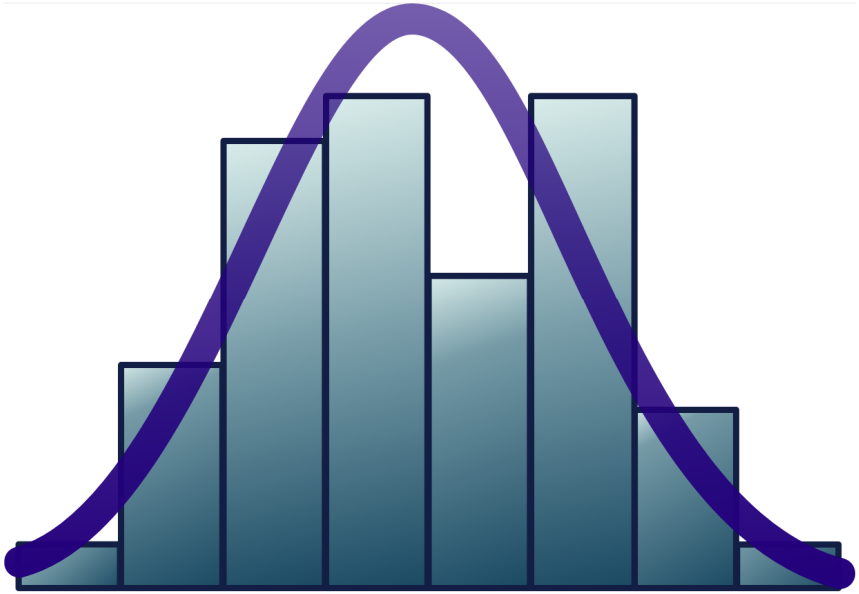
Математические методы анализа технологической информации



Лектор: А.А. Ефремов

Томский политехнический университет, 2021

Контактная информация

ЕФРЕМОВ Александр Александрович

Старший преподаватель,
Отделение автоматизации и робототехники,
ИШИТР

Ауд. 115а, 10к.
email: *alexyefremov@tpu.ru*

Временной ресурс

Лекции	8 часов
Практики	24 часа
Лабораторные	16 часов
Самостоятельная работа	60 часов

Промежуточная аттестация: ЗАЧЕТ

Рекомендованная литература

Храмов, А. Г. Методы и алгоритмы интеллектуального анализа данных : учебное пособие / *Самара : СамГУ, 2019. — 176 с. —URL: https://e.lanbook.com/book/148603— Режим доступа: для авториз. пользователей.*

Низаметдинов, Ш. У. Анализ данных : учебное пособие / *М: НИЯУ МИФИ, 2012. — 288 с. —URL: https://e.lanbook.com/book/75847 — Режим доступа: для авториз. пользователей.*

Лекция 1

Введение в математические методы анализа данных

Введение в математические методы анализа данных

Анализ данных — область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных.

Анализ данных — процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений.

Введение в математические методы анализа данных

Интеллектуальный анализ данных (*ИАД, Data Mining, Knowledge Discovery in Databases, KDD*) - собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Введение в математические методы анализа данных

Методы ИАД

статистические

- корреляционный анализ
- регрессионный анализ
- дисперсионный анализ
- анализ временных рядов
- и др.

иные

- классификация
- кластеризация
- деревья решений
- прогнозирование
- нечёткая логика
- нейронные сети
- генетические алгоритмы

Введение в математические методы анализа данных

Основная польза разрабатываемых методов анализа данных заключается в некоторой предсказательной способности:

проанализировав некоторый набор данных, информационная система анализа данных должна обучиться для дальнейшего распознавания или прогнозирования некоторых участков данных в ситуациях, когда часть данных утеряна или неизвестна.

Кроме этого, системы анализа данных могут решать задачи

- редуцирования объёма данных с целью устранения избыточности,
- визуализации данных для их удобного восприятия человеком,
- моделирования новых данных по имеющимся данным
- и др.

Введение в математические методы анализа данных

Смежная область: Машинное обучение (*Machine Learning*) – раздел искусственного интеллекта, математическая дисциплина, использующая математическую статистику, численные методы оптимизации, теорию вероятностей, выделяющая знания из данных.

