1 **Что подразумевают под функциональной и статистической зависимостью? Назовите частные случаи статистичесокй зависимости**.

Функциональная зависимость – это связь, при которой каждому значению независимой переменной *x* соответствует точно определенЕ ное значение *y*. В экономических процессах такой вид зависимости между переменными встречается редко. Для этих процессов характерно взаимодействие случайных факторов. Существующая зависимость между признаками может проявляться не в каждом отдельном случае, а лишь «в общем и среднем» при большом количестве наблюдений.

2) *Статистическая зависимость* – это связь, при которой каждому значению независимой переменной *x* соответствует множество значений зависимой переменной *y*, причем заранее не известно, какое именно значение примет *y*. Частным случаем статистической зависиЕ мости является *корреляционная зависимость*.

Корреляционная зависимость – это связь, при которой каждому значению независимой переменной *x* соответствует определенное средЕ нее значение (математическое ожидание) зависимой переменной *y*.

В регрессионном анализе рассматриваются односторонние зависиЕ мости случайной переменной *Y* от неслучайной независимой переменЕ ной *X*. Такая зависимость может возникнуть, когда при каждом фиксированном значении *X* соответствующие значения *Y* подвержены случайному разбросу за счет действия ряда неконтролируемых фактоЕ ров. Тогда зависимость между *X* и *Y*, представленную в виде соотноЕ шения $Y = f(X)$ называют *модельным уравнением регрессии* (или просто *уравнением регрессии*). Функциональная и корреляционная связь в зависимости от направЕ ления действия бывает *прямая* и *обратная*. По аналитическому выражеЕ нию зависимость может быть прямолинейной (*линейной*) и криволиЕ нейной (*нелинейной*). В зависимости от количества признаков, включенных в модель, корреляционные связи делят на однофакторные (*парные,* $Y = f(X)$) и многофакторные (*множественные,* $Y = f(X_1, X_2, ..., X_n)$

2 **Что же такое Data Mining? Что обусловило его развитие?**

Data Mining - мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теориябаз данных и др., см. рис. 1.1.

Data Mining - это процесс поддержки принятия решений, основанный на поиске в данныхскрытых закономерностей (шаблонов информации) [3].

Технологию Data Mining достаточно точно определяет Григорий Пиатецкий-Шапиро(Gregory Piatetsky-Shapiro) - один из основателей этого направления:

Data Mining - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Возникновение и развитие Data Mining обусловлено различными факторами, основныесреди которых являются следующие [2]:

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи данных;
- накопление большого количества ретроспективных данных;
- совершенствование алгоритмов обработки информации.

3 **Как можно охарактеризовать суть и цель технологии Data Mining**

Суть и цель технологии Data Mining можно охарактеризовать так: это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

Неочевидных - это значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

Объективных - это значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

Практически полезных - это значит, что выводы имеют конкретное значение, которомуможно найти практическое применение.

Знания - совокупность сведений, которая образует целостное описание, соответствующее некоторому уровню осведомленности об описываемом вопросе, предмете, проблеме и т.д.

Использование знаний (knowledge deployment) означает действительное применение найденных знаний для достижения конкретных преимуществ (например, в конкурентнойборьбе за рынок).

4 Понятие Business Intelligence. Состав рынка BI

Business Intelligence - программные средства, функционирующие в рамках предприятия и обеспечивающие функции доступа и анализа информации, которая находится в хранилище данных, а также обеспечивающие принятие правильных и обоснованных управленческих решений.

Понятие BI объединяет в себе различные средства и технологии анализа и обработкиданных масштаба предприятия.

На основе этих средств создаются BI-системы, цель которых - повысить качество информации для принятия управленческих решений.

BI-системы также известны под названием Систем Поддержки Принятия Решений (СППР,DSS, Decision Support System). Эти системы превращают данные в информацию, на основекоторой можно принимать решения, т.е. поддерживающую принятие решений.

Gartner Group определяет состав рынка систем Business Intelligence как набор программных продуктов следующих классов:

- средства построения хранилищ данных (data warehousing, ХД);
- системы оперативной аналитической обработки (OLAP);
- информационно-аналитические системы (Enterprise Information Systems, EIS);
- средства интеллектуального анализа данных (data mining);
- инструменты для выполнения запросов и построения отчетов (query and reporting tools).

