Конспект курса Geometric Methods of Machine Learning

Михаил Кузнецов (@mmkuznecov)

Содержание

1 Классификация методов понижения размерности

1.1 Линейные методы

Предполагают, что данные лежат в линейном подпространстве меньшей размерности:
- Метод главных компонент (PCA)
- Метод независимых компонент (ICA)
- Преследование проекций (Projection Pursuit)

1.2 Нелинейные методы

1.2.1 Ядерные методы

Используют преобразование данных в пространство признаков с помощью ядерных функций:
- Ядерный метод главных компонент (Kernel PCA)

1.2.2 Нейросетевые методы

Используют нейронные сети для нелинейного понижения размерности:
- Автоассоциативные нейронные сети (автоэнкодеры)

1.2.3 Методы многообразий (Manifold Learning)

Предполагают, что данные лежат на или вблизи нелинейного многообразия низкой размерности:

Методы сохранения локальной геометрии:
- Локально-линейное вложение (LLE)
- Собственные карты Лапласа (Laplacian Eigenmaps)

Методы сохранения расстояний:
- ISOmetric MAPping (ISOMAP)

Методы согласования вероятностных распределений:
- Стохастическое вложение соседей (SNE)
- t-распределенное стохастическое вложение соседей (t-SNE)
- Равномерное приближение и проекция многообразия (UMAP)

Методы сохранения дифференциальной структуры:
- Римановы многообразия / Log-map
- Грассманово и Штифелево отображение собственных значений (Grassmann & Stiefel Eigenmaps)

1.2.4 Топологические методы

Используют методы топологии для характеристики формы данных:
- Топологический анализ данных (TDA)

2 Сводная таблица методов понижения размерности

Метод	Основная идея	Преимущества	Недостатки	Примечания
PCA (Метод главных компонент)	Находит ортогональные направления максимальной дисперсии в данных	• Простота и эффективность • Сохраняет глобальную структуру • Устраняет корреляцию	• Эффективен только для линейных данных • Чувствителен к масштабированию • Предполагает линейность структуры данных	Можно рассматривать как оптимальное линейное приближение, оптимальное решение линейного понижения размерности, максимизацию дисперсии

Метод	Основная идея	Преимущества	Недостатки	Примечания
ICA (Метод независимых компонент)	Находит статистически независимые компоненты в данных	• Эффективно разделяет смешанные сигналы • Подходит для задач слепого разделения источников • Учитывает независимость, а не только отсутствие корреляции	• Неприменим к гауссовым источникам (не более одного) • Требует предположения о независимости • Не определяет порядок компонент	Часто используется в обработке сигналов и изображений (например, проблема “коктейльной вечеринки”)
Projection Pursuit	Находит интересные низкоразмерные проекции оптимизацией индекса проекции	• Гибкая структура для разных типов проекций • Может обнаруживать кластеры и выбросы • Применим для нелинейных данных	• Сложная оптимизация • Требует выбора индекса проекции • Вычислительно затратный	PCA и ICA можно рассматривать как частные случаи с конкретными индексами проекции
Kernel PCA	Применяет PCA в неявном пространстве признаков высокой размерности с использованием ядерного трюка	• Обрабатывает нелинейные данные • Вычислительно эффективен за счёт ядерного трюка • Хорошо работает с данными сложной структуры	• Сложный выбор ядра • Трудная интерпретация • Проблемы с масштабируемостью для больших наборов данных	Преобразует нелинейные данные, делая их более линейными в пространстве признаков

Метод	Основная идея	Преимущества	Недостатки	Примечания
Автоассоциативные нейронные сети (автоэнкодеры)	Нейросеть с узким внутренним слоем, обученная воспроизводить вход через представление меньшей размерности	• Может изучать сложные нелинейные отображения • Гибкая архитектура • Может работать с любыми типами данных	• Сложность обучения • Склонность к локальным минимумам • Требует настройки гиперпараметров	Современные варианты включают вариационные, сверточные и разреженные автоэнкодеры
LLE (Локально-линейное вложение)	Сохраняет локальную геометрию, представляя точки как линейные комбинации соседей	• Сохраняет локальную структуру • Минимум параметров (только размер окрестности) • Хорошо работает с гладкими многообразиями	• Чувствителен к размеру окрестности • Плохо обрабатывает “дыры” • Проблемы с масштабируемостью	Основан на идее, что локально многообразие можно аппроксимировать линейным подпространством
ISOMAP	Расширяет MDS, используя геодезические расстояния вместо евклидовых	• Сохраняет глобальную структуру многообразия • Обрабатывает нелинейные данные • Теоретически обоснован	• Чувствителен к шуму • Требует, чтобы многообразие было изометрично евклидову пространству • Проблемы с “дырами” в данных	Использует графовые расстояния как приближение к геодезическим

Метод	Основная идея	Преимущества	Недостатки	Примечания
Laplacian Eigenmaps	Аппроксимирует многообразие графом смежности, сохраняет отношения близости	• Сильное теоретическое обоснование • Сохраняет локальную структуру • Связан с дифференциальными операторами на многообразиях	• Сохраняет только локальную структуру • Не показывает глобальную структуру • Требует построения графа	Основан на операторе Лапласа-Бельтрами на многообразиях
SNE (Стохастическое вложение соседей)	Преобразует сходства в условные вероятности, сопоставляет распределения в исходном и вложенном пространствах	• Хорошо сохраняет локальную структуру • Эффективен для визуализации • Выявляет кластеры	• Вычислительно интенсивен • “Проблема скученности” • Трудности с выбором перплексии	Базовый метод для t-SNE и UMAP
t-SNE	Модификация SNE, использующая t-распределение в пространстве меньшей размерности	• Превосходные результаты визуализации • Сохраняет кластеры • Решает “проблему скученности”	• Вычислительно затратный • Акцент на локальной структуре • Сложность интерпретации	Стандартный метод визуализации высокоразмерных данных
UMAP	Использует иную вероятностную модель, лучше сохраняет глобальную структуру	• Быстрее t-SNE • Лучше сохраняет глобальную структуру • Теоретически обоснован (алгебраическая топология)	• Сложный алгоритм • Много параметров • Сложная интерпретация	Более новый метод (2018), становится популярным как альтернатива t-SNE

Метод	Основная идея	Преимущества	Недостатки	Примечания
Riemannian Manifold Learning	Использует римановы нормальные координаты для представления точек многообразия	• Хорошо сохраняет геометрию многообразия • Учитывает кривизну • Теоретически обоснован	• Чувствителен к выбору базовой точки • Проблемы с границей среза • Сложен в реализации	Представляет точки их геодезическим расстоянием и направлением от базовой точки
Grassmann & Stiefel Eigenmaps	Решает задачу обучения касательного расслоения многообразия, сохраняя как точки, так и касательные пространства	• Точная реконструкция • Сохраняет дифференциальную структуру • Эффективен для последующих задач регрессии на многообразии	• Вычислительно сложный • Труден для понимания • Требует большой выборки	Трехэтапный метод: аппроксимация касательных пространств, вложение многообразия, реконструкция
TDA (Топологический анализ данных)	Количественно характеризует форму данных, используя топологические методы	• Захватывает глобальные особенности формы • Устойчив к шуму • Инвариантен к деформациям	• Абстрактное представление • Вычислительно интенсивный • Сложная интерпретация	Использует фильтрацию и персистентную гомологию для выявления топологических особенностей

3 Иерархия и классификация методов

```
1. Линейные методы
    PCA (максимизация дисперсии)
    ICA (статистическая независимость)
    Projection Pursuit (общая структура)

2. Нелинейные методы
    2.1. Ядерные методы
        Kernel PCA

    2.2. Нейросетевые методы
        Автоассоциативные нейронные сети (автоэнкодеры)

    2.3. Методы обучения многообразий
```

2.3.1. Сохранение локальной геометрии
Locally Linear Embedding (LLE)
Laplacian Eigenmaps

2.3.2. Сохранение расстояний
ISOMAP (геодезические расстояния)

2.3.3. Согласование вероятностных распределений
SNE
t-SNE
UMAP

2.3.4. Сохранение дифференциальной структуры
Riemannian Manifold Learning
Grassmann & Stiefel Eigenmaps

2.4. Топологические подходы
Топологический анализ данных (TDA)

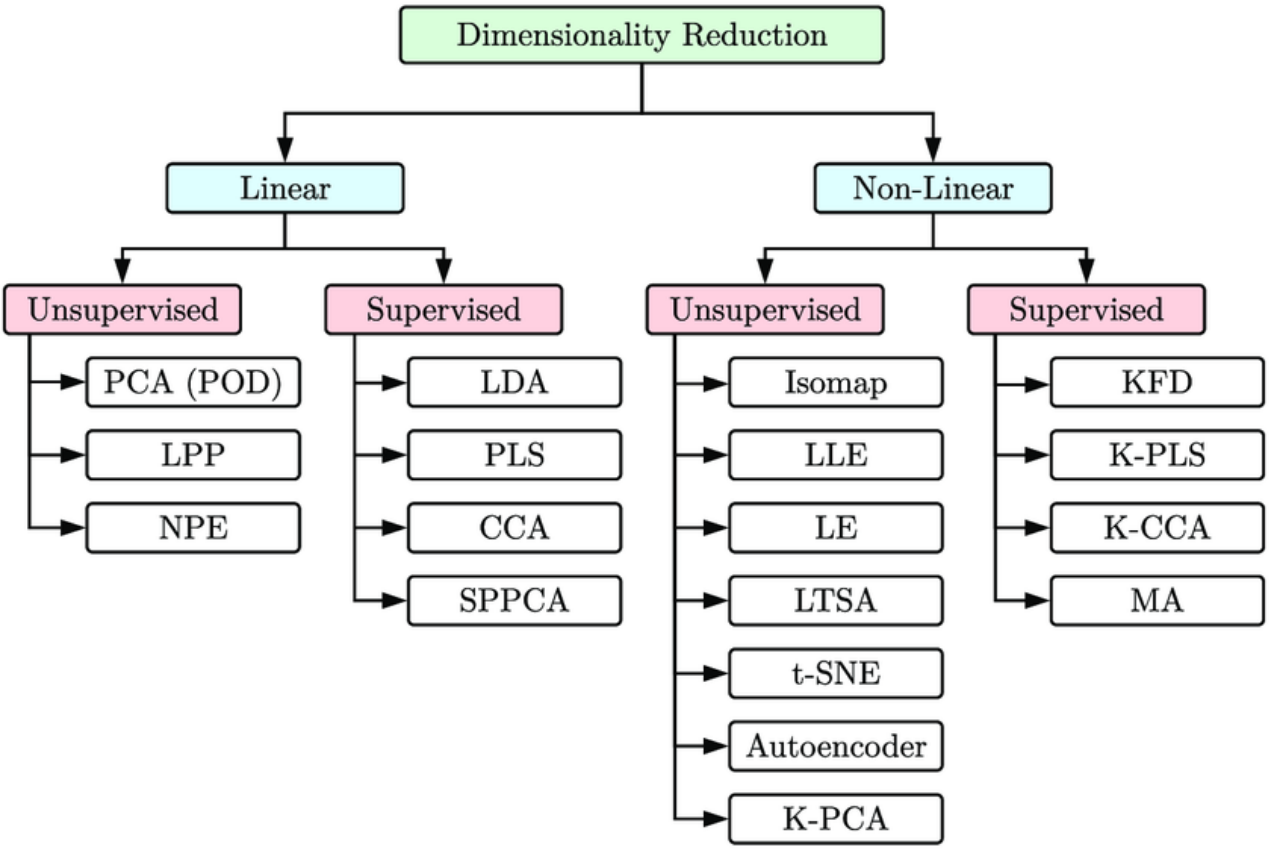

Рис. 1: Таксономия методов понижения размерности

4 Ответы на билеты

4.1 Empty space phenomenon: details and examples

Явление пустого пространства – фундаментальная проблема в геометрии высокоразмерных пространств, заключающаяся в том, что большая часть объема концентрируется вблизи границы, оставляя "середину" пространства практически пустой.

