

Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Магистерская программа: Науки о данных
Кафедра: Технологий моделирования сложных систем

КУРСОВАЯ РАБОТА

на тему:

Тестирование гипотезы о многообразии для
реальных данных

Студент группы ТМСС,
Буланов Олег Викторович

Руководитель КР:
Старший преподаватель
Янович Юрий Александрович

Москва, 2017

Содержание

1	Введение	2
2	Определения внутренней размерности	3
3	Описание методов	5
3.1	Метод Левиной-Бикеля	6
3.2	Метод Isomap	7
3.3	Метод Граната-Карневале	8
4	Численные эксперименты	10
4.1	Применение методов к многомерным сферам	10
4.2	Применение к Swiss roll, Moons, S curve	12
4.3	Применение к реальным данным	13
5	Заключение	15
	Список литературы	16

1. Введение

Современные данные часто имеют большое количество признаков, то есть большую размерность. Многие из этих признаков зависят друг от друга не всегда линейно. Предполагается, что исходные данные $\Omega \subset \mathbb{R}^d$ лежат на многообразии меньшей размерности $\Omega \subseteq \mathbb{R}^M \subset \mathbb{R}^d, M < d$ [1]. В связи с этим возникает задача снижения размерности, то есть нахождение методов проецирующих исходные данные на многообразие меньшей размерности без потери информации. Это важная задача, потому что многие алгоритмы машинного обучения работают дольше на данных большой размерности, так как вычислительная сложность зависит от размерности иногда даже экспоненциально [2]. Также с увеличением размерности для корректной работы методов требуется больше данных, это называют проклятием размерности [3]. Многие методы снижения размерности, например, PCA [4], Isomap [5], LLE [6] и другие, требуют внутреннюю размерность M в качестве входного параметра, поэтому задача оценивания этого параметра является достаточно актуальной.

Эта работа посвящена исследованию методов оценки внутренней размерности. Основными задачами являются: реализация методов и тестирование их на модельных данных, то есть данных с заранее известной внутренней размерностью, и реальных данных. В частности, были рассмотрены три метода оценивания внутренней размерности: метод Левиной-Бикеля [7], метод основанный на алгоритме Isomap [5] и метод Граната-Карневале [8].

2. Определения внутренней размерности

Прежде чем описывать методы стоит ввести понятие внутренней размерности. Стоит отметить, что существует несколько определений внутренней размерности, которые при некоторых условиях являются эквивалентными. Таким образом есть возможность создавать новые методы на основании разных формулировок. Первое определение достаточно общее, но оно даёт понимание того, что мы хотим сделать

Определение 1 Выборка $\Omega \subseteq \mathbb{R}^N$ имеет **внутреннюю размерность** M , если все элементы выборки лежат в M -мерном многообразии в \mathbb{R}^N $M < N$ без потери информации.

Например, по этому определению размерность сферы в трехмерном пространстве равна 2, потому что любая координата выражается через функцию от 2 параметров (параметрическое представление сферы в сферических координатах).

Теперь введем одно из первых определений размерности, которое достаточно общее, но его частные случаи часто используются для определения размерности.

Определение 2 Мера Хаусдорфа — $\Gamma_H^m = \lim_{r \rightarrow 0} \inf_{s_i} \underbrace{\sum_i (r_i)^m}_{\Gamma_H^m(r)}$, где s_i — ячейки

диаметра $r_i < r$ покрывающие все элементы выборки. Существует критическое значение $m_H : \Gamma_H^m = \infty \forall m < m_H$ и $\Gamma_H^m = 0 \forall m > m_H$, которое называется **размерностью Хаусдорфа для множества** (Hausdorff dimension of the set).

Для примера со сферой. Мы пытаемся покрыть площадь сферы m -мерными ячейками размера r . Получаем, что при $m \leq 1$ мы пытаемся покрыть поверхность сферы отрезками или точками. Конечным числом таких ячеек это сделать не получится, поэтому $\Gamma_H^{0/1} = \infty$. При $m = 2$ мы пытаемся покрыть по-

верхность сферы 2-мерными ячейками, например, кругами, это можно сделать конечным числом таких ячеек, следовательно, $0 < \Gamma_H^2 < \infty$. То есть размерность Хаусдорфа для сферы будет 2. Следующая размерность является частным случаем размерности Хаусдорфа.