Различают два типа обучения:

1. Обучение по прецедентам, или *индуктивное обучение*, – основано на выявлении закономерностей в эмпирических данных.
2. Дедуктивное обучение – предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний; относится к области экспертных систем, поэтому термины *машинное обучение* и *обучение по прецедентам* можно считать синонимами.

Введение в математические методы анализа данных

Данные. Типы данных

При выполнении анализа данных мы рассматриваем некое пространство объектов, обычно довольно большое. Однако, мы имеем доступ к малой части объектов из этого пространства.

Целью анализа данных является получение информации из имеющихся данных, которая будет применима к большому объему данных, недоступных для нас.

Введение в математические методы анализа данных

Каждый объект описывается некоторым количеством переменных, отражающих его свойства. Часто эти переменные называют атрибутами.

Множество значений переменных (атрибутов), соответствующих каждому из объектов, называется экземпляром.

Набор анализируемых данных часто представлен в виде таблицы, строки которой соответствуют экземплярам, а столбцы – атрибутам.

Введение в математические методы анализа данных

Атрибут 1	Атрибут 2	Атрибут 3	Атрибут 4	Класс
A	B	10	низкий	3
A	B	8	средний	1
B	C	12	средний	2
A	A	-2	высокий	2
C	A	0	высокий	3
…	…	…	…	…
C	A	2	низкий	3

Если некоторому атрибуту придается особое значение, и целью анализа является предсказание этого значения, то такой набор данных называется *размеченным*.

Введение в математические методы анализа данных

Интеллектуальный анализ данных с использованием размеченных данных известен как обучение с учителем (supervised learning).

Распространенные задачи:

- классификация;
- регрессия (искусственные нейронные сети).

Введение в математические методы анализа данных

Если же данные являются неразмеченными, интеллектуальный анализ данных с их использованием известен как обучение без учителя (unsupervised learning).

Распространенные задачи:

- поиск ассоциативных правил;
- кластеризация.

Введение в математические методы анализа данных

В общем, существует большое количество типов переменных, использующихся для измерения (характеризации) свойств объектов.

Недостаточное понимание различий между типами данных может привести к серьезным проблемам при выполнении анализа данных.

Можно выделить по крайней мере шесть различных типов.

Категориальные (номинальные) переменные

Переменные используются для разделения объектов на категории, например, по цвету.

Категориальные переменные могут быть представлены числами, но значения в этом случае не имеют математической интерпретации.

К примеру, мы можем обозначить 10 человек числами от 1 до 10, но любые математические операции с этими числами являются бессмысленными.

Двоичные (бинарные) переменные

Частный случай категориальных переменных. Принимают два возможных значения: истина-ложь, 1 или 0.

Порядковые (ранговые) переменные

Порядковые переменные схожи с категориальными, за исключением того, что их значения могут быть упорядочены в некотором осмысленном порядке, например, «малый», «средний», «большой».

Целочисленные переменные

В отличие от категориальных переменных, выраженных в численной форме, математические операции имеют смысл.

Переменные шкалы интервалов

Переменные выражаются численными значениями, представляющими расстояние от некоторой начальной точки отсчета. Однако, точка начала не подразумевает истинное отсутствие измеряемой характеристики и может быть выбрана произвольно.

Переменные шкалы отношений (абсолютной шкалы)

Понятие схоже с переменными шкалы интервалов, однако, точка начала означает отсутствие измеряемой характеристики.

Зачастую, во многих приложениях анализа данных достаточно выделить лишь два типа переменных:

- категориальные;
- непрерывные.

Очистка данных

Данные, содержащиеся в базе данных, могут содержать ошибочные значения.

Ошибки в данных допускаются по различным причинам:

- ошибки измерений;
- субъективные суждения;
- сбои регистрирующей автоматики;
- и др.

Ошибки в данных могут быть разделены на два класса:

- ошибочные возможные значения атрибутов;
- невозможные значения.

Примеры:	**шум**	**ошибочное значение**
6,972 :	69,72	6,s72
«красный»:	«синий»	«крассный»

Даже очевидные ошибки в данных можно заметить с трудом, если объем данных очень велик.

Для выявления ошибок часто используются программные средства, в частности, средства визуализации, помогающие обнаружить выбросы или необычные концентрации значений.

Однако, даже простейшая сортировка по возрастанию может выявить неожиданные результаты.