4.1.1 Математическая формализация

Объем гиперкуба и вписанного шара: Для p-мерного куба $C(R,p) = [-R,R]^p$ с объемом $V(C(R,p)) = (2R)^p$ и вписанного шара $B(R,p) = \{x \in \mathbb{R}^p : \|x\| \leq R\}$ с объемом $V(B(R,p)) = \frac{\pi^{p/2}R^p}{\Gamma(\frac{p}{2}+1)}$:

$$\lim_{p\to\infty} \frac{V(B(R,p))}{V(C(R,p))} = 0$$

Например:
- При $p = 6$: отношение ≈ 0.1 (шар занимает только 10% объема куба)
- При $p = 10$: отношение ≈ 0.0025 (0.25%)

Концентрация объема на границе: Для сферической оболочки толщины ε вблизи поверхности единичного шара:

$$\lim_{p\to\infty} \frac{V(B(1,p)) - V(B(1-\varepsilon,p))}{V(B(1,p))} = 1$$

Этот результат показывает, что почти весь объем шара сконцентрирован в тонком сферическом слое вблизи поверхности.

Диагонали гиперкуба: Для p-мерного куба $C(1,p) = [-1,1]^p$, угол θ_p между любой полудиагональю E (от центра до вершины) и произвольной координатной осью E_k:

$$\cos\theta_p = \frac{(E,E_k)}{\|E\| \cdot \|E_k\|} = \pm\frac{1}{\sqrt{p}}$$

При $p \to \infty$, $\theta_p \to \frac{\pi}{2}$, что означает: полудиагонали становятся почти ортогональными ко всем координатным осям.

4.1.2 Последствия в анализе данных

1. **Разреженность выборки:** Высокоразмерное пространство $\mathbb{R}^p$ очень "просторно" - для покрытия значительной части объема требуется экспоненциально большое количество точек с ростом размерности p.

2. **Проблемы с ближайшими соседями:** В высоких размерностях расстояния между точками "выравниваются", что усложняет различение близких и далеких соседей:

$$\frac{\max_{x,y\in X} \|x-y\| - \min_{x,y\in X, x\neq y} \|x-y\|}{\min_{x,y\in X, x\neq y} \|x-y\|} \to 0 \text{ при } p \to \infty$$

3. **Эффект проекции:** Проекция кластеров данных, лежащих вблизи диагоналей гиперкуба, на координатные оси приводит к их отображению вблизи начала координат, что искажает восприятие истинной структуры данных.

4.1.3 Примеры

1. **Равномерное распределение в гиперкубе:** Если $X \sim \mathcal{U}[-1,1]^p$, то вероятность, что точка находится в ε-окрестности центра:

$$P(\|X\| \leq \varepsilon) \to 0 \text{ при } p \to \infty \text{ для любого фиксированного } \varepsilon$$

2. **Пример с изображениями лиц:** Лица, описываемые векторами 10^6 пикселей (1024×1024 изображения), занимают лишь крошечную часть всего пространства изображений, сконцентрированную вблизи некоторого многообразия низкой размерности (на практике порядка 100).

3. **Явление в робототехнике:** При локализации робота по панорамному изображению (размерность 163,840), фактически изображения лежат на 3-мерном многообразии, соответствующем положению и ориентации робота.

4.2 Principal Component Analysis as the best linear approximation, including the solution to the Eigenvector problem

PCA как оптимальное линейное приближение

Дано: высокоразмерный датасет $\{X_1, X_2, \ldots, X_n\} \subset \mathbb{R}^p$

Задача: найти линейное аффинное подпространство $L(q)$ размерности $q < p$, которое наилучшим образом аппроксимирует данные.

Формально: минимизировать функционал

$$J(L(q)) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \mathrm{Pr}_{L(q)}(X_i)\|^2$$

где $\mathrm{Pr}_{L(q)}(X_i)$ - ортогональная проекция точки X_i на подпространство $L(q)$.

Решение:

1. $L(q) = L(q, \bar{X}, E)$ - аффинное q-мерное подпространство:
   - проходящее через точку $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (среднее)
   - натянутое на ортонормированные векторы $\{e_1, e_2, \ldots, e_q\}$
2. Ортогональная проекция: $\mathrm{Pr}_{L(q)}(X) = \bar{X} + \sum_{k=1}^{q}(X - \bar{X}, e_k) \times e_k$
3. Функционал можно переписать как:

$$J(L(q, \bar{X}, E)) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \bar{X}\|^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}((X_i - \bar{X}, e_k))^2$$

4. Минимизация J эквивалентна максимизации:

$$\Phi(E) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{q}((X_i - \bar{X}, e_k))^2 = \sum_{k=1}^{q} e_k^T \Sigma e_k$$

   где $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$ - выборочная ковариационная матрица.

5. **Решение проблемы собственных векторов**:
   - Максимизация квадратичной формы $\Phi(E)$ при ограничении $E^T E = I_q$

- Решение: столбцы $e_1, e_2, \ldots, e_q$ матрицы E_{PCA} - собственные векторы матрицы Σ, соответствующие q наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$
- $\Sigma e_k = \lambda_k e_k$, $k = 1, 2, \ldots, q$
- Максимальное значение функционала: $\max \Phi(E) = \sum_{k=1}^{q} \lambda_k$

6. Оптимальное подпространство: $L_{PCA}(q) = L(q, \bar{X}, E_{PCA})$

4.3 Principal Component Analysis from Singular Value Decomposition technique

PCA через сингулярное разложение (SVD)

1. Центрированная матрица данных: $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n) \in \mathbb{R}^{p \times n}$, где $\bar{X}_i = X_i - \bar{X}$

2. Сингулярное разложение (SVD): $\bar{\mathbf{X}} = U_p \Sigma_p V_p^T$:
   - U_p - ортогональная $p \times p$ матрица с ортонормированными столбцами $e_1, e_2, \ldots, e_p$
   - Σ_p - диагональная $p \times p$ матрица с неотрицательными элементами $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$
   - V_p - ортогональная $n \times p$ матрица: $V_p^T V_p = I_p$

3. Связь с ковариационной матрицей:

$$\Sigma = \frac{1}{n} \bar{\mathbf{X}} \bar{\mathbf{X}}^T = \frac{1}{n} U_p \Sigma_p^2 U_p^T$$

4. Собственные значения и собственные векторы:
   - Собственные значения ковариационной матрицы: $\lambda_k = \frac{\sigma_k^2}{n}$
   - Собственные векторы ковариационной матрицы: столбцы e_k матрицы U_p

5. PCA-представление: $Y_{PCA} = E_{PCA}^T \bar{\mathbf{X}} = \Sigma_q V_q^T$, где:
   - E_{PCA} - $p \times q$ матрица, составленная из первых q собственных векторов $\{e_1, e_2, \ldots, e_q\}$
   - Σ_q - диагональная $q \times q$ матрица с первыми q сингулярными числами
   - V_q - первые q столбцов матрицы V_p

4.4 Principal Component Analysis as a solution to the Dimensionality reduction problem

PCA как решение задачи снижения размерности

Задача снижения размерности: найти отображения $h : \mathbb{R}^p \to \mathbb{R}^q$ и $g : \mathbb{R}^q \to \mathbb{R}^p$, минимизирующие ошибку восстановления:

$$\varepsilon(h, g) = \left(\frac{1}{n} \sum_{i=1}^{n} \|\hat{X}_i - X_i\|^2 \right)^{1/2}$$

где $\hat{X}_i = g(h(X_i))$ - восстановленные данные.

Линейное снижение размерности:

1. Линейное отображение вложения: $y_i = W(X_i - \bar{X}) \in \mathbb{R}^q$, где W - ортогональная $q \times p$ матрица: $WW^T = I_q$

2. Линейное отображение восстановления: $\hat{X}_i = \bar{X} + V y_i$, где V - $p \times q$ матрица

3. Минимизация ошибки восстановления:
   - Оптимальная матрица восстановления при заданной W: $V(W) = W^T$
   - Восстановленное значение: $\hat{X}_i(W) = \bar{X} + W^T W(X_i - \bar{X}) = \bar{X} + P(W)(X_i - \bar{X})$ где $P(W) = W^T W$ - ортогональная проекционная матрица: $P^T(W) = P^2(W) = P(W)$

4. Квадрат ошибки:

$$\varepsilon^2(W) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \bar{X} - P(W)(X_i - \bar{X})\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|X_i - \bar{X}\|^2 - \mathrm{Tr}(W\Sigma W^T)$$

5. Минимизация $\varepsilon^2(W)$ эквивалентна максимизации $\mathrm{Tr}(W\Sigma W^T)$

6. Решение: $W_{PCA} = E_{PCA}^T$, где E_{PCA} - матрица собственных векторов ковариационной матрицы Σ, соответствующих q наибольшим собственным значениям.