Классификация Gartner базируется на методе функциональных задач, где программныепродукты каждого класса выполняют определенный набор функций или операций с использованием специальных технологий.

5 Отличия Data Mining от других методов анализа данных

Традиционные методы анализа данных (статистические методы) и OLAP в основном ориентированы на проверку заранее сформулированных гипотез (verification-driven data mining) и на "грубый" разведочный анализ, составляющий основу оперативной аналитической обработки данных (OnLine Analytical Processing, OLAP), в то время как одно из основных положений Data Mining - поиск неочевидных закономерностей.

Инструменты Data Mining могут находить такие закономерности самостоятельно

и также самостоятельно строить гипотезы о взаимосвязях. Поскольку именно формулировка гипотезы относительно зависимостей является самой сложной задачей, преимущество Data Mining по сравнению с другими методами анализа является очевидным.

Большинство статистических методов для выявления взаимосвязей в данных используют концепцию усреднения по выборке, приводящую к операциям над несуществующими величинами, тогда как Data Mining оперирует реальными значениями.

OLAP больше подходит для понимания ретроспективных данных, Data Mining опираетсяна ретроспективные данные для получения ответов на вопросы о будущем.

6 Задачи Data Mining. Виды классификации

- системное распределение изучаемых предметов, явлений, процессов породам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация - упорядоченное по некоторому принципу множество объектов, которыеимеют сходные классификационные признаки (одно или несколько свойств), выбранныхдля определения сходства или различия между этими объектами.

Классификация требует соблюдения следующих правил:

- в каждом акте деления необходимо применять только одно основание;
- деление должно быть соразмерным, т.е. общий объем видовых понятий должен равняться объему делимого родового понятия;
- члены деления должны взаимно исключать друг друга, их объемы не должны перекрещиваться;
- деление должно быть последовательным.

Различают:

- вспомогательную (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- естественную классификацию, которая производится по существенным признакам, характеризующим внутреннюю общность предметов и явлений. Она является результатоми важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятийклассификация может быть:

- простой - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");
- сложной - применяется для деления одного понятия по разным основаниям и синтезатаких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Под классификацией будем понимать отнесение объектов (наблюдений,

событий) кодному из заранее известных классов.

Классификация — это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы. Таким образом, для проведения классификации должны присутствовать признаки, характеризующие группу, к которойпринадлежит то или иное событие или объект (обычно при этом на основании анализа уже классифицированных событий формулируются некие правила).

Классификация относится к стратегии обучения с учителем (supervised learning), котороетакже именуют контролируемым или управляемым обучением.

Задачей классификации часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборкинепрерывных и/или категориальных переменных.

Например, можно предсказать, кто из клиентов фирмы является потенциальным покупателем определенного товара, а кто - нет, кто воспользуется услугой фирмы, а кто - нет, и т.д. Этот тип задач относится к задачам бинарной классификации, в них зависимая переменная может принимать только два значения (например, да или нет, 0 или 1).

Другой вариант классификации возникает, если зависимая переменная может принимать значения из некоторого множества предопределенных классов. Например, когда необходимо предсказать, какую марку автомобиля захочет купить клиент. В этих случаяхрассматривается множество классов для зависимой переменной.

Классификация может быть одномерной (по одному признаку) и многомерной (по двум иболее признакам).

Многомерная классификация была разработана биологами при решении проблем дискриминации для классифицирования организмов. Одной из первых работ, посвященных этому направлению, считают работу Р. Фишера (1930 г.), в которой организмы разделялись на подвиды в зависимости от результатов измерений их физических параметров. Биология была и остается наиболее востребованной и удобной средой для разработки многомерных методов классификации.

7 Процесс классификации

состоит из двух этапов [21]: конструирования модели и ееиспользования.

1. Конструирование модели: описание множества предопределенных классов.
   - Каждый пример набора данных относится к одному предопределенному классу.
   - На этом этапе используется обучающее множество, на нем происходитконструирование модели.
   - Полученная модель представлена классификационными правилами, деревомрешений или математической формулой.
2. Использование модели: классификация новых или неизвестных значений.
   - Оценка правильности (точности) модели.
     1. Известные значения из тестового примера сравниваются с результатамииспользования полученной модели.
     2. Уровень точности - процент правильно

классифицированных примеров втестовом множестве.

3. Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.

– Если точность модели допустима, возможно использование модели дляклассификации новых примеров, класс которых неизвестен.

8 Методы, применяемые для решения задач классификации

Для классификации используются различные методы. Основные из них:

- классификация с помощью деревьев решений;
- байесовская (наивная) классификация;
- классификация при помощи искусственных нейронных сетей;
- классификация методом опорных векторов;
- статистические методы, в частности, линейная регрессия;
- классификация при помощи метода ближайшего соседа;
- классификация CBR-методом;
- классификация при помощи генетических алгоритмов.

9 Задача кластеризации. Цели, области применения

Задача кластеризации сходна с задачей классификации, является ее логическим продолжением, но ее отличие в том, что классы изучаемого набора данных заранее не предопределены.

Синонимами термина "кластеризация" являются "автоматическая классификация","обучение без учителя" и "таксономия".

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Если данные выборки представить как точки в признаковом пространстве, то задача кластеризации сводится к определению "сгущенийточек".

Цель кластеризации - поиск существующих структур.

Кластеризация является описательной процедурой, она не делает никаких статистическихвыводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Само понятие "кластер" определено неоднозначно: в каждом исследовании свои "кластеры". Переводится понятие кластер (cluster) как "скопление", "гроздь".

Кластер можно охарактеризовать как группу объектов, имеющих общие свойства. Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Вопрос, задаваемый аналитиками при решении многих задач, состоит в том, как организовать данные в наглядные структуры, т.е. развернуть таксономии.

10 Оценка качества кластеризации

Оценка качества кластеризации может быть проведена на основе следующих процедур:

– ручная проверка;
– установление контрольных точек и проверка на полученных кластерах;
– определение стабильности кластеризации путем добавления в модель новых переменных;
– создание и сравнение кластеров с использованием различных методов. Разные методы кластеризации могут создавать разные кластеры, и это является нормальным явлением. Однако создание схожих кластеров различными методами указывает на правильность кластеризации.

Опишите цикл аналитики больших данных

можно описать следующим этапом -

- Определение бизнес-проблемы
- Research
- Оценка человеческих ресурсов
- Получение данных
- Изменение данных
- Хранилище данных
- Исследовательский анализ данных
- Подготовка данных для моделирования и оценки
- Modeling
- Implementation

11 Роль и компетенции специалиста по данным

Основные навыки, которыми должен обладать компетентный аналитик данных, перечислены ниже:

- Деловое понимание
- SQL программирование
- Дизайн и реализация отчета
- Разработка дашбордов

Роль специалиста по данным обычно связана с такими задачами, как прогнозное моделирование, разработка алгоритмов сегментации, рекомендательные системы, фреймворки A / B-тестирования и часто работа с необработанными неструктурированными данными.

Характер их работы требует глубокого понимания математики, прикладной статистики и программирования. Между аналитиком данных и специалистом по анализу данных есть несколько общих навыков, например, способность запрашивать базы данных. Оба анализируют данные, но решение специалиста по данным может иметь большее влияние на организацию.

Вот набор навыков, которые обычно необходимы специалисту по данным:

- Программирование в статистическом пакете, таком как: R, Python, SAS, SPSS или Julia
- Возможность очищать, извлекать и исследовать данные из разных источников
- Исследование, разработка и внедрение статистических моделей

- Глубокие статистические, математические и компьютерные знания

12 **Сформулируйте свойства метрики**. **Какие метрики часто используют при кластеризации данных, измеренных в количественной шкале?**

Остановимся на метриках и сформулируем **свойства метрики**.

Пусть i, j, k – некоторые точки, а $d(i, j)$ – метрика, расстояние от точки i до точки j. Тогда:

1. $d(i, i) = 0$;
2. $d(i, j) \geq 0$;
3. $d(i, j) = d(j, i)$;
4. $d(i, k) \leq d(i, j) + d(j, k)$.