Определение 3 Информационная размерность (*information dimension*) вероятностной меры P

$$m_H(P) = \inf_B \{m_H(B) : P(B) = 1\}$$

Эту размерность можно определить в точке как:

Определение 4 Поточечная размерность. Пусть $B_r(x)$ — шар радиуса r с центром $x \in \mathbb{R}^N$ и P — вероятностная мера, такая что предел существует. Тогда поточечная размерность это

$$M = \lim_{r \rightarrow 0} \frac{\ln P(B_r(x))}{\ln r}$$

Видно, что эта размерность получена из размерности Хаусдорфа логарифмированием меры Хаусдорфа и заменой квадратных ячеек шарами. В примере со сферой поточечная размерность будет тоже 2. Достаточно популярной версией поточечной размерности является корреляционная размерность, в которой в качестве вероятностной меры взят корреляционный интеграл.

Определение 5 Корреляционная размерность. Пусть $C(r)$ корреляционный интеграл

$$C(r) = \frac{2}{|\Omega|(|\Omega| - 1)} \sum_{i < j} 1(|x_i - x_j| > r)$$

Тогда корреляционная размерность это

$$M = \lim_{r \rightarrow 0} \frac{\ln C(r)}{\ln r}$$

3. Описание методов

Среди методов, дающих оценку размерности данных, можно выделить три основные группы: глобальные, локальные и поточечные [1]. Глобальные методы предполагают, что данные лежат на одном многообразии с зафиксированной размерностью. К этой группе относятся большинство существующих методов. В частности, к этой группе относится наиболее часто используемый метод основанный на методе главных компонент, но он склонен давать завышенные оценки, поэтому в этой работе рассматривался метод Граната-Карневале [8], опирающийся на определение корреляционной размерности.

В глобальных методах присутствует предположение о том, что данные лежат на одном многообразии, однако для некоторых данных это может и не выполняться. В связи с этим возникают локальные методы. Эти методы предполагают, что данные могут лежать на нескольких многообразиях и поэтому методы оценивают размерность многообразия на нескольких подмножествах данных (у каждого подмножества своя размерность). В основном методы опираются на принцип многомерного шкалирования, то есть на предположение, что близкие точки в пространстве большой размерности являются близкими в пространстве более низкой размерности. В этой работе рассмотрен локальный метод основанный на алгоритме Isomap [5]. Стоит отметить, что способы оценки, основанные на многомерном шкалировании, используют аналогичные рассматриваемому способу процедуры.

С помощью поточечных методов можно оценивать как локальные размерности подмножеств данных, так и глобальные. Основное отличие этой группы методов в том, что размерность оценивается основываясь на определении поточечной размерности. Эта размерность вычисляется в окрестности каждой точки, а затем усредняется в группах точек или во всей выборке, и, в зависимости от этого, находится локальная или глобальная размерность соответственно. В данной работе рассмотрен поточечный метод Левиной-Бикеля [7].

Стоит отметить, что оценка может получаться не целым числом, но для

задачи снижения размерности можно её округлить. Рассмотрим упомянутые методы более подробно.

3.1. Метод Левиной-Бикеля

Алгоритм предполагает, что наблюдения $X_1, \dots, X_n \in \mathbb{R}^p$ это независимые и одинаково распределенные случайные величины такие, что $X_i = g(Y_i)$, где $Y_i \in \mathbb{R}^m$, $m \leq p$ распределены с некоторой гладкой плотностью f , а g — непрерывная и достаточно гладкая. Это гарантирует, что близкие Y_i отображаются в близкие X_i . Сделав такие предположения, можно применить метод максимального правдоподобия (ММП) для оценки параметра m .

