinsic dimension is essentially lower than dimension of the space. Popular model for such data is a Manifold one in accordance with which data lie on or near an unknown low-dimensional Data manifold (DM) embedded in an ambient high-dimensional space. Predictive Modeling tasks studied under this assumption are referred to as the manifold learning ones whose general goal is discovering a low-dimensional structure of high-dimensional manifold valued data from a given dataset. If dataset points are sampled according to an unknown probability measure on the DM, we face with statistical problems about manifold valued data. In this paper we make a short review of statistical problems regarding high-dimensional manifold valued data and their solutions.

KEYWORDS: dimension reduction, manifold learning, predictive modeling.