4.5 Principal Component Analysis as the best solution to Metric Multi Dimensionality Scaling

PCA как оптимальное решение метрического многомерного шкалирования (MDS)

MDS: найти низкоразмерные представления $\{y_1, y_2, \ldots, y_n\} \subset \mathbb{R}^q$, сохраняющие попарные евклидовы расстояния:

$$\mathrm{APD}(W) = \sum_{i,j=1}^{n} \left(\|X_i - X_j\|^2 - \|y_i - y_j\|^2\right)^2 \to \min$$

Учитывая, что $\|X_i - X_j\|^2 = \|X_i - \bar{X}\|^2 + \|X_j - \bar{X}\|^2 - 2(X_i - \bar{X}, X_j - \bar{X})$, задача MDS эквивалентна минимизации:

$$\mathrm{MDS}(W) = \|\bar{\mathbf{X}}^T\bar{\mathbf{X}} - Y^T Y\|_F^2$$

где Y - матрица с низкоразмерными координатами: $Y = W\bar{\mathbf{X}}$

Решение SVD для матрицы $\bar{\mathbf{X}}^T\bar{\mathbf{X}}$:

$$\bar{\mathbf{X}}^T\bar{\mathbf{X}} = V_p\Sigma_p^2 V_p^T$$

Оптимальное приближение ранга q:

$$\bar{\mathbf{X}}^T\bar{\mathbf{X}} \approx V_q\Sigma_q^2 V_q^T = (\Sigma_q V_q^T)^T(\Sigma_q V_q^T)$$

Оптимальное решение MDS:

$$Y_{MDS} = \Sigma_q V_q^T = W_{MDS}\bar{\mathbf{X}} = W_{PCA}\bar{\mathbf{X}}$$

Таким образом, $W_{MDS} = W_{PCA}$, что показывает эквивалентность PCA и метрического MDS.

4.6 Principal Component Analysis: as Maximum variance preserving

PCA как метод максимального сохранения дисперсии

PCA можно рассматривать как поиск направлений, вдоль которых наблюдается максимальная дисперсия данных:

1. Задача нахождения первой главной компоненты:
   - Найти единичный вектор e_1, максимизирующий $\mathrm{Var}(e_1^T X) = e_1^T \Sigma e_1$
   - Решение: e_1 - собственный вектор ковариационной матрицы Σ, соответствующий наибольшему собственному значению λ_1
   - Максимальная дисперсия: $\mathrm{Var}(e_1^T X) = \lambda_1$
2. Задача нахождения k-ой главной компоненты ($k > 1$):
   - Найти единичный вектор e_k, ортогональный ко всем предыдущим $\{e_1, e_2, \ldots, e_{k-1}\}$, максимизирующий $\mathrm{Var}(e_k^T X) = e_k^T \Sigma e_k$
   - Решение: e_k - собственный вектор матрицы Σ, соответствующий k-му по величине собственному значению λ_k
   - Дисперсия k-ой главной компоненты: $\mathrm{Var}(e_k^T X) = \lambda_k$
3. Общая дисперсия данных и доля сохраненной дисперсии:
   - Полная дисперсия исходных данных: $\sum_{k=1}^{p} \lambda_k = \mathrm{Tr}(\Sigma)$
   - Дисперсия, сохраненная в первых q главных компонентах: $\sum_{k=1}^{q} \lambda_k$
   - Доля сохраненной дисперсии: $\frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$
4. Свойства PCA-преобразования:
   - Декорреляция: $\mathrm{Cov}(y_i, y_j) = 0$ для $i \neq j$
   - Упорядочение по дисперсии: $\mathrm{Var}(y_1) \geq \mathrm{Var}(y_2) \geq \ldots \geq \mathrm{Var}(y_q)$
   - Оптимальность: из всех линейных преобразований, PCA обеспечивает максимальное сохранение дисперсии при заданной размерности q

4.7 Independent component analysis: assumptions and general approach

Метод независимых компонент (ICA): предположения и общий подход

ICA предполагает, что наблюдаемые данные $X \in \mathbb{R}^p$ являются линейной смесью независимых источников $S \in \mathbb{R}^q$:

$$X = AS$$

где A - неизвестная $p \times q$ матрица смешивания.

4.7.1 Основные предположения

1. Компоненты источника $s_1, s_2, \ldots, s_q$ статистически независимы: $p(s_1, s_2, \ldots, s_q) = p_1(s_1) \cdot p_2(s_2) \cdot \ldots \cdot p_q(s_q)$
2. Не более одной компоненты может иметь гауссово распределение
3. Число наблюдаемых сигналов не меньше числа источников: $p \geq q$

4.7.2 Ограничения и неоднозначности

1. Невозможно определить порядок независимых компонент
2. Невозможно определить масштаб (дисперсию) независимых компонент

4.7.3 Общий подход

1. Предварительная обработка данных:
   - Центрирование: $X \leftarrow X - \mathbb{E}[X]$
   - Отбеливание: $X_V = VX$, где V - матрица отбеливания: $\mathrm{Cov}(X_V) = I$
2. Поиск разделяющей матрицы W, такой что $\hat{S} = WX$ дает оценку независимых компонент
   - После отбеливания, W - ортогональная матрица: $WW^T = I$
3. Критерии для нахождения W:
   - Максимизация негауссовости компонент
   - Максимизация правдоподобия
   - Минимизация взаимной информации
4. Оптимизация выбранного критерия итеративными методами (градиентный спуск, FastICA)

4.8 Independent component analysis based on Kurtosis

ICA на основе куртозиса

Куртозис как мера негауссовости случайной величины ξ с $\mathbb{E}[\xi] = 0$ и $\mathbb{E}[\xi^2] = 1$:

$$\mathrm{Kurt}(\xi) = \mathbb{E}[\xi^4] - 3(\mathbb{E}[\xi^2])^2 = \mathbb{E}[\xi^4] - 3$$

Свойства куртозиса:
- Для гауссовой случайной величины: $\mathrm{Kurt}(\mathrm{Gauss}) = 0$
- Для супергауссовых распределений (с “тяжелыми хвостами”): $\mathrm{Kurt}(\xi) > 0$
- Для субгауссовых распределений: $\mathrm{Kurt}(\xi) < 0$

Подход к ICA через максимизацию куртозиса:

1. Задача: найти вектор w такой, что $\hat{s} = w^T X$ имеет максимальный по модулю куртозис при $\|w\| = 1$
2. Функционал: $F(w) = |\mathbb{E}[(w^T X)^4] - 3|$
3. Оценка на конечной выборке: $F_n(w) = \left|\frac{1}{n}\sum_{t=1}^{n}(w^T X_t)^4 - 3\right|$

Алгоритм градиентного спуска:
- Если $\mathrm{Kurt}(w^T X) > 0$: $\Delta w \propto \mathbb{E}[X(w^T X)^3]$
- Если $\mathrm{Kurt}(w^T X) < 0$: $\Delta w \propto -\mathbb{E}[X(w^T X)^3]$

FastICA алгоритм (версия на основе куртозиса):

1. Случайная инициализация w с последующей нормализацией: $w \leftarrow w/\|w\|$
2. Итерации:
   - $w \leftarrow \mathbb{E}[X(w^T X)^3] - 3w$
   - $w \leftarrow w/\|w\|$
3. Повторение итераций до сходимости

Для нахождения нескольких независимых компонент применяется процедура декорреляции (ортогонализации Грама-Шмидта) после каждой итерации.

4.9 Independent component analysis based on Negentropy

ICA на основе негэнтропии

Негэнтропия как мера негауссовости:

$$J(\xi) = H(\xi_{\text{Gauss}}) - H(\xi)$$

где $H(\xi) = -\int p(y) \log p(y) dy$ - дифференциальная энтропия, ξ_{Gauss} - гауссова случайная величина с тем же математическим ожиданием и дисперсией, что и ξ.

Свойства негэнтропии:

- $J(\xi) \geq 0$
- $J(\xi) = 0$ тогда и только тогда, когда ξ имеет гауссово распределение
- Инвариантна к линейным преобразованиям

Аппроксимации негэнтропии:

1. На основе куртозиса: $J^*(\xi) = \frac{1}{12}\mathbb{E}[\xi^3]^2 + \frac{1}{48}\text{Kurt}(\xi)^2$
   - Для симметричных распределений с $\mathbb{E}[\xi^3] = 0$: $J^*(\xi) = \frac{1}{48}\text{Kurt}(\xi)^2$
2. Аппроксимация через нелинейные функции:

$$J^{**}(\xi) = \sum_{i=1}^{m} k_i [\mathbb{E}[h_i(\xi)] - \mathbb{E}[h_i(\xi_{\text{Gauss}})]^2$$

   Часто используемые функции:
   - $h_1(y) = \frac{1}{\alpha_1} \log \cosh(\alpha_1 y)$, $1 \leq \alpha_1 \leq 2$
   - $h_2(y) = -\exp(-y^2/2)$

Алгоритм FastICA на основе негэнтропии:

1. Для функции $h(y) = \log \cosh(y)$ с производной $g(y) = \tanh(y)$: $\Delta w \propto \mathbb{E}[X g(w^T X)] - \mathbb{E}[g'(w^T X)]w$
2. Итерации FastICA:
   - $w \leftarrow \mathbb{E}[X g(w^T X)] - \mathbb{E}[g'(w^T X)]w$
   - $w \leftarrow w/\|w\|$
3. Критерий сходимости: малое изменение направления w между итерациями

Сравнение с подходом на основе куртозиса:

- Более устойчив к выбросам
- Лучше работает с распределениями разных типов
- Требует выбора подходящих нелинейных функций

4.10 Independent component analysis based on Likelihood maximization and Mutual Information Minimization

ICA на основе максимизации правдоподобия и минимизации взаимной информации

4.10.1 Максимизация правдоподобия

Плотность вероятности наблюдаемых данных при заданной разделяющей матрице $B = A^{-1}$:

$$p_X(x) = |\det(B)| \cdot \prod_{j=1}^{q} p_j(b_j^T x)$$

где p_j - плотность вероятности j-й независимой компоненты.

Логарифмическая функция правдоподобия для выборки $\{X_1, X_2, \ldots, X_n\}$:

$$L(B, p_S) = \sum_{t=1}^{n} \sum_{i=1}^{q} \log p_i(b_i^T X_t) + n \log |\det(B)|$$

Максимизация $L(B, p_S)$ по B (при известных или аппроксимированных p_i):

$$\frac{\partial L}{\partial B} = \sum_{t=1}^{n} g(BX_t) \cdot X_t^T + n[B^T]^{-1}$$

где $g(s) = [g_1(s_1), g_2(s_2), \ldots, g_q(s_q)]^T$, $g_i(s_i) = \frac{d}{ds_i} \log p_i(s_i)$

Итерации градиентного спуска:

$$B \leftarrow B + \eta \left(\frac{1}{n} \sum_{t=1}^{n} g(BX_t) \cdot X_t^T + [B^T]^{-1} \right)$$

FastICA для максимизации правдоподобия приводит к тем же обновлениям, что и в случае максимизации негэнтропии.