Для того, чтобы применить ММП рассмотрим биномиальный процесс $N(x, t)$, считающий количество точек выборки, попавших в шар $S_x(R)$ с центром в x и радиусом R .

$$N(x, t) = \sum_{i=1}^n 1(X_i \in S_x(t))$$

Предполагая, что плотность $f(x)$ в малой окрестности x постоянна, можно аппроксимировать биномиальный процесс $N(t, x)$ процессом Пуассона $N(t)$ с параметром $\lambda(t)$. Из свойств процесса Пуассона следует:

$$\lambda(t) = f(x)V(m)mt^{m-1}$$

Так как площадь поверхности сферы это $V(m)mt^{m-1} = \frac{d}{dt}[V(m)t^m]$, где $V(m)$ объём единичного m -мерного шара. Взяв $\theta = f(x)$, можно написать функцию правдоподобия:

$$L(m, \theta) = \int_0^R \log \lambda(t) dN(t) - \int_0^R \lambda(t) dt$$

Из условий первого порядка следует оценка ММП:

$$\hat{m}_R(x) = \left(\frac{1}{N(x, R)} \sum_{j=1}^{N(x, R)} \log \frac{R}{T_j(x)} \right)^{-1}$$

Где $T_j(x)$ — это евклидово расстояние от x до j -го соседа x . На практике вместо радиуса R проще фиксировать количество соседей, тогда формула принимает вид:

$$\hat{m}_k(x) = \left(\frac{1}{k-1} \sum_{j=1}^k \log \frac{T_k(x)}{T_j(x)} \right)^{-1}$$

Мы получили оценку в точке, для оценки размерности выборки усредним по всем точкам выборки. Поскольку k тоже влияет на оценку, авторы предлагают усреднять и по количеству соседей. Итоговая оценка получается следующей:

$$\hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \frac{1}{n} \sum_{i=1}^n \hat{m}_k(X_i)$$

Где $k_1, k_2 \in \mathbb{N}$ являются параметрами метода (в экспериментах использовались $k_1 = 3, k_2 = 20$). Для тестирования алгоритма использовалась немного оптимизированная реализация [9].

3.2. Метод Isomap

Основная идея Isomap заключается в том, что расстояния на многообразии низкой размерности можно приблизить кратчайшими расстояниями на графе соседей. Вершинами этого графа являются точки выборки. Ребро между двумя точками выборки i, j появляется, либо, если евклидово расстояние между ними меньше или равно некоторого порога ϵ , либо i является одним из K ближайших соседей j . У каждого ребра есть вес равный евклидову расстоянию между точками. Алгоритм снижения размерности Isomap может находить низкоразмерные представления данных, принимая на вход размерность

и K или ϵ . Опишем схему алгоритма.

1. Строим граф соседей G
2. Вычисляем матрицу кратчайших расстояний D_G (например, с помощью алгоритма Дейкстры)
3. Строим низкоразмерное представление данных с помощью одного из методов многомерного шкалирования

Для оценивания размерности необходимо составить две матрицы: матрицу кратчайших расстояний на графе D_G и матрицу евклидовых расстояний между всеми точками $D(m)$, к которым применен метод снижения размерности Isomap для некоторой размерности m . Затем необходимо оценить корреляцию $\rho(D_G, D(m))$ между матрицами для нескольких размерностей и построить график дисперсии ошибок $1 - \rho^2(D_G, D(m))$ в зависимости от размерности m . Размерность определяется как такое m , при котором произошел резкий спад или, начиная с которого, значение дисперсии ошибок уже не сильно меняется.

Аналогичным способом можно оценивать размерность с помощью других методов снижения размерности. Для этого достаточно определить понятие ошибки аналогичной дисперсии ошибок в Isomap. Часто методы снижения размерности формулируются как оптимизационная задача, поэтому ошибку можно, например, измерять в значениях целевой функции для разных размерностей.

3.3. Метод Граната-Карневале

Метод основан на оценке размерности при помощи корреляционной размерности. Сначала уточняется выражение для корреляционного интеграла.

$$C(r) = \frac{2}{|\Omega|(|\Omega| - 1)} \sum_{i < j} H(r - \|x_i - x_j\|) =$$

$$\begin{aligned}
&= \frac{2}{|\Omega|(|\Omega| - 1)} \int_0^r dx \sum_{i < j} \delta(r - \|x_i - x_j\|) = \\
&= \int_0^r \hat{p}(x) dx
\end{aligned}$$