Примеры:

- Численная переменная принимает только шесть значений, расположенных достаточно далеко друг от друга. Возможно, будет лучше рассматривать эти данные как категориальные.

- Все значения одинаковые. Переменную можно проигнорировать.

Примеры:

- Все значения, кроме одного, одинаковые. Необходимо выяснить, является ли это единственное значение ошибкой. Если нет, можно представить атрибут двоичной переменной.

Примеры:

- Некоторое значение атрибута встречается необычно часто.

К примеру, при заполнении регистрационной формы на сайте отмечается, что в поле «Страна» в 20% случаев указано «Албания».

Примеры:

- Некоторые значения находятся далеко за пределами нормального диапазона изменения переменной. Например,

  [200; 5000]: 22654,8 38597 44625,7

Необходимо отметить, что аномальные значения могут в действительности быть *подлинными выбросами* – истинными значениями, отличающимися от обычных.

Распознавание выбросов и их значения может быть ключом к открытиям, в особенности в таких сферах деятельности, как физика или медицина.

Пропуски в данных

Во многих реальных базах данных могут быть записаны не все значения атрибутов. Это может произойти, например, по причине того, что не все атрибуты применимы для всех экземпляров.

К примеру, некоторые медицинские данные применимы только к пациентам-женщинам или к пациентам определенной возрастной группы.

В таких случаях целесообразно разделить набор данных на две (или более) части и анализировать их отдельно.

Также возможно, что отсутствуют значения атрибутов, которые нормально должны присутствовать. Это может произойти по разным причинам.

Например,

- сбой регистрирующего оборудования;
- изменение форм регистрации данных в ходе их сбора;
- информация не доступна по объективным причинам.

Существуют несколько возможных стратегий при обнаружении пропущенных данных.

Две самые очевидные из них:

- исключить экземпляры с отсутствующими данными;
- заполнить пробелы приблизительными значениями.

Исключение экземпляров с отсутствующими значениями атрибутов представляется простейшим способом выхода из ситуации: попросту удалить все записи, в которых пропущено хотя бы одно значение какого-нибудь атрибута.

Эта стратегия очень консервативна.

С одной стороны, мы избегаем внесения ошибочных данных. С другой, исключая неполные данные, мы можем значительно уменьшить надежность результатов анализа.

Менее осторожной стратегией является оценивание пропущенных значений по имеющимся данным.

Простым, но эффективным способом при работе с категориальными атрибутами является замена отсутствующих значений на наиболее часто встречающееся из имеющихся.

Значение атрибута	A	B	C
Частота	85%	10%	5%

Значение атрибута	A	B	C
Частота	**85%**	**10%**	**5%**

Если значения атрибутов распределены более равномерно, адекватность такого подхода является спорной.

В случае числовых данных в этом случае зачастую используется среднее значение.

Замена отсутствующих значений атрибутов на их оценки, разумеется, может внести шум в данные, но если доля отсутствующих значений сравнительно мала, эффект от искажения может быть также невелик.

В любом случае, необходимо с осторожностью использовать методы оценивания пропущенных данных.

Сокращение числа атрибутов

Во многих областях интеллектуального анализа данных доступность хранилищ информации большого объема при их постоянно снижающейся цене приводит к тому, что для каждого экземпляра может записываться огромное количество атрибутов, например

- информация о покупках каждого покупателя за полгода;
- большой объем детализированных данных о пациентах в больнице.

Иногда, число атрибутов может быть в 10 и даже в 100 раз больше, чем количество экземпляров.

Хотя хранение огромного количества информации для каждого экземпляра может показаться хорошей идеей (можно не задумываться, какая информация в действительности нужна), в этом подходе присутствуют определенные риски.

Предположим, имеется 10000 позиций данных о каждом покупателе супермаркета. Необходимо предсказать, какие покупатели купят новый бренд еды для собак.

Число атрибутов, имеющих отношение к этой задаче, вероятно, очень мало.

В лучшем случае обработка ненужных данных увеличит время работы алгоритма. В худшем – приведет к искажению результатов.

Можно возразить, что не всегда известно, какие данные являются (или будут являться) релевантными для конкретной задачи. Безопаснее собирать большой объем данных и хранить его.

Даже если проблема хранения и обработки больших объемов данных не является критичной, всегда существует риск того, что полученные результаты будут менее основательными, чем результаты, полученные при обработке небольшого количества атрибутов.