4.10.2 Минимизация взаимной информации

Взаимная информация между компонентами $s_1, s_2, \ldots, s_q$:

$$I(s_1, s_2, \ldots, s_q) = \sum_{i=1}^{q} H(s_i) - H(s_1, s_2, \ldots, s_q)$$

Свойства:

- $I(s_1, s_2, \ldots, s_q) \geq 0$
- $I(s_1, s_2, \ldots, s_q) = 0$ тогда и только тогда, когда $s_1, s_2, \ldots, s_q$ статистически независимы

Для преобразования $S = BX$:

$$I(s_1, s_2, \ldots, s_q) = \sum_{i=1}^{q} H(s_i) - H(X) - \log |\det(B)|$$

Для ортогональной матрицы B после отбеливания:

$$I(s_1, s_2, \ldots, s_q) = \sum_{i=1}^{q} H(s_i) - H(X)$$

Минимизация взаимной информации эквивалентна минимизации суммы энтропий отдельных компонент, что в свою очередь эквивалентно максимизации суммы негэнтропий компонент.

Эквивалентность подходов:

- Максимизация правдоподобия
- Минимизация взаимной информации (при правильном выборе функций)
- Максимизация негауссовости через негэнтропию

Это показывает, что различные формулировки ICA приводят к одним и тем же алгоритмам, хотя и с разных теоретических позиций.

4.11 Projection pursuit: an approach and examples

Projection Pursuit: подход и примеры

Projection Pursuit (преследование проекций) - методы, которые ищут интересные низкоразмерные линейные проекции многомерных данных путем оптимизации определенной целевой функции, называемой индексом проекции.

4.11.1 Общий подход

1. Выбор индекса проекции $I(b)$, который измеряет "интересность" проекции данных на направление b, $\|b\| = 1$
2. Поиск направления b^*, максимизирующего выбранный индекс: $b^* = \arg\max_b I(b)$
3. Для многомерных проекций: последовательный или одновременный поиск нескольких направлений

4.11.2 Примеры индексов проекции

1. **Дисперсия** (приводит к PCA):

$$I_{PCA}(b) = \mathrm{Var}(b^T X) = b^T \Sigma b$$

2. **Куртозис** (приводит к одному из вариантов ICA):

$$I_{ICA-1}(b) = |\mathrm{Kurt}(b^T X)| = |E[(b^T X)^4] - 3(E[(b^T X)^2])^2|$$

3. **Негэнтропия** (другой вариант ICA):

$$I_{ICA-2}(b) = J(b^T X) \approx [E[G(b^T X)] - E[G(\xi_{Gauss})]]^2$$

где G - некоторая нелинейная функция, ξ_{Gauss} - стандартная гауссова случайная величина.

4. **Энтропия** (минимизация вместо максимизации):

$$I_H(b) = -H(b^T X)$$

5. **Индекс Фридмана-Таки** (оригинальная мера "дырчатости"):

$$I_{FT}(b) = \sum_{i=1}^{n} \sum_{j=1}^{n} [f(b^T(X_i - X_j)) - f_0]^2$$

где f - гауссово ядро, f_0 - константа нормализации.

6. **Индекс для дискриминантного анализа:**

$$I_{LDA}(b) = \frac{b^T \Sigma_{between} b}{b^T \Sigma_{within} b}$$

4.11.3 Примеры применения

1. **Кластерный анализ**: поиск проекций, выявляющих кластерную структуру данных
   - Проекции, максимизирующие индексы на основе смеси распределений
   - Визуализация мультимодальных данных
2. **Выявление выбросов**: направления, максимизирующие асимметрию или куртозис
3. **Исследование нелинейных структур**: направления, максимизирующие любые отклонения от нормальности
4. **Дискриминантный анализ**: направления, максимизирующие разделение классов
   - LDA как частный случай Projection Pursuit
5. **Классификация банкнот (швейцарский пример)**: обнаружение подделок через проекции, максимизирующие индекс "Hermit"
6. **Анализ многомерных данных**: поиск интересных зависимостей в сложных наборах данных
   - Например, выявление бивариантной смеси двух гауссианов в 10-мерном зашумленном пространстве

4.11.4 Оптимизация индексов проекции

1. Для некоторых индексов (PCA, LDA) - аналитическое решение через собственные значения
2. Для других индексов - численные методы оптимизации:
   - Градиентный спуск
   - Методы Ньютона
   - Методы поиска без производных

PCA и ICA можно рассматривать как частные случаи Projection Pursuit с конкретными индексами проекции.

5 Внутренняя размерность (Intrinsic Dimension)

5.1 Математические определения внутренней размерности

Внутренняя размерность характеризует минимальное число параметров, необходимых для описания данных без потери информации. Существует несколько формальных математических определений:

5.1.1 Размерность Хаусдорфа

Для ограниченного множества $C \subset \mathbb{R}^p$ и $S(X, r)$ - r-шара с центром в X:

Пусть $S(r) = \{S(X_i, r_i)\}$ - набор шаров такой, что:

- $r_i \leq r$
- $\cup_i S(X_i, r_i) \supset C$

Определяем величину:

$$H_r^d = \inf_S \sum_i r_i^d \tag{1}$$

где инфимум берется по всем возможным покрытиям S.

Размерность Хаусдорфа D_H определяется как критическое значение:

$$H^d = \lim_{r\to 0} H_r^d = \begin{cases} \infty, & \text{если } d < D_H \\ 0, & \text{если } d > D_H \\ \text{константа} \neq 0, & \text{если } d = D_H \end{cases} \tag{2}$$

Пример: Для отрезка $C = [0, 1]$ при $r = 1/(2n)$:

$$H_r^d = \sum_{i=1}^{n} \left(\frac{1}{2n}\right)^d = \frac{1}{2^d} \cdot n^{1-d} \tag{3}$$

Откуда следует $D_H = 1$.

5.1.2 Колмогоровская ёмкость (Box-counting dimension)

Для множества $C \subset \mathbb{R}^p$ обозначим $N(C, r)$ - минимальное число шаров радиуса r, необходимых для покрытия C.

Ёмкостная размерность определяется как:

$$D_{Cap} = \lim_{r\to 0} \frac{\ln N(C, r)}{\ln \frac{1}{r}} \tag{4}$$

Для отрезка $C = [0, 1]$ при $r = 1/(2n)$: $N(C, r) = n$, откуда $D_{Cap} = 1$.

5.1.3 Размерность Хаусдорфа-Безиковича

Пусть $E_r = \{[kr, kr + r)^p, k = 0, \pm 1, \pm 2, \ldots\}$ - множество кубов в $\mathbb{R}^p$ с ребром r.

Обозначим $N(C, r) = \#\{Q \in E_r : C \cap Q \neq \emptyset\}$ - число кубов, пересекающихся с C.

Если существуют числа V и D такие, что:

$$\lim_{r\to 0} \frac{N(C, r)}{V \cdot r^{-D}} = 1 \tag{5}$$

то $D_{HB}(C) = D$ называется **размерностью Хаусдорфа-Безиковича**, а $V_{HB}(C) = V$ - соответствующим объемом.

Если C измеримо по Жордану, то $D_{HB}(C)$ и $V_{HB}(C)$ равны топологической размерности и объему соответственно.

5.1.4 Корреляционная размерность

Пусть $\{X_1, X_2, \ldots, X_n\}$ - выборка из C, а $m(n) = n(n-1)/2$ - число пар.

Корреляционный интеграл:

$$C_n(r) = \frac{1}{m(n)} \sum_{1\leq i<j\leq n} I(||X_i - X_j|| \leq r) \tag{6}$$

где $I(A)$ - индикатор события A.

Корреляционная размерность определяется как:

$$D_{corr} = \lim_{r\to 0} \frac{\ln C_n(r)}{\ln r} \tag{7}$$

при $n \to \infty$.

5.1.5 Взаимосвязь размерностей

Для "хороших"множеств эти размерности совпадают, но в общем случае:

$$D_{corr} \leq D_{HB} \leq D_{Cap} \leq D_H \tag{8}$$

Эти определения имеют теоретическое значение, но напрямую их сложно использовать для оценки размерности по конечной выборке данных.

5.2 Оценка внутренней размерности: геометрические методы

Геометрические методы оценки внутренней размерности основаны на анализе пространственных отношений между точками выборки.

5.2.1 Методы на основе адаптации математических определений

Метод оценки корреляционной размерности Для конечной выборки $X_1, X_2, \ldots, X_n$ вычисляется корреляционный интеграл для различных значений r:

$$C_n(r) = \frac{1}{m(n)} \sum_{1 \leq i < j \leq n} I(||X_i - X_j|| \leq r) \tag{9}$$

График зависимости $\ln C_n(r)$ от $\ln r$ аппроксимируется прямой линией, наклон которой дает оценку размерности:

$$D_{corr} \approx \frac{\ln C_n(r_2) - \ln C_n(r_1)}{\ln r_2 - \ln r_1} \tag{10}$$

Практическая реализация:

1. Вычислить $C_n(r)$ для нескольких значений r
2. Построить график $\ln C_n(r)$ от $\ln r$
3. Найти линейный участок и определить его наклон

Box-counting метод

1. Разбить пространство сеткой с размером ячейки r
2. Подсчитать число ячеек $N(r)$, содержащих хотя бы одну точку выборки
3. Построить график зависимости $\ln N(r)$ от $\ln(1/r)$
4. Определить наклон линейного участка как оценку размерности:

$$D_{Cap} \approx \frac{\ln N(r_2) - \ln N(r_1)}{\ln(1/r_2) - \ln(1/r_1)} \tag{11}$$

5.2.2 Метод PCA и его локальные вариации

Глобальный PCA

1. Вычислить ковариационную матрицу $\Sigma = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T$
2. Найти собственные значения $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$
3. Оценить размерность как число значимых собственных значений:

$$D_{PCA} = \min \left\{ q : \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} > \text{threshold} \right\} \tag{12}$$

где threshold обычно выбирается как 0.9 или 0.95.