Где $H(x)$ — функция Хевисайда, $\delta(x)$ — дельта функция и $\hat{p}(x)$ — оценка плотности попарных расстояний ($\mathbb{E}\hat{p}(x) \rightarrow p(x)$, $|\Omega| \rightarrow \infty$). Затем корреляционная размерность расписывается по правилу Лопиталья:

$$M = \lim_{r \rightarrow 0} \frac{d \ln C(r)}{d \ln r} = \lim_{r \rightarrow 0} \frac{d \ln C(r)}{dr} r = \lim_{r \rightarrow 0} \frac{p(r)}{C(r)} r = \lim_{r \rightarrow 0} D(r)$$

Авторы утверждают, что $p(r)$ несёт такую же информацию о размерности, что и $D(r)$, поэтому концентрируют своё исследование на свойствах $p(r)$. Основная идея метода заключается в том, что на модельных данных графики зависимости $\ln p(r)/p_{max}$ от $\ln r/r_{max}$, где $p_{max} = p(r_{max}) = \max_r p(r)$, для поверхностей одной размерности практически совпадают в окрестности r_{max} . Авторы задают эту окрестность интервалом $[r_{max} - 2s, r_{max}]$, где s — стандартное отклонение попарных расстояний. Исходя из этого предположения, можно взять в качестве эталонной плотности — плотность на M -мерной сфере, то есть:

$$p(r) = C \sin^{M-1}(r), C = const$$

Предлагается два способа искать размерность. Первый заключается в нахождении M (D_{fit}) методом наименьших квадратов, используя гистограмму в качестве оценки плотности. Второй предполагает подбирать размерность сферы (D_{min}) так, чтобы корень из среднего значения квадратов разностей (RMDS) между гистограммой и распределением на сфере был минимальным. Стоит отметить, что плотность на сфере и оценка плотности (гистограмма) сравниваются на интервале $[r_{max} - 2s, r_{max}]$. Для тестирования алгоритма использовалась немного исправленная реализация [10].

4. Численные эксперименты

Для тестирования описанных выше методов использовались как модельные данные с известной размерностью, так и реальные данные. Для реальных данных нет точного значения для внутренней размерности, поэтому для определения их размерности лучше использовать несколько методов, чтобы определить границы, в которых наиболее вероятно лежит истинная размерность.

В качестве модельных данных были сгенерированы несколько наборов данных. Во-первых, были сгенерированы выборки, по 1000 точек в каждой, с поверхностей сфер в пространствах размерности от 2 до 30. Истинная размерность таких выборок на 1 меньше размерности пространства. Во-вторых, было сгенерировано 3 выборки по 200 точек с трёхмерных поверхностей: Swiss roll (фигура, полученная при движении спирали по одной из осей), moons (два смещённых полуцилиндра) и S curve (фигура, полученная при движении буквы S по одной из осей). Истинные размерности этих поверхностей равны двум. Также для этих трёх поверхностей были сгенерированы точки, к которым был добавлен шум.

В качестве реальных данных были рассмотрены шесть датасетов:

- MNIST — чёрно-белые изображения рукописных цифр размера 28×28
- Iris — характеристики размеров ирисов
- Diabetes — записи показателей анализов больных диабетом
- Boston house-prices — характеристики недвижимости в Бостоне
- Olivetti faces — чёрно-белые фотографии лиц людей размера 20×20
- California housing — характеристики недвижимости в Калифорнии

4.1. Применение методов к многомерным сферам

Рассмотрим, как алгоритмы работают на данных сгенерированных с поверхностей сфер. Запишем результаты в таблицу 1. Метод Левиной-Бикеля хо-

рошо работает для размерностей меньших 10, но достаточно сильно занижает оценку при больших значениях истинной размерности. Как можно видеть на рисунке 1, резких скачков не происходит, поэтому метод Isomap выдаёт размерность пространства. Метод Граната-Карнивале иногда завышает, иногда занижает оценку, но, в целом, работает лучше метода Isomap на размерностях не выше 21. По результатам видно, что метод Isomap всегда даёт завышенную оценку, в то время как метод Левиной-Бикеля всегда даёт оценку снизу. По данному эксперименту можно сделать вывод, что при маленьких размерностях, лучше работает метод Левиной, при размерностях от 10 до 21 лучше работает метод Граната-Карнивале, а при больших размерностях лучше использовать Isomap.

