**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

Доклад

**Тема: Методы редукции данных**

Выполнили: Чепасов Д.В, Щедрин А.А.

Группа:2382

Санкт-Петербург

2023

Оглавление

[PCA (Метод главных компонентов) 3](#_Toc133663150)

[**1.1)** **Математическое обоснование** 3](#_Toc133663151)

[**1.2)** **Роль в машинном обучении** 4](#_Toc133663152)

[**1.3)** **Реализация в прикладной задаче** 5](#_Toc133663153)

[**1.4)** **Плюсы и минусы метода** 6](#_Toc133663154)

[ISOMAP 7](#_Toc133663155)

[**2.1) Математическая теория** 7](#_Toc133663156)

[Метод максимального правдоподобия 12](#_Toc133663158)

[**3.1) Математическая теория** 12](#_Toc133663159)

[**3.2) Роль в машинном обучении** 13](#_Toc133663160)

[**3.3) Реализация для прикладной задачи** 14](#_Toc133663161)

[**3.4) Плюсы и минусы метода** 15](#_Toc133663162)

[Сравнение методов 16](#_Toc133663163)

**PCA (Метод главных компонентов)**

* 1. **Математическое обоснование**
  2. **Роль в машинном обучении**
  3. **Реализация в прикладной задаче**
  4. **Плюсы и минусы метода**

**ISOMAP**

**2.1) Математическая теория**

В современном мире в датасетах может быть слишком много признаком и данных, которые замедляют и делают менее эффективными машинное обучение. Для лучшей работы программы, были придуманы методы редукции данных, которые позволяют уменьшить размерность данных, не теряя точности обучения. Рассмотрим один из таких методов ISOMAP (Isometric mapping).

Этот метод был придуман еще в 20 веке. Он является одним из самых первых методов нелинейной редукции данных.

Рассмотрим основные моменты теории, для понимания работы алгоритма.

Многие методы нуждаются в центрированном датасете. Это эквивалентно удалению средних значений из каждой характеристики датасета.

Предположим, у нас есть датасет . Возьмем столбец и вычтем из него . Здесь – средние значения характеристик столбца, – единичный вектор n. Проделаем это для каждого столбца нашего датасета. Возьмем стандартное отклонение j-того столбца , чтобы получить *стандартизированную матрицу*, необходимо для каждого столбца применить следующее действие: .

Симметричная матрица называется *центрированной*, когда её умножение на матрицу X дает такой же эффект, что и вычитание среднего значения из каждого элемента. Считается по формуле для каждого столбца:

. Здесь I – симметричная матрица, J –квадратная матрица размерностью n, состоящая из единиц.

СКО – среднеквадратичное отклонение. Если СКО = 0, то элементы в столбцах одинаковые.

Ковариация - . Показывает связь между элементами. Если *cov(X,Y) > 0,* то увеличение X соответствует увеличению Y. Если *cov(X,Y) < 0*, то увеличение X соответствует уменьшению Y. Если *cov(X,Y) = 0*, то X и Y не связаны.

Ковариационная матрица случайного вектора – матрица размерностью Вида . Где .

Спектральное разложение матрицы: . Здесь S – диагональная матрица из собственных значений, V – Базис собственный векторов.

SVD (Singular Value Decomposition). Для любой вещественной матрицы существуют две ортогональные матрицы U и V (n x n) такие, что . Где L – диагональная матрица, где .

**ТОПОЛОГИЯ.**

Топологическое пространство – множество, снабженное топологической структурой.

Топологическая структура(τ) – некоторое множество подмножеств X. Подмножества множества, входящие в топологию, называют открытыми. При этом должны выполняться следующие свойства:

1)Объединение любого наборам открытых множеств открыто

2)Пересечение конечного набора открытых множеств открыто

3)Пустое множество и все множество X открыто

Топологическое пространство *X*называется *локально евклидовым размерности n****,***если всякая его точка *х0*обладает окрестностью **U**, гомеоморфной ***n***-мерному евклидову пространству  или полупространству.