Локальный PCA

1. Для каждой точки X_i найти ее окрестность $U(X_i)$ (k ближайших соседей или ϵ-окрестность)
2. Применить PCA к каждой окрестности $U(X_i)$, получив локальные собственные значения $\{\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ip}\}$
3. Усреднить собственные значения по всем точкам: $\lambda_t = \frac{1}{n}\sum_{i=1}^{n} \lambda_{it}$
4. Оценить размерность аналогично глобальному PCA:

$$D_{LPCA} = \min\left\{q : \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k} > \text{threshold}\right\} \tag{13}$$

5.2.3 Метод на основе углов между векторами

1. Для каждой точки X_i найти ее k ближайших соседей $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$
2. Построить подпространство $L(X_i, k) = \text{Span}(X_{(1)} - X_i, X_{(2)} - X_i, \ldots, X_{(k)} - X_i)$
3. Измерить угол $\alpha_i^{(k)}$ между вектором $X_{(k+1)} - X_i$ к $(k+1)$-му соседу и его проекцией на $L(X_i, k)$
4. Увеличивать k пока средний угол $\bar{\alpha}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \alpha_i^{(k)}$ не превысит выбранный порог
5. Оценка размерности: $D = \min\{k : \bar{\alpha}^{(k)} < \text{threshold}\}$

5.2.4 Метод на основе собственных значений оператора Лапласа

Для графа соседства с весами W_{ij}:

1. Построить матрицу Лапласа $L = D - W$, где D - диагональная матрица степеней
2. Найти собственные значения $0 = \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$
3. Применить асимптотическую формулу: $\log(\mu_j) \approx A + \frac{2}{q}\log(j)$, где q - внутренняя размерность
4. Оценить размерность методом регрессии из соотношения:

$$D_{Lap} \approx \frac{2}{\text{наклон графика } \log(\mu_j) \text{ от } \log(j)} \tag{14}$$

5.2.5 Преимущества и недостатки геометрических методов

Преимущества:
- Интуитивная геометрическая интерпретация
- Некоторые методы не требуют явных предположений о распределении данных
- Возможность анализировать локальную размерность в разных частях данных

Недостатки:
- Чувствительность к выбору параметров (размеру окрестности, порогу)
- Проблемы с масштабируемостью на больших наборах данных
- Некоторые методы чувствительны к шуму и выбросам

5.3 Оценка внутренней размерности: регрессионный подход

Регрессионные методы оценки внутренней размерности основаны на моделировании отношений между расстояниями до ближайших соседей и их количеством или рангом.

5.3.1 Теоретическое обоснование

Пусть X - многообразие с внутренней размерностью d, вложенное в $\mathbb{R}^p$, с плотностью вероятности $f(X)$.

Для малого шара $B(X, \varepsilon)$ с центром в точке $X \in X$ и радиусом ε:

Вероятность, что случайная точка X' попадет в этот шар:

$$P(X, \varepsilon) = P(X' \in B(X, \varepsilon) \cap X) \approx f(X) \cdot \varepsilon^d \cdot V(d) \tag{15}$$

где $V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ - объем d-мерного единичного шара.

5.3.2 Метод регрессии на k ближайших соседях

1. Для каждой точки X_i из выборки $\{X_1, X_2, \ldots, X_n\}$:
   - Найти k ближайших соседей $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$
   - Вычислить расстояния $r_j(X_i) = \|X_{(j)} - X_i\|$ для $j = 1, 2, \ldots, k$
2. Для шара $B(X_i, r_k(X_i))$ точно k точек из n попадает внутрь, следовательно:
$$\frac{k}{n} \approx P(X_i, r_k(X_i)) \approx f(X_i) \cdot (r_k(X_i))^d \cdot V(d) \tag{16}$$
3. Логарифмируя обе части:
$$\ln k - \ln n \approx \ln f(X_i) + d \cdot \ln r_k(X_i) + \ln V(d) \tag{17}$$
4. Перегруппировав:
$$\ln k \approx A_i + d \cdot T_k(i) \tag{18}$$
где $A_i = A_i(d) = \ln n - \ln V(d) - \ln f(X_i)$ и $T_k(i) = -\ln r_k(X_i)$
5. Для каждой точки X_i и различных значений k получается система регрессионных уравнений:
$$\ln k \approx A_i + d \cdot T_k(i), \quad k = 1, 2, \ldots, k_{max} \tag{19}$$
6. Оценка размерности получается методом наименьших квадратов:
$$d = \frac{\sum_{i=1}^{n} \sum_{k=1}^{k_{max}} (T_k(i) - \bar{T}_i)(\ln k - \overline{\ln k})}{\sum_{i=1}^{n} \sum_{k=1}^{k_{max}} (T_k(i) - \bar{T}_i)^2} \tag{20}$$
где $\bar{T}_i = \frac{1}{k_{max}} \sum_{k=1}^{k_{max}} T_k(i)$ и $\overline{\ln k} = \frac{1}{k_{max}} \sum_{k=1}^{k_{max}} \ln k$.

5.3.3 Вариации регрессионного подхода

Метод Levina-Bickel (2005) Для каждой точки X_i строится локальная оценка:

$$\hat{d}_i = \left(\frac{1}{k} \sum_{j=1}^{k} \ln \frac{r_k(X_i)}{r_j(X_i)} \right)^{-1} \tag{21}$$

Глобальная оценка получается усреднением:

$$\hat{d} = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_i \tag{22}$$

Метод построения кривой масштабирования

1. Вычислить среднее расстояние до k-ого соседа:

$$\bar{r}_k = \frac{1}{n}\sum_{i=1}^{n} r_k(X_i) \tag{23}$$

2. Построить зависимость $\ln \bar{r}_k$ от $\ln k$

3. Наклон этой зависимости равен $1/d$:

$$\hat{d} = \frac{\ln k_2 - \ln k_1}{\ln \bar{r}_{k_2} - \ln \bar{r}_{k_1}} \tag{24}$$

5.3.4 Практические соображения

1. **Выбор значения** k:
   - При малых k оценки нестабильны из-за локальных флуктуаций
   - При больших k могут нарушаться предположения о локальной линейности многообразия
   - Рекомендуется анализировать устойчивость оценки при различных k

2. **Робастность**:
   - Для повышения устойчивости к выбросам можно использовать медианное усреднение вместо среднего
   - Возможно применение робастных методов регрессии

3. **Коррекция смещения**:
   - При малом размере выборки методы могут давать смещенные оценки
   - Предложены различные методы коррекции смещения, например, путем введения поправочных коэффициентов

5.3.5 Преимущества и недостатки регрессионного подхода

Преимущества:
- Статистически обоснованные оценки
- Возможность получения локальных оценок размерности
- Относительная простота реализации

Недостатки:
- Чувствительность к выбору параметров (особенно k)
- Предположение о равномерном распределении точек в локальных окрестностях
- Проблемы со сходимостью при наличии шума или неравномерной выборки

5.4 Оценка внутренней размерности: подход максимального правдоподобия

Метод максимального правдоподобия (MLE) для оценки внутренней размерности основан на моделировании локального распределения расстояний между точками данных.

5.4.1 Теоретическое обоснование

Предположения:

1. Данные лежат на или вблизи многообразия размерности d в $\mathbb{R}^p$
2. Локально многообразие можно аппроксимировать евклидовым пространством $\mathbb{R}^d$
3. Точки выборки распределены локально равномерно с плотностью $f(X)$

5.4.2 Метод Levina-Bickel (2005)

Вероятностная модель В окрестности точки X вероятность обнаружить точку на расстоянии от t до $t + dt$ моделируется пуассоновским процессом с интенсивностью:

$$\lambda(t) = f(X) \cdot V(d) \cdot d \cdot t^{d-1} \tag{25}$$

где:

- $f(X)$ - плотность в точке X
- $V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ - объем d-мерного единичного шара
- $d \cdot t^{d-1}$ - производная от t^d

Функция правдоподобия Для точки X и ее соседей на расстояниях $r_1 < r_2 < \ldots < r_k$ функция правдоподобия:

$$F(f(X), d) = \exp\left(\int_0^{r_k} \ln \lambda(t) dN(X,t) - \int_0^{r_k} \lambda(t) dt\right) \tag{26}$$

где $N(X,t)$ - число точек в шаре радиуса t с центром в X.
При дифференцировании по t, интеграл преобразуется в сумму:

$$\ln F(f(X), d) = \sum_{j=1}^{k-1} \ln \lambda(r_j) - \int_0^{r_k} \lambda(t) dt \tag{27}$$

Максимизация правдоподобия Оценка максимального правдоподобия для размерности d в точке X при $R = r_k$:

$$D_R(X) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \ln \frac{R}{r_j}\right)^{-1} \tag{28}$$

Используя k ближайших соседей и обозначая $R = r_k(X)$:

$$D_k(X) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \ln \frac{r_k(X)}{r_j(X)}\right)^{-1} \tag{29}$$

Глобальная оценка получается усреднением по всем точкам:

$$D_k = \frac{1}{n} \sum_{i=1}^{n} D_k(X_i) \tag{30}$$

5.4.3 Расширения метода MLE

MLE с переменным k (Levina-Bickel, 2005) Для улучшения устойчивости оценки:

$$D_{k_1,k_2} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_2-k_1+1}\sum_{k=k_1}^{k_2}D_k(X_i) \tag{31}$$

MLE с коррекцией смещения (Mackay-Ghahramani, 2005) Учитывает систематическое смещение при малых k:

$$D_k^{corrected} = \frac{k-2}{k-1}\cdot D_k \tag{32}$$

Геодезический MLE (Chen et al., 2013) Заменяет евклидовы расстояния на геодезические для улучшения оценки на искривленных многообразиях:

$$D_k^{geo}(X) = \left(\frac{1}{k-1}\sum_{j=1}^{k-1}\ln\frac{d_{geo}(X,X_{(k)})}{d_{geo}(X,X_{(j)})}\right)^{-1} \tag{33}$$

где $d_{geo}(X,Y)$ - аппроксимация геодезического расстояния (например, через графовые расстояния).

5.4.4 Оценка плотности

MLE также позволяет оценить локальную плотность $f(X)$:

$$\hat{f}_R(X) = \frac{k-1}{V(D_R(X))\cdot R^{D_R(X)}} \tag{34}$$

или при использовании k ближайших соседей:

$$\hat{f}_k(X) = \frac{k-1}{V(D_k(X))\cdot (r_k(X))^{D_k(X)}} \tag{35}$$

Это позволяет оценить энтропию распределения:

$$\hat{J}(f) = \frac{1}{n}\sum_{i=1}^{n}\ln(\hat{f}_k(X_i)) \tag{36}$$

5.4.5 Выбор оптимального k

Ключевой вопрос - как выбрать оптимальное значение k. Предложены различные подходы:

1. **Метод стабилизации**: выбрать k, при котором оценка D_k стабилизируется
2. **Метод минимальной вариации**: минимизировать дисперсию локальных оценок $D_k(X_i)$
3. **Кросс-валидация**: оптимизировать прогностическую способность модели
4. **Байесовский подход**: усреднить оценки с различными k с соответствующими весами

5.4.6 Преимущества и недостатки подхода MLE

Преимущества:
- Статистически обоснованный метод с теоретическими гарантиями
- Возможность получения как локальных, так и глобальных оценок
- Одновременная оценка размерности и плотности
- Высокая эффективность на данных, близких к предположениям модели

Недостатки:
- Предположение о локально равномерном распределении данных
- Чувствительность к выбору числа соседей k
- Вычислительные затраты на поиск k ближайших соседей
- Потенциальные проблемы на границах многообразия

5.4.7 Практические рекомендации:

1. Использовать несколько значений k и анализировать стабильность оценки
2. Сравнивать результаты с другими методами оценки размерности
3. Визуализировать локальные оценки для выявления областей с различной локальной размерностью
4. При наличии шума, предварительно применять методы его подавления

MLE подход к оценке внутренней размерности является одним из наиболее теоретически обоснованных и широко используемых на практике.

6 Методы обучения многообразий (Manifold Learning)

6.1 Locally Linear Embedding (LLE)

Локально-линейное вложение (LLE) – метод нелинейного снижения размерности, сохраняющий локальную геометрию данных путем представления точек как линейных комбинаций их соседей.

6.1.1 Теоретическое обоснование

LLE основан на предположении, что локально (в малой окрестности) многообразие можно аппроксимировать линейным подпространством. Следовательно, каждая точка может быть представлена как линейная комбинация своих ближайших соседей.

6.1.2 Алгоритм LLE

1. **Определение соседства**:
   - Для каждой точки $X_i \in \mathbb{R}^p$ найти k ближайших соседей $\{X_1(i), X_2(i), \ldots, X_k(i)\}$
2. **Вычисление весовых коэффициентов**:
   - Для каждой точки X_i найти веса $\{w_1(i), w_2(i), \ldots, w_k(i)\}$, которые минимизируют:

$$E_i = \left\| X_i - \sum_{j=1}^{k} w_j(i) \times X_j(i) \right\|^2 \tag{37}$$

- При условии: $\sum_{j=1}^{k} w_j(i) = 1$ (для инвариантности к сдвигам)
- Решение находится аналитически через систему линейных уравнений:

$$G_{ts}(i) = (X_t(i) - X_i) \cdot (X_s(i) - X_i) \tag{38}$$

$$w_t(i) = \frac{\sum_{s=1}^{k} G_{ts}^{-1}(i)}{\sum_{t,s=1}^{k} G_{ts}^{-1}(i)} \tag{39}$$

3. **Построение низкоразмерного вложения**:
   - Найти q-мерные координаты $\{y_1, y_2, \ldots, y_n\} \subset \mathbb{R}^q$, минимизирующие:

$$\Phi(Y) = \sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{k} w_j(i) \times y_j(i) \right\|^2 \tag{40}$$

   - С ограничениями: $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$ и $\frac{1}{n}\sum_{i=1}^{n} y_i y_i^T = I_q$
   - В матричной форме: $L(Y) = \mathrm{Tr}(Y(I_n - W)^T(I_n - W)Y^T)$
4. **Решение задачи собственных значений**:
   - Минимизация $L(Y)$ эквивалентна нахождению q собственных векторов матрицы:

$$M_{LLE} = (I_n - W)^T(I_n - W) \tag{41}$$

   - соответствующих q наименьшим ненулевым собственным значениям
   - Собственный вектор с нулевым собственным значением (константный) отбрасывается

6.1.3 Преимущества и недостатки LLE

Преимущества:
- Сохраняет локальную структуру данных
- Требует настройки всего одного параметра (числа соседей k)
- Имеет аналитическое решение (задача собственных значений)
- Эффективно работает для данных, лежащих на гладком многообразии

Недостатки:
- Чувствителен к выбору числа соседей k
- Может давать топологические "сшивания"и "складки"при сложной геометрии данных
- Плохо работает с негладкими многообразиями или при наличии "дыр"
- Не всегда хорошо сохраняет глобальную структуру данных
- Проблемы с масштабируемостью для очень больших наборов данных

6.2 ISOmetric MAPping (ISOMAP)

ISOmetric MAPping (ISOMAP) – метод нелинейного снижения размерности, который сохраняет геодезические расстояния между точками многообразия.

6.2.1 Теоретическое обоснование

ISOMAP основан на предположении, что евклидовы расстояния в исходном пространстве не отражают истинную структуру нелинейного многообразия. Вместо них используются геодезические расстояния – длины кратчайших путей по многообразию.

6.2.2 Алгоритм ISOMAP

1. **Построение графа соседства**:
   - Для каждой точки X_i определить соседей (через k ближайших соседей или ϵ-окрестность)
   - Построить взвешенный неориентированный граф $\Gamma(X_n)$, где веса рёбер равны евклидовым расстояниям между соседними точками
2. **Вычисление геодезических расстояний**:
   - Используя алгоритм Дейкстры (или Флойда-Уоршалла), найти длины кратчайших путей $D(X_i, X_j)$ между всеми парами точек в графе
   - Эти расстояния аппроксимируют геодезические расстояния на многообразии
3. **Применение многомерного шкалирования (MDS)**:
   - Построить матрицу квадратов расстояний $\mathbf{D}_X$ с элементами $[D(X_i, X_j)]^2$
   - Выполнить операцию двойного центрирования: $\mathbf{S}_X = -\frac{1}{2}H\mathbf{D}_X H$, где $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ – центрирующая матрица
   - Найти q собственных векторов $v_1, v_2, \ldots, v_q$ матрицы $\mathbf{S}_X$, соответствующих q наибольшим собственным значениям $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$
   - q-мерные координаты: $Y = \Lambda^{1/2}V^T$, где $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_q)$ и $V = [v_1, v_2, \ldots, v_q]$

6.2.3 Теоретические гарантии

Если многообразие изометрично евклидову пространству, то при достаточно плотной выборке и правильном выборе параметров ISOMAP гарантированно восстанавливает истинную геометрию многообразия с точностью до изометрии.

6.2.4 Преимущества и недостатки ISOMAP

Преимущества:
- Сохраняет глобальную геометрическую структуру данных
- Теоретически обоснован для изометрично вложенных многообразий
- Интуитивно понятный подход через геодезические расстояния
- Хорошо работает для данных, лежащих на "развертываемых"многообразиях

Недостатки:
- Чувствителен к шуму и выбросам
- Требует точной аппроксимации геодезических расстояний
- Проблемы с "дырами"в данных и негладкими многообразиями
- Вычислительно затратен для больших наборов данных ($\mathrm{O(n^3)}$ для MDS)
- Ограничен многообразиями, которые можно изометрически отобразить в евклидово пространство

6.3 Laplacian Eigenmaps

Laplacian Eigenmaps (собственные карты Лапласа) – метод нелинейного снижения размерности, основанный на теории спектральных графов и дифференциальной геометрии.

6.3.1 Теоретическое обоснование

Метод основан на связи между лапласианом графа и оператором Лапласа-Бельтрами на многообразии. Основная идея заключается в том, что собственные функции оператора Лапласа-Бельтрами образуют оптимальное вложение многообразия, сохраняющее его локальную структуру.

6.3.2 Алгоритм Laplacian Eigenmaps

1. **Построение графа соседства**:
   - Для каждой точки X_i определить соседей (через k ближайших соседей или ϵ-окрестность)
   - Построить неориентированный граф $\Gamma(X_n)$, где рёбра соединяют соседние точки
2. **Взвешивание рёбер** (опционально):
   - Определить веса рёбер, например, используя тепловое ядро:

$$W_{ij} = \begin{cases} \exp\left(-\frac{\|X_i - X_j\|^2}{t}\right), & \text{если } X_i \text{ и } X_j \text{ соседи} \\ 0, & \text{иначе} \end{cases} \tag{42}$$

   - Где $t > 0$ – параметр ширины ядра
3. **Построение лапласиана графа**:
   - Вычислить матрицу степеней D – диагональную матрицу с элементами $D_{ii} = \sum_j W_{ij}$
   - Построить лапласиан $L = D - W$
4. **Спектральное разложение**:
   - Решить обобщённую задачу собственных значений:

$$Lv = \lambda Dv \tag{43}$$

   - Отбросить собственный вектор v_0, соответствующий нулевому собственному значению $\lambda_0 = 0$
   - Использовать q собственных векторов $v_1, v_2, \ldots, v_q$, соответствующих q наименьшим ненулевым собственным значениям $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_q$, для формирования q-мерного вложения:

$$y_i = (v_{1i}, v_{2i}, \ldots, v_{qi})^T, \quad i = 1, 2, \ldots, n \tag{44}$$

6.3.3 Связь с оператором Лапласа-Бельтрами

Дискретный лапласиан L аппроксимирует оператора Лапласа-Бельтрами Δ_M на многообразии M:

$$\Delta_M f(x) = \operatorname{div}(\nabla_M f(x)) = -\sum_{k=1}^{p} \frac{\partial^2 f(x)}{\partial x_k^2} \tag{45}$$

Минимизация функционала $\Phi(f) = \int_M \|\nabla_M f(x)\|^2 d\mu(x)$ приводит к собственным функциям оператора Δ_M, что в дискретном случае соответствует минимизации:

$$\Phi(y) = \sum_{i,j=1}^{n} W_{ij} \|y_i - y_j\|^2 \tag{46}$$

6.3.4 Преимущества и недостатки Laplacian Eigenmaps

Преимущества:
- Сильное теоретическое обоснование через спектральную теорию графов и дифференциальную геометрию
- Хорошо сохраняет локальную структуру данных
- Устойчив к шуму и выбросам (при правильном выборе весов)
- Вычислительно эффективнее, чем некоторые другие методы (особенно при разреженной матрице весов)
- Естественно обобщается на полууправляемое обучение

Недостатки:
- Чувствительность к выбору параметров (числа соседей, параметра теплового ядра)
- Не всегда хорошо сохраняет глобальную структуру данных
- Проблемы с "дырами"в данных и разрывными многообразиями
- Отсутствие явного отображения для новых точек (out-of-sample extension)
- Возможные вычислительные проблемы при работе с очень большими наборами данных

6.4 Log Map

Log map (логарифмическое отображение) – метод снижения размерности, основанный на римановой геометрии многообразий, который использует экспоненциальное и логарифмическое отображения для параметризации многообразия.