Переводя эти свойства на менее формальный язык, получим довольно логичные утвержде- ния: расстояние от точки до самой себя равно 0, расстояние не бывает отрицательным, рас- стояние от точки i до точки j – то же самое, что расстояние от точки j до точки i, расстояниеот точки i до точки k меньше суммы расстояния от точки i до точки j и расстояния от точки j до точки k (неравенство треугольника).

Давайте рассмотрим основные **виды метрик**, которые часто используют при кластеризации данных, измеренных **в количественной шкале**.

Для определённости давайте зафиксируем, что в p-мерном пространстве у нас есть две точки$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ и $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$. Если обозначения кажутся не совсем понятны- ми, посмотрите, как у нас записан массив X в постановке задачи (x_i и x_j – просто строки, соответствующие i-тому и j-тому наблюдению).

Итак, виды метрик для данных в количественной шкале:

про

1. Евклидово расстояние,
2. Квадрат евклидова расстояния, также обозначается L2 squared, необходимо для неко-торых методов агломерации, в частности, для метода Уорда (Варда):

3. Манхэттенское расстояние, оно же блочное расстояние, также обозначается L1:
4. Чебышева

13 **Назовите параметры кластеризации**

Кластерный анализ (скорее всего не нужно, по логике имеется ввиду иерархическая кластеризация):

Лекция: Анализ БД_лекция 6

Процесс кластеризации зависит от выбранного метода и почти всегда является итеративным. Он может стать увлекательным процессом и включать множество экспериментов по выбору разнообразных параметров, например, меры расстояния, типа стандартизации переменных, количества кластеров и т.д.

В таблице 5.2 приведено сравнение некоторых параметров задач классификации и кластеризации.

Таблица 5.2. Сравнение классификации и кластерзации

Характеристика	Классификация	Кластеризация
Контролируемость обучения	Контролируемое обучение	Неконтролируемое обучение
Стратегия	Обучение с учителем	Обучение без учителя
Наличие метки класса	Обучающее множество сопровождается меткой, указывающей класс, к которому относится наблюдение	Метки класса обучающего множества неизвестны
Основание для классификации	Новые данные классифицируются на основании обучающего множества	Дано множество данных с целью установления существования классов или кластеров данных

Пояснения:

Контролируемость обучения: Кластеризация не предполагает наличия заранее заданных меток классов.

Стратегия: Метки классов не предоставляются, алгоритмы пытаются определить структуры или группы данных самостоятельно.

Наличие метки класса: Данные не имеют предварительно определённых меток классов, что делает задачу неконтролируемой.

Основание для классификации: Основной целью является выявление групп или кластеров в данных, что помогает понять их внутреннюю структуру.

Иерархический кластерный анализ:

Лекция: Иерархический_кластерный_анализ_lecture09_1.pdf

В основе данного вида кластерного анализа лежат два предположения:

1. На самом первом шаге кластерного анализа количество кластеров совпадает с количеством наблюдений (имеем n кластеров, состоящих ровно из одного наблюдения).

2. Количество кластеров заранее неизвестно, мы объединяем точки в кластеры до тех пор, пока не получим один большой кластер. Так, на первом шаге иерархического кластерного анализа у нас есть n кластеров, на втором шаге (n−1) кластеров, на третьем уже (n−2) кластеров, и так далее, а на последнем шаге остаётся один кластер. Другими словами, мы выстраиваем некоторую иерархию из кластеров, вложенных друг в друга, а потом решаем, на каком делении, более детальном (много маленьких кластеров) или более общем (мало больших кластеров), стоит остановиться.

У иерархического кластерного анализа есть **два параметра кластеризации:**

1. Метрика: мера расстояния между точками (наблюдениями).

2. Метод агломерации или метод агрегирования: алгоритм, который позволяет решать, каким образом объединять точки в кластеры на основе выбранной метрики (про метрики в 12 вопросе)

14 **Назовите основные методы агломерации**

Лекция: Иерархический_кластерный_анализ_lecture09_1.pdf

Метод агломерации или метод агрегирования: алгоритм, который позволяет решать, каким образом объединять точки в кластеры на основе выбранной метрики.

Можно выделить следующие основные методы агломерации.