Таблица 1: Оценки размерности многомерных сфер разными алгоритмами

Размерность пространства	2	3	4	5	6	7	8	9	10	11
Левиной	0.99	2.01	3.01	3.96	5.05	5.7	6.6	7.5	7.99	8.85
Isomap	2	3	4	5	6	7	8	9	10	11
D_{min}	-	3.00	3.00	5.00	6.00	8.00	9.00	11.00	12.00	14.00
D_{fit}	-	2.86	3.43	4.76	6.22	7.69	9.22	10.57	11.99	13.55
Размерность пространства	12	13	14	15	16	17	18	19	20	21
Левиной	9.58	10.41	10.9	11.67	12.42	12.85	13.37	14.36	14.75	15.18
Isomap	12	13	14	15	16	17	18	19	20	21
D_{min}	15.00	13.00	14.00	15.00	16.00	16.00	17.00	18.00	19.00	20.00
D_{fit}	14.99	13.23	14.16	14.83	15.74	16.43	17.08	18.01	18.57	19.57
Размерность пространства	22	23	24	25	26	27	28	29	30	
Левиной	15.42	16.49	16.6	17.34	18.12	18.36	19.8	19.7	20.06	
Isomap	22	23	24	25	26	27	28	29	30	
D_{min}	20.00	21.00	21.00	22.00	23.00	23.00	24.00	24.00	25.0	
D_{fit}	19.85	20.69	21.07	21.66	22.99	23.27	23.52	23.92	25.27	

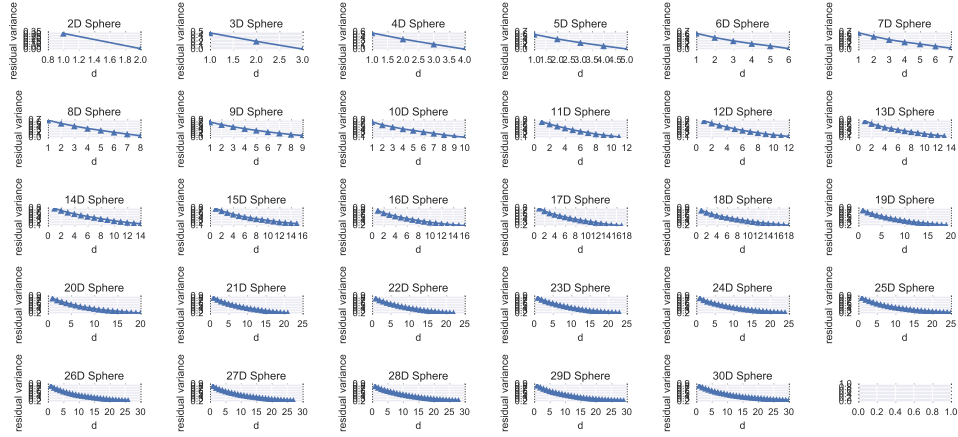


Рис. 1: Применение метода Isomap к многомерным сферам

4.2. Применение к Swiss roll, Moons, S curve

Для данных Swiss roll, Moons и S curve истинная размерность равна 2. Результаты применения методов приведены в таблице 2

Таблица 2: Оценки размерности для Swiss roll, Moons, S curve. В скобках результат для зашумлённых данных

Данные	Левиной	Isomap	D_{fit}	D_{min}	Размер выборки	Число признаков
Swiss roll	2.02 (2.46)	2 (2)	2.73 (2.71)	3 (3)	200	3
Moons	1.91 (2.49)	2 (2)	2.68 (2.41)	3 (2)	200	3
S curve	1.91 (2.48)	2 (2)	2.68 (2.42)	3 (2)	200	3

По результатам экспериментов видно, что метод Левиной-Бикеля даёт правильные результаты, однако подвержен влиянию шума. Метод Isomap (рисунок 2) имеет перегиб в точке 2, поэтому даёт точный результат. Метод Граната-Карневале выдаёт завышенную оценку, однако на Moons и S curve с шумом оценка совпадает с истинной. Возможно, это происходит, потому что гистограмма не достаточно гладкая или на гистограмме существуют локальные максимумы. Авторы предлагают смотреть на гистограмму, график $\ln p(r)/p_{max}$ от $\ln r/r_{max}$ и график значений RMDS от размерности, чтобы