6.4.1 Теоретические основы

1. **Экспоненциальное отображение (Exponential map)**:
   - Для точки $p \in M$ на многообразии и вектора $v \in T_pM$ в касательном пространстве, экспоненциальное отображение $\exp_p(v) = X \in M$ определяет точку на многообразии, достигаемую движением по геодезической линии в направлении v на расстояние $\|v\|$
2. **Логарифмическое отображение (Log map)**:
   - Обратное к экспоненциальному: $\log_p(X) = v \in T_pM$ – вектор в касательном пространстве, такой что $\exp_p(v) = X$
   - Геометрически $\log_p(X)$ указывает направление и расстояние для движения от p к X по геодезической линии
3. **Римановы нормальные координаты (RNC)**:
   - Параметризация окрестности точки p на многообразии через координаты в касательном пространстве T_pM
   - Для точки X вблизи p: $Y(X) = \log_p(X) \in \mathbb{R}^q$ – координаты X в системе RNC с центром в p

6.4.2 Алгоритм Log map / Рианмановы нормальные координаты (RNC)

1. **Выбор базовой точки** p:
   - Выбрать точку $p \in M$ (например, среднее значение или точку с минимальным геодезическим радиусом)

2. **Оценка геодезических расстояний**:
   - Построить граф соседства и вычислить приближенные геодезические расстояния $D(p, X_i)$ от базовой точки до каждой точки X_i (как в ISOMAP)

3. **Оценка направлений в касательном пространстве**:
   - Для каждой точки X_i определить направление $e(X_i) \in S^q(p)$ (единичную сферу в касательном пространстве)
   - Использовать локальную аппроксимацию: $f(Z, X) = (d_M(Z, X))^2$ – квадрат геодезического расстояния
   - Построить полином второго порядка $\phi(Z, X)$, аппроксимирующий $f(Z, X)$ для $Z \in U$ (окрестность p)
   - Направление $e(X) = g(X)/\|g(X)\|$, где $g(X) = \nabla_Z \phi(Z, X)|_{Z=p}$ – градиент в точке p

4. **Построение отображения**:
   - Для каждой точки X_i вычислить RNC: $y(X_i) = d_M(p, X_i) \cdot e(X_i) \in \mathbb{R}^q$
   - $y(X_i)$ – q-мерное представление точки X_i

6.4.3 Варианты и улучшения

1. **LOGMAP (Brun et al., 2005)**:
   - Фокусируется на эффективном вычислении направлений в касательном пространстве
   - Использует полиномиальную аппроксимацию геодезических расстояний

2. **Riemannian Manifold Learning (RML) (Lin et al., 2006)**:
   - Более точное оценивание касательного пространства
   - Улучшенная процедура для точек, далеких от базовой

3. **Improved RML (Lin, Zha, 2008)**:
   - Использует локальные квадратичные аппроксимации геодезических линий
   - Лучше работает с сильно искривленными многообразиями

6.4.4 Преимущества и недостатки

Преимущества:
- Строгое математическое обоснование в рамках римановой геометрии
- Сохраняет как локальную, так и глобальную структуру многообразия
- Обеспечивает естественную параметризацию многообразия
- Хорошо работает для многообразий с высокой кривизной

Недостатки:
- Чувствительность к выбору базовой точки
- Проблемы с точками вблизи разреза (cut locus)
- Вычислительная сложность определения направлений
- Ограниченная область применимости из-за необходимости аппроксимировать градиенты
- Сложность в интерпретации и реализации

6.5 Riemannian Manifold Learning

Riemannian Manifold Learning (RML) – это расширение и уточнение метода Log map, которое более тщательно учитывает риманову геометрию многообразий.

6.5.1 Теоретические основы

RML опирается на те же концепции римановой геометрии, что и Log map:
- Экспоненциальное и логарифмическое отображения
- Римановы нормальные координаты
- Геодезические линии как кратчайшие пути на многообразии

Однако RML предлагает более точные методы аппроксимации этих объектов.

6.5.2 Алгоритм RML (Lin, Zha, Lee, 2006)

1. **Оценка касательного пространства**:
   - Выбрать базовую точку p
   - Найти $q+1$ точек $X_0, X_1, \ldots, X_q$ из окрестности базовой точки p, лежащих в общем положении
   - Построить линейное подпространство $L(p) = X_0 + \mathrm{Span}(X_1 - X_0, X_2 - X_0, \ldots, X_q - X_0)$ – оценка касательного пространства T_pM
   - Построить ортонормированный базис $\{e_1, e_2, \ldots, e_q\}$ в $L(p)$

2. **Вычисление римановых нормальных координат для близких точек**:
   - Для точки X из окрестности U базовой точки p:
     - Найти проекцию $z = (z_1, z_2, \ldots, z_q)$ точки X на $L(p)$
     - Решить задачу минимизации: $y = \arg\min_y \|X - (p + \sum_{i=1}^{q} y_i \cdot e_i)\|^2$
     - Нормализовать: $y = \frac{\|z\|_q}{\|z\|_p} \cdot y$ – римановы нормальные координаты точки X

3. **Вычисление римановых нормальных координат для далеких точек**:
   - Для точки X вне окрестности U:
     - Найти предыдущую точку b на кратчайшем пути от p к X
     - Определить k промежуточных точек $c_1, c_2, \ldots, c_k$ с известными координатами
     - Вычислить углы между векторами в исходном пространстве: $\alpha_i = \angle((X-b), (c_i - b))$
     - Найти координаты $y(X)$, сохраняющие эти углы в пространстве меньшей размерности:

$$y(X) = y(b) + \|X - b\| \cdot \frac{\sum_{i=1}^{k} \cos(\alpha_i) \cdot (y(c_i) - y(b))}{\|\sum_{i=1}^{k} \cos(\alpha_i) \cdot (y(c_i) - y(b))\|} \tag{47}$$

6.5.3 Улучшенный RML (Lin, Zha, 2008)

1. **Локальная квадратичная аппроксимация геодезических**:
   - Использует второй порядок разложения в ряд Тейлора для более точной аппроксимации геодезических линий
   - Учитывает кривизну многообразия через метрический тензор и символы Кристоффеля

2. **Адаптивный выбор окрестностей**:

- Размер окрестности зависит от локальной кривизны многообразия
- Меньшие окрестности в областях высокой кривизны

3. **Улучшенная процедура распространения координат**:
   - Более устойчивый алгоритм для точек, далёких от базовой
   - Учет нескольких возможных путей для повышения робастности

6.5.4 Преимущества и недостатки

Преимущества:
- Более точное отражение римановой геометрии многообразия
- Лучшая аппроксимация геодезических линий
- Работает с многообразиями высокой кривизны
- Сохраняет как локальные, так и глобальные свойства многообразия
- Теоретически обоснованный подход

Недостатки:
- Высокая вычислительная сложность
- Сложность реализации и настройки параметров
- Чувствительность к выбору базовой точки
- Проблемы с многообразиями сложной топологии
- Трудности с масштабированием на большие наборы данных

6.6 Вероятностные методы: SNE, t-SNE, UMAP

Эти методы основаны на вероятностном подходе к сохранению структуры соседства и относятся к современным алгоритмам визуализации и снижения размерности данных.

6.6.1 Stochastic Neighbor Embedding (SNE)

Основная идея: преобразовать расстояния между точками в условные вероятности, которые представляют сходство точек, и создать низкоразмерное вложение, сохраняющее эти вероятности.

Алгоритм SNE:

1. **Определение вероятностей в исходном пространстве**:
   - Для каждой точки X_i определить условные вероятности выбора других точек в качестве соседей:
   $$P_{j|i} = \frac{\exp(-\|X_i - X_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|X_i - X_k\|^2/2\sigma_i^2)} \tag{48}$$
   - Параметр σ_i выбирается так, чтобы перплексия распределения $P_{j|i}$ была равна заданному значению

2. **Определение вероятностей в пространстве меньшей размерности**:
   - Для точек $y_i, y_j \in \mathbb{R}^q$:
   $$Q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k\neq i}\exp(-\|y_i - y_k\|^2)} \tag{49}$$

3. **Минимизация расхождения Кульбака-Лейблера**:

- Минимизировать:

$$C_{SNE} = \sum_i KL(P_{\cdot|i}||Q_{\cdot|i}) = \sum_i \sum_j P_{j|i} \log \frac{P_{j|i}}{Q_{j|i}} \tag{50}$$

- Используя градиентный спуск:

$$\frac{\partial C_{SNE}}{\partial y_i} = 2 \sum_j (P_{j|i} - Q_{j|i} + P_{i|j} - Q_{i|j}) \cdot (y_i - y_j) \tag{51}$$

6.6.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Основные улучшения t-SNE по сравнению с SNE:
- Использование симметричных вероятностей: $P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}$
- Использование t-распределения с одной степенью свободы (распределение Коши) в пространстве меньшей размерности:

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \tag{52}$$

Преимущества t-SNE:
- Решает проблему "скученности"(crowding problem) благодаря "тяжелым хвостам"t-распределения
- Лучше сохраняет как локальную, так и глобальную структуру данных
- Более устойчивое и интерпретируемое вложение

Алгоритм t-SNE:

1. Вычислить матрицу сходства P в исходном пространстве
2. Инициализировать точки y_i случайным образом или через другой метод (например, PCA)
3. Минимизировать KL-дивергенцию между P и Q с помощью градиентного спуска с моментом

6.6.3 Uniform Manifold Approximation and Projection (UMAP)

Основная идея: использование принципов алгебраической топологии и теории категорий для создания топологического представления данных и последующего вложения в низкоразмерное пространство.

Теоретические основы UMAP:
- Представление данных как нечеткого топологического множества
- Аппроксимация метрики на многообразии через локально-однородное распределение
- Использование теории графов для дискретизации представления

Алгоритм UMAP:

1. **Построение взвешенного графа в исходном пространстве**:
   - Для каждой точки X_i найти k ближайших соседей

- Определить локальный масштаб ρ_i и σ_i такие, что:

$$\sum_{j=1}^{k} \exp\left(-\frac{\max(0, \|X_i - X_j\| - \rho_i)}{\sigma_i}\right) = \log_2(k) \tag{53}$$

- Построить направленный граф с весами:

$$v_{ij} = \exp\left(-\frac{\max(0, \|X_i - X_j\| - \rho_i)}{\sigma_i}\right) \tag{54}$$

- Преобразовать в неориентированный граф: $w_{ij} = v_{ij} + v_{ji} - v_{ij} \cdot v_{ji}$

2. **Построение графа в низкоразмерном пространстве**:
   - Определить функцию сходства для точек $y_i, y_j \in \mathbb{R}^q$:

$$q_{ij} = (1 + a\|y_i - y_j\|^{2b})^{-1} \tag{55}$$

   - где a и b – параметры, зависящие от размерности q

3. **Оптимизация вложения**:
   - Минимизировать кросс-энтропию:

$$C_{UMAP} = \sum_{i,j} w_{ij} \log \frac{w_{ij}}{q_{ij}} + (1 - w_{ij}) \log \frac{1 - w_{ij}}{1 - q_{ij}} \tag{56}$$

   - с помощью стохастического градиентного спуска и отрицательного сэмплирования

6.6.4 Сравнение методов

Аспект	**SNE**	**t-SNE**	**UMAP**
Теоретическое обоснование	Вероятностное	Вероятностное	Топологическое
Скорость работы	Медленная	Медленная	Быстрая
Сохранение локальной структуры	Хорошее	Отличное	Отличное
Сохранение глобальной структуры	Плохое	Среднее	Хорошее
Масштабируемость	Плохая	Средняя	Хорошая
Параметры	Перплексия	Перплексия	Число соседей, min_di
Год появления	2002	2008	2018

Таблица 2: Сравнительная характеристика методов SNE, t-SNE и UMAP

6.6.5 Преимущества и недостатки

SNE:
- **Преимущества**: Интуитивный вероятностный подход
- **Недостатки**: Проблема "скученности асимметричные вероятности, медленная сходимость

t-SNE:
- **Преимущества**: Решает проблему "скученности хорошо сохраняет кластеры, стандарт де-факто для визуализации
- **Недостатки**: Вычислительная сложность O(n²), плохо сохраняет глобальную структуру, чувствительность к перплексии

UMAP:

- **Преимущества**: Теоретически обоснован, быстрее t-SNE, лучше сохраняет глобальную структуру, масштабируемость
- **Недостатки**: Сложность алгоритма, чувствительность к параметрам, сложная интерпретация

6.7 Grassmann & Stiefel Eigenmaps

Grassmann&Stiefel Eigenmaps (GSE) – метод обучения многообразий, который сохраняет не только точки многообразия, но и касательные пространства к нему, решая задачу обучения касательного расслоения многообразия.