1 *Метод ближнего соседа, он же метод одиночной связи (single linkage).*

Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между ближайшими точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем выбираем минимальное из посчитанных – это и будет расстояние между кластерами A и B.

Особенности. Имеет недостаток: склонен образовывать кластеры, состоящие из одного наблюдения (монокластеры).

2. *Метод дальнего соседа, он же метод полной связи (complete linkage).*

Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между дальними точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем выбираем максимальное из посчитанных – это и будет расстояние между кластерами A и B.

Особенности. Вполне надёжный метод, используется в качестве метода агломерации по умолчанию функцией hclust() в R.

3. Метод средней связи (average linkage).

Реализация. Расстояние между двумя кластерами A и B определяется как среднее расстояние между точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем считаем среднее арифметическое – это и будет расстояние между кластерами A и B.

Особенности. Особых примет нет, тоже вполне надёжный метод.

4. Метод центроидной связи (centroid linkage).

Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между центроидами (центрами тяжести) кластеров. Центроид – средний вектор кластера, его координаты считаются как средние арифметические соответствующих переменных6.

Особенности. Имеет недостаток: может вызывать инверсию–ситуацию, когда на последующем шаге кластеризации объединение в кластеры происходит на расстоянии меньшем, чем на предыдущем шаге. Инверсия противоречит самой логике иерархического кластерного анализа: для минимальной потери информации об исходных наблюдениях (а мы теряем её, переходя к группам), мы должны на каждом шаге кластеризации объединять точки в более крупные кластеры, которые в большей степени удалены друг от друга, а здесь мы «скачем» от большего расстояния к меньшему.

5. Метод Уорда, он же метод Варда (Ward's linkage).

Реализация. На каждом шаге обновления кластеров точка присоединяется к тому кластеру, присоединение к которому приводит к минимально возможному увеличению внутригрупповой дисперсии этого кластера. Внутригрупповую дисперсию можно определить следующим образом:

$$SS = \sum_{i \in G} (x_i - \bar{x}_G)^2,$$

где G – кластер, а $\bar{x}_G$ – средний вектор, центроид этого кластера. Соответственно, на каждом шаге для каждого кластера оценивается текущее значение SS, возможное значение SS в случае добавления точки в кластер, и точка добавляется к тому кластеру, где изменение SS минимально.

Особенности. Считается одним из самых эффективных методов7 агломерации, требует использования только одной метрики – квадрата евклидова расстояния.

15 **Что является визуализатором результатов иерархического кластерного анализа? Выбор числа кластеров.**

Лекция: Иерархический_кластерный_анализ_lecture09_1.pdf

Запустим кластерный анализ и построим *дендрограмму* – график, который визуализирует результаты иерархического кластерного анализа и позволяет увидеть все возможные варианты кластеризации, от наиболее детальной, где много маленьких кластеров, до наиболее общей, где мало больших кластеров.

По горизонтальной оси на дендрограмме отмечаются сами наблюдения, по вертикальной – расстояния между наблюдениями или кластерами на момент объединения их в более крупный кластер.

Пример построения:

На первом шаге у нас 5 кластеров, каждый кластер состоит из одной точки. В иерархическом кластерном анализе всегда первый шаг будет таким, вне зависимости от выбранного способа агломерации. Теперь объединим в кластер те точки, которые ближе всего друг к другу. Это точки A и B, расстояние между ними 2. Соединим эти точки, а на вертикальной оси отметим расстояние 2.

Рис. 2: Дендрограмма: шаг 2

Нужно снова объединить точки в более крупные кластеры. Для этого необходимо определить попарные расстояния между точками C, D и E, а также расстояние между кластером A + B и точкой C, между кластером A + B и точкой D, между кластером A + B и точкой E.

Мы выбрали метод ближнего соседа, поэтому нам надо посчитать расстояние от всех точек кластера A + B до всех точек кластера E и выбрать минимальное. Как можно догадаться, при выборе метода дальнего соседа мы будем брать максимальное значение из посчитанных, а при выборе метода средней связи – среднее расстояние. Проделаем такую операцию для остальных точек и получим матрицу расстояний.

$$D = \begin{bmatrix} & \boldsymbