**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

**Введение**

Бывает такое, что объем данных слишком большой и просто обучать модель по ним нерационально, существуют много способов устранить эту проблему, в данной работе будут рассмотрены некоторые алгоритмы, которые помогут решить данную проблему, они созданы для редукции данных, а также будет рассказано их теоретическое и математическое устройство. Алгоритмы для понижения размерности очень важная вещь в машинном обучении, так как они позволяют ускоряют обучение модели. Уменьшая размерность и количество данных разными способами, не только с помощью редукции, уменьшается и выборка, по которой будет обучаться модель, однако важно, чтобы результат оставался таким же точным, как и до алгоритмов, поэтому будет проверка точности предсказаний в программе, чтобы убедиться в работоспособности алгоритмов.

**PCA (Метод главных компонентов)**

Используется в задачах, где есть много параметров (например, распознавание изображений) или когда нужно визуализировать данные

1. Центрирование данных: Вычитание среднего значения каждого признака из соответствующего значения в данных. Это делается для того, чтобы получить нулевое среднее значение и уменьшить корреляцию между признаками.

2. Расчет ковариационной матрицы: Вычисление ковариационной матрицы для центрированных данных. Ковариация - . Показывает направление и степень взаимосвязи 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– матрица размерностью Вида . Где .

3. Расчет главных компонент: Поиск собственных векторов (и соответствующие им собственные значения) для ковариационной матрицы. Собственные векторы показывают направления, вдоль которых данные имеют наибольшую дисперсию, а собственные значения - меру этой дисперсии.

4. Сортировка главных компонент: Сортировка собственные значения в порядке убывания, чтобы определить, какие главные компоненты являются наиболее значимыми.

5. Проекция данных на новое пространство признаков: Построение матрицы проекции, состоящей из выбранных главных компонент. Умножение центрированных данных на эту матрицу проекции для получения новых координат в пространстве главных компонент.

Метод PCA как правило применяется, если нужно визуализировать данные в 2 или 3 мерном пространстве, если данные требуется сжать, чтобы избавиться от шума. Полезен в задачах, где признаков очень много (например, распознавание изображений).

К минусам метода относится возможность потери важных данных и сложность интерпретации результатов.

**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.

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 набора 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, т.е. пространство, в котором для любых двух точек существуют непересекающиеся окрестности.

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

Множество A = называется атласом на многообразии M, если и – гомеоморфизм на открытое подмножество . Класс эквивалентности гладких атласов называется *гладкой структурой*. Многообразие с гладкой структурой называется *гладким многообразием.*

**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**

Метод Isomap используется в машинном обучении для нелинейного снижения размерности данных. Он применяется в случаях, когда данные имеют сложную нелинейную структуру и не могут быть эффективно представлены в пространстве меньшей размерности с помощью линейных методов.

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.

Данный алгоритм, мы можем применять только в некоторых случаях:

1. Когда датасет геодезически выпуклый, т.е. любые две точки в нем, могут быть соединены кратчайшим путем, который находится в самом датасете. Если это не будет выполняться, то будут образовываться дырки,в результате чего получается граф окрестностей с большими несвязными областями, из-за которых требуются пути с больший кривизной.
2. Данные лежат на низкоразмерном многообразии, иначе будут получены плохие низкоразмерные представления.

**Метод максимального правдоподобия**

Метод максимального правдоподобия используется для оценки параметров распределений, проверки статистических гипотез, прогнозирования.

1. Определение вероятностной модели: Определяется вероятностная модель, описывающая процесс генерации данных. Например, если данные являются независимыми и одинаково распределенными (i.i.d.), то можно использовать параметрическую модель, такую как нормальное распределение.

2. Формулирование функции правдоподобия: Формулируется функция правдоподобия, которая описывает вероятность получения наблюдаемых данных при заданных параметрах модели. Функция правдоподобия представляет собой произведение вероятностей каждого наблюдения в выборке.

3. Логарифмирование функции правдоподобия: Функция правдоподобия может иметь очень маленькие значения, что затрудняет ее оптимизацию. Поэтому обычно логарифмируют функцию правдоподобия, что упрощает вычисления и не изменяет положения ее максимума.

4. Оптимизация функции правдоподобия: Цель состоит в том, чтобы найти значения параметров модели, при которых функция правдоподобия принимает наибольшее значение. Это можно сделать путем дифференцирования логарифма функции правдоподобия по параметрам модели и приравнивания производных к нулю, а затем решением системы уравнений.

Метод максимального правдоподобия может применяться для различных видов распределения. Точность предсказаний, как правило, получается высокой. Но, этот метод можно использовать лишь тогда, когда можно сделать предположение о распределении данных.