Топологическое пространство **X** называется *n-мерным топологическим многообразием****,***если оно:

**a)** локально евклидово размерности ***n***;

**б )** хаусдорфово;

**в)** обладает покрытием, состоящим из не более чем счетного числа евклидовых открытых множеств.

Множество A = называется атласом на многообразии M, если .

**ГРАФЫ.**

Пусть у нас есть пара (V,E), где V – множество объектов, названные вершинами( узлы, точки), E – семейство элементов, названные, дугами. G = (V, E) – граф.

Простой граф – граф без петель и кратных ребер.

Степень вершины – количество ребер, концов которых является эта вершина.

Матрица смежности – один из способов представления графа в виде матрицы:

**ВСПОМОГАТЕЛЬНЫЕ АЛГОРИТМЫ.**

**Алгоритм Флойда – Уоршелла**.

Представим у нас есть взвешенный, ориентированный граф с количеством ребер n (граф, в котором все ребра имеют вес). Его мы записываем в матрицу смежности размерностью n на n A. Все элементы главной диагонали в этой матрице равны нулю. Если петля направлена из i в i вершину, то такой путь равен 0. Если между вершинами нет ребра, то в эту ячейку матрицы записывается бесконечность. Далее с помощью формулы: . Применив данный алгоритм, мы получим матрицу, в который находятся кратчайшие пути между двумя вершинами.

**Алгоритм Дейкстры.**

Один из популярных алгоритмов, и не очень трудных. Суть его заключается в том, чтобы найти самые кратчайшие пути и вершины i во все остальные вершины. Чтобы это сделать возьмем ориентированный взвешенный граф, где каждый весь больше нуля. Шаги:

1. Создать список вершин и пометить все вершины, кроме начальной, как не посещенные.

2. Установить начальную вершину и ее расстояние до себя равным 0.

3. Для каждой смежной вершины, которая еще не была посещена, вычислить расстояние до нее от начальной вершины через текущую вершину.

4. Если это расстояние меньше, чем текущее расстояние до этой вершины, то обновить текущее расстояние.

5. Пометить текущую вершину как посещенную и выбрать следующую вершину с наименьшим расстоянием из списка непосещенных вершин.

6. Повторять шаги 3-5 для всех непосещенных вершин до тех пор, пока не будут посещены все вершины.

**MDS.**

Рассмотрим классический MDS. По шагам:

1. Составляем матрицу из Евклидовых расстояний между двумя точками.
2. Используем двойное центрирование
3. Найдем собственные числа , и собственные вектора (m – желаемая размерность выходящей матрицы)
4. Считаем . Здесь – матрица собственных векторов, – диагональная матрица, собственных чисел.

**K-NN.**

С помощью данного алгоритма K-NN (Поиск ближайших соседей) мы можем определить категорию или класс определенного датасета. Рассмотрим шаги алгоритма:

1. Выберем число K соседей
2. Посчитаем Евклидовы расстояния между ними
3. Возьмем K ближайших соседей в соответствии рассчитанным евклидовым расстоянием.
4. Среди этих k соседей посчитаем количество точек в каждой категории.
5. Отнесем новые точки к той категории, для которых количество соседей является максимальным.

**ISOMAP**

Шаги алгоритма:

1. Строим взвешенный граф G, из евклидовых расстояний и находим подграф применяя K-NN алгоритм для на граф G.
2. Считаем кратчайшие пути между всеми парами узлов, используя алгоритм Флойда – Уоршелла или Дейкстры.
3. Понижаем размерность, используя MDS.

**Метод максимального правдоподобия**

**3.1) Математическая теория**

**3.2) Роль в машинном обучении**

**3.3) Реализация для прикладной задачи**

**3.4) Плюсы и минусы метода**

**Сравнение методов**