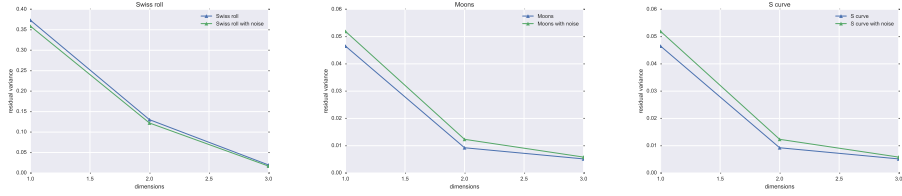


Рис. 2: Применение метода Isomap к Swiss roll, Moons, S curve

оценивать качество оценки. Рассмотрим такие графики для Moons с шумом (рисунок 3). По гистограмме можно судить о гладкости, по графику видно, что гистограмма не очень гладкая и имеет локальные максимумы. По второму графику видно, что оценка плотности не совпадает с плотностью сферы в окрестности максимума, поэтому результат работы метода может быть не очень точный. По последнему графику видно, что минимум достигается в точке 2, но значение в точке 3 не намного больше, поэтому алгоритм и ошибается на данных без шума.

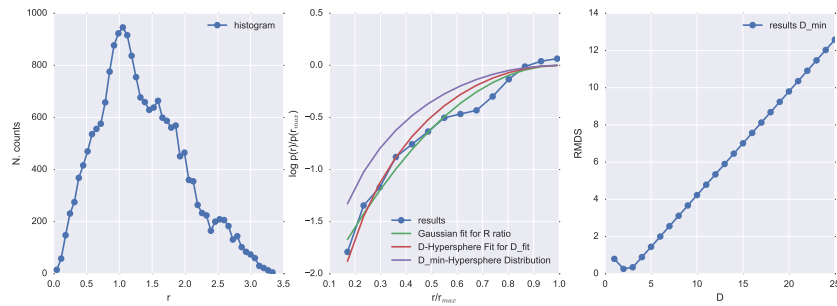


Рис. 3: Гистограмма, график $\ln p(r)/p_{max}$ от $\ln r/r_{max}$ и значение RMDS от размерности для Moons с шумом

4.3. Применение к реальным данным

Для рассмотренных данных результаты представлены в таблице 3.

Поскольку истинная размерность для реальных данных не известна, по-

Таблица 3: Оценки размерности для реальных данных

Данные	Левиной	Isomap	D_{fit}	D_{min}	Размер выборки	Число признаков
MNIST	12.69	13	21.63	22	1000	784
Iris	3.09	4	1.61	2	150	4
Diabetes	6.09	6	6.62	7	422	10
Boston house-prices	2.73	8	1.78	2	506	13
Olivetti	7.23	12	15.49	15	400	4096
California housing	3.53	5	-	-	20640	8

