ДИСЦИПЛИНА	Прикладная математика
ИНСТИТУТ	ИПТИП
КАФЕДРА	Индустриального программирования
ВИД УЧЕБНОГО МАТЕРИАЛА	Методические указания по дисциплине
ПРЕПОДАВАТЕЛЬ	Астафьев Рустам Уралович
СЕМЕСТР	1 семестр, 2025/2026 уч. год

Ссылка на материал:
https://github.com/astafiev-rustam/applied-mathematics/tree/lecture-1-3

Лекция №3: Кластеризация данных

Введение в кластеризацию данных

Кластеризация данных является одним из основных методов машинного обучения без учителя. Ее цель состоит в том, чтобы разделить множество объектов на группы, называемые кластерами, таким образом, чтобы объекты внутри одного кластера были схожи между собой, а объекты из разных кластеров существенно различались. В отличие от классификации, заранее заданные метки классов отсутствуют. Алгоритм сам обнаруживает скрытую структуру данных, основываясь solely на их свойствах. Этот метод находит применение в сегментации клиентов, биоинформатике для группировки генов и в социальных науках.

Задача кластеризации данных

Формально задача кластеризации ставится следующим образом. Имеется набор данных, представленный в виде точек в многомерном пространстве признаков. Требуется найти такое разбиение этого набора на k групп, которое бы минимизировало некоторый функционал качества. Чаще всего этим функционалом выступает сумма внутрикластерных дисперсий. Чем меньше разброс точек внутри каждого кластера, тем более однородными и компактными они являются. Качество кластеризации также оценивается по тому, насколько хорошо кластеры separated друг от друга.

Методы классификации данных

Хотя термины "кластеризация" и "классификация" иногда используются как синонимы, они обозначают принципиально разные задачи. Классификация это задача обучения с учителем. В этом случае алгоритму предоставляется размеченная обучающая выборка, где каждому объекту присвоена метка класса. Цель алгоритма научиться на основе этих примеров присваивать правильные метки новым, ранее не виденным объектам. Примером классификации является определение категории электронного письма как "спам" или "не спам". Таким образом, ключевое различие заключается в наличии заранее известных ответов для обучения в задаче классификации и их отсутствии в задаче кластеризации.

Примеры классификации данных

Методы классификации широко применяются в самых разных областях. В финансах кредитный скоринг использует классификацию для отнесения заемщиков к категориям "низкий риск" или "высокий риск" на основе их кредитной истории и доходов. В медицине диагностические системы классифицируют медицинские изображения, например, определяя наличие опухоли на снимке МРТ. В сфере безопасности системы распознавания лиц классифицируют изображения, идентифицируя человека по чертам лица. Во всех этих случаях алгоритм обучается на большом массиве размеченных исторических данных.

Подходы к кластеризации

Среди множества алгоритмов кластеризации можно выделить несколько ключевых семейств. Иерархическая кластеризация строит древовидную структуру кластеров, которая может быть полезной для визуального анализа. Алгоритмы на основе центроидов, такие как k-means, стремятся найти центры кластеров и отнести каждый объект к ближайшему центру. Плотностные алгоритмы, например DBSCAN, формируют кластеры из плотно расположенных точек, что позволяет находить кластеры произвольной формы и выделять выбросы. Выбор конкретного метода зависит от природы данных и поставленной задачи.

Примеры и реализация

Рассмотрим примеры по теме лекционного занятия:

Пример 1

Пример 2

Пример 3