6.7.1 Теоретическое обоснование

Метод основан на концепции касательного расслоения многообразия – объединения всех касательных пространств в каждой точке многообразия.

Цель GSE – найти отображения $h : \mathbb{R}^p \to \mathbb{R}^q$ и $g : \mathbb{R}^q \to \mathbb{R}^p$, обеспечивающие:

1. **Близость многообразий**: $\hat{X} = g(h(X)) \approx X$ для всех $X \in \mathcal{M}$
2. **Близость касательных пространств**: $L_{h,g}(\hat{X}) = \mathrm{Span}(J_g(h(X))) \approx L(X)$

где $L(X)$ – касательное пространство к исходному многообразию в точке X, а $L_{h,g}(\hat{X})$ – касательное пространство к реконструированному многообразию в точке $\hat{X}$.

6.7.2 Алгоритм GSE

GSE состоит из трех последовательных этапов:

1. Аппроксимация касательных пространств (Tangent Manifold Learning)

1. **Оценка касательных пространств**:
   - Для каждой точки X_i применить локальный PCA к её окрестности $U(X_i)$
   - Получить ортогональную матрицу $Q_{PCA}(X_i)$ с q столбцами, аппроксимирующими базис касательного пространства $L(X_i)$
2. **Согласование базисов** (Aligning Problem):
   - Найти вращения $v(X_i)$ для каждого базиса, чтобы минимизировать:

$$\Phi_n = \frac{1}{2} \sum_{i,j=1}^{n} K(X_i, X_j) \| Q_{PCA}(X_i) \times v(X_i) - Q_{PCA}(X_j) \times v(X_j) \|_F^2 \tag{57}$$

   - где $K(X_i, X_j)$ – ядро, отражающее близость точек и их касательных пространств
   - Решается как обобщенная задача собственных значений
3. **Построение согласованных базисов**:
   - $H(X_i) = Q_{PCA}(X_i) \times v(X_i)$ – согласованная аппроксимация базиса касательного пространства в точке X_i

2. Вложение многообразия (Manifold Embedding)

1. **Определение весов**:
   - Минимизировать:

$$\sum_{i,j=1}^{n} K(X_i, X_j)\|(X_j - X_i) - H(X_i) \times (h(X_j) - h(X_i))\|^2 \tag{58}$$

   - где $h(X_i)$ – искомые низкоразмерные координаты
2. **Аналитическое решение**:
   - Замкнутая форма для $h(X_i)$ через решение линейной системы
3. **Обобщение на новые точки** (Out-of-sample extension):
   - Для новой точки X:

$$h(X) = \frac{1}{K(X)} \sum_{i=1}^{n} K(X, X_i) \left[h(X_i) + \frac{1}{K(X)} \sum_{j=1}^{n} K(X, X_j) v^{-1}(X_j)(Q_{PCA}(X_j))^T \times (X - X_j) \right] \tag{59}$$

   где $K(X) = \sum_{j=1}^{n} K(X, X_j)$

3. Реконструкция многообразия (Manifold Reconstruction)

1. **Построение ядра** $k(y, y_j)$ в пространстве $\mathbb{R}^q$ для аппроксимации $K(X, X_j)$
2. **Восстановление оригинальных координат**:
   - Для точки $y \in \mathbb{R}^q$:

$$g(y) = \frac{1}{k(y)} \sum_{j=1}^{n} k(y, y_j) \left[X_j + \frac{1}{k(y)} \sum_{j=1}^{n} k(y, y_j) \times H(X_j) \times (y - y_j) \right] \tag{60}$$

   где $k(y) = \sum_{j=1}^{n} k(y, y_j)$

6.7.3 Теоретические гарантии

GSE обеспечивает асимптотическую близость многообразий и касательных пространств:
- $d_H(\mathcal{M}_{h,g}, \mathcal{M}) = O(n^{-2/(q+2)})$ – совпадает с асимптотически минимаксной нижней границей
- $d_{P,2}(L(X), L_{h,g}(X)) = O(n^{-1/(q+2)})$

6.7.4 Преимущества и недостатки

Преимущества:
- Сохраняет дифференциальную структуру многообразия
- Обеспечивает точную реконструкцию
- Имеет теоретические гарантии оптимальности
- Эффективен для последующих задач регрессии на многообразии
- Обеспечивает естественное обобщение на новые точки

Недостатки:
- Вычислительная сложность
- Сложность реализации
- Множество параметров, требующих настройки
- Требует большой обучающей выборки
- Сложен для интуитивного понимания

6.8 Manifold Learning Regression

Manifold Learning Regression (MLR) – подход к задаче регрессии, который учитывает многообразную структуру пространства входных данных для построения более точных предсказательных моделей.

6.8.1 Задача регрессии на многообразии

Пусть:
- $X \subset \mathbb{R}^q$ – входное многообразие размерности q
- $f : X \to \mathbb{R}^m$ – неизвестная гладкая функция, которую нужно аппроксимировать
- Дан набор пар "вход-выход": $\{(x_i, y_i = f(x_i))\}_{i=1}^n$

Задача: построить функцию $f^* : X \to \mathbb{R}^m$, аппроксимирующую f с минимальной ошибкой.

6.8.2 Основные проблемы регрессии на многообразиях

1. **Проклятие размерности**: в высокоразмерных пространствах данные разрежены, что затрудняет обучение
2. **Нестационарность градиентов**: традиционные методы с стационарными ядрами плохо работают для функций с сильно изменяющимися градиентами
3. **Игнорирование геометрии**: стандартные методы регрессии не учитывают геометрию входного пространства

6.8.3 Подходы к регрессии на многообразиях

1. Двухэтапный подход

1. **Снижение размерности**:
   - Применить метод снижения размерности к входным данным: $z_i = h(x_i) \in \mathbb{R}^d$, $d < q$
   - Использовать методы вроде PCA, Kernel PCA, ISOMAP, LLE и т.д.
2. **Регрессия в пространстве меньшей размерности**:
   - Построить регрессионную модель $g : \mathbb{R}^d \to \mathbb{R}^m$ для пар (z_i, y_i)
   - Итоговая функция: $f^*(x) = g(h(x))$

Недостатки: снижение размерности выполняется независимо от задачи регрессии, что может привести к потере важной для прогнозирования информации.

2. Manifold Learning Regression (MLR) Подход, предложенный Bernstein, Kuleshov (2015-2018), основан на использовании регрессионного многообразия.

1. **Регрессионное многообразие**:
   - Неизвестная функция $f : X \subset \mathbb{R}^q \to \mathbb{R}^m$ определяет отображение:
   $$F(x) = \begin{pmatrix} x \\ f(x) \end{pmatrix} \in \mathbb{R}^{q+m} \tag{61}$$
   - $M(f) = \{F(x) = (x, f(x)) \in \mathbb{R}^{q+m} : x \in X \subset \mathbb{R}^q\}$ – q-мерное регрессионное многообразие
2. **Обучение многообразия**:
   - Обучающая выборка: $Z_n = \{Z_i = F(x_i) = (x_i, y_i = f(x_i))\}_{i=1}^n \subset M(f)$
   - Применение GSE к Z_n для получения отображений h_{GSE} и g_{GSE}
   - Аппроксимация регрессионного многообразия: $M_{GSE} \approx M(f)$
3. **Оценка функции регрессии**:
   - Для входа $x \in X$, вычислить выход:
   $$f_{MLR}(x) = g_{GSE,y}(\phi_{MLR}(x)) \tag{62}$$
   - где $\phi_{MLR}(x)$ – оценка точки $h_{GSE}(F(x))$ без знания $f(x)$
   - $g_{GSE,y}$ – компонента отображения g_{GSE}, соответствующая выходу y
4. **Оценка матрицы Якоби**:
   $$G_{MLR}(x) = G_{GSE,y}(\phi_{MLR}(x)) \times G^{-1}_{GSE,x}(\phi_{MLR}(x)) \tag{63}$$
   - где $G_{GSE,y}$ и $G_{GSE,x}$ – компоненты матрицы Якоби G_{GSE}

6.8.4 Преимущества MLR

1. **Нелинейное перепараметризование**:
   - MLR неявно находит параметризацию $\phi(x)$, которая делает задачу регрессии более линейной
   - Оценка функции $g_y(\phi(x)) \approx f(x)$ в перепараметризованном пространстве
2. **Использование дифференциальной структуры**:
   - Учет касательных пространств позволяет лучше аппроксимировать локальное поведение функции
   - Особенно полезно для функций с сильно изменяющимися градиентами
3. **Теоретические гарантии**:
   - Доказаны сходимость и оптимальность метода
   - Сравнимая или лучшая асимптотическая скорость сходимости по сравнению с традиционными методами

6.8.5 Эмпирические результаты

Метод MLR показывает значительное превосходство над традиционными методами регрессии (гауссовские процессы, RKHS) для:
- Функций с сильно изменяющимися градиентами
- Данных с нелинейной структурой входного пространства
- Задач с высокой размерностью входа

6.8.6 Расширения и варианты

1. **Регрессия на неизвестных входных многообразиях**:
   - Входные данные лежат на неизвестном многообразии меньшей размерности $M \subset \mathbb{R}^p$
   - Одновременное оценивание структуры входного многообразия и функции регрессии

2. **Регрессия на многообразии с неоднородными ядрами**:
   - Использование нестационарных ядер, адаптирующихся к локальной геометрии данных
   - Комбинация с методами отбеливания и локальной нормализации

3. **Плотность распределения на многообразии**:
   - Оценка плотности распределения данных на многообразии
   - Использование в задачах обнаружения аномалий и кластеризации