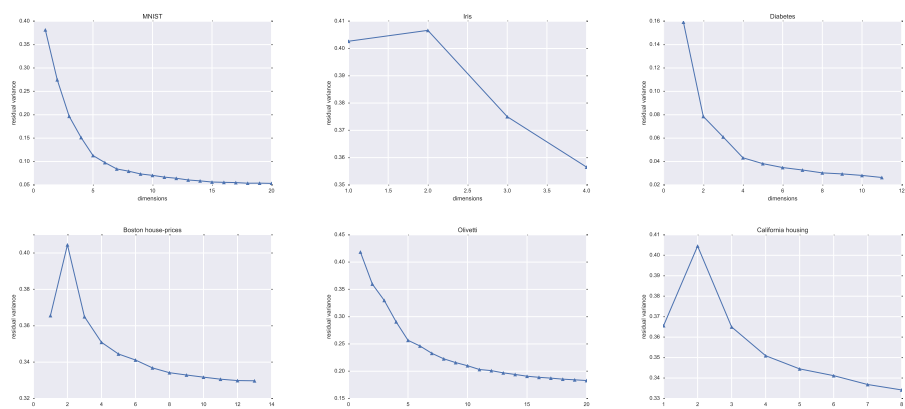


Рис. 4: Применение метода Isomap к реальным данным

попробуем оценить эту размерность, основываясь на свойствах методов, полученных ранее. При этом не будем рассматривать результат Граната-Карневале (Г-К), если он выше чем результат Isomap. То есть будем считать результат метода Isomap (рисунок 4) за верхнюю границу, а результат метода Левиной-Бикеля за нижнюю границу. Для данных MNIST и Diabetes верхняя и нижняя границы совпадают поэтому можно взять 13 и 6 в качестве истинных оценок. Для данных Isis и California housing результаты отличаются на 1. Для данных Iris Г-К оценка меньше оценок других методов, при этом Isomap выдаёт просто размерность пространства, поэтому можно взять в качестве истинной размерности 3. Для данных California housing Г-К оценка не была получена, потому что из-за большого объёма выборки метод требует очень большого

количества оперативной памяти. Поскольку верхняя оценка достаточно низкая, можно доверять методу Левиной-Бикеля и утверждать, что истинная размерность равна 4. По такому же принципу можно сказать, что размерность Boston house-prices равна 3. Для данных Olivetti оценки достаточно сильно разнятся. Можно только сказать, что размерность лежит в интервале $[7, 12]$. Для уточнения оценки, например, в задачах обучения с учителем лучше принимать решение на основе кросс-валидации для размерностей от 7 до 12.

5. Заключение

В работе были рассмотрены три метода оценивания размерности: метод Левиной-Бикеля, метод Isomap и метод Граната-Карневале. Первый метод показал хорошие результаты на модельных данных при размерностях не больше 10, на больших размерностях метод выдаёт нижнюю границу размерности. Стоит также отметить, что этот метод работал быстрее остальных. В методе Isomap необходимо оценивать размерность по графику зависимости дисперсии ошибки от размерности. Дисперсия ошибки убывает с увеличением размерности, поэтому часто у графика может и не быть излома, но при больших размерностях она изменяется значительно меньше чем при маленьких. Исходя из этого оценка может быть завышенной, что было видно на модельных данных. Последний метод основан на империческом наблюдении поведения функций плотности около максимального значения. Из-за этого метод часто выдаёт не точные значения, но близкие к истинным. На метод можно полагаться, если истинная размерность примерно от 10 до 20, что можно определить, например, по верхней оценке метода Isomap.

В заключении стоит отметить, что поскольку нет методов, всегда дающих точную оценку, особенно для больших размерностей, лучше использовать несколько методов для оценивания. Среди них полезно брать методы дающие верхние и нижние границы. Если эти границы близки вывод о размерности можно сделать почти однозначно, иначе лучше смотреть на показатели

качества оценки. Например, в методе Граната-Карневале можно смотреть на гладкость и похожесть графиков плотности. Если по таким показателям тоже сложно что-либо сказать и снижение размерности делается для задачи обучения с учителем можно использовать кросс-валидацию по интервалу, полученному из верхней и нижней границы. В любом случае методы оценивания размерности помогают локализовать истинное значение размерности, что значительно ускоряет и упрощает выбор оптимальных параметров для методов снижения размерности.

Список литературы

- [1] F. Camastra and A. Staiano, “Intrinsic dimension estimation: Advances and open problems,” 2015.
- [2] E. Chavez, G. Navarro, R. Baeza-Yates, and J. L. Marroquin, “Searching in metric spaces,” *ACM Comput. Surv.*, 2001.
- [3] R. Bellman, “Adaptive control processes: A guided tour,” *Princeton University Press*, 1961.
- [4] C. Bishop, “Neural networks for pattern recognition,” *Cambridge University Press*, 1995.
- [5] J. B. Tenenbaum, V. de Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, 2000.
- [6] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, 2000.
- [7] E. Levina and P. Bickel, “Maximum likelihood estimation of intrinsic dimension,” *Advances in Neural Information Processing Systems*, 2005.

- [8] D. Granata and V. Carnevale, “Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets,” *Sci. Rep.*, 2016.
- [9] Levina-bickel method implementation. [Online]. Available: https://codegists.com/snippet/python/intdim_mlepy_mehdide_python
- [10] Granata-carnevale method implementation. [Online]. Available: <https://github.com/dgranata/Intrinsic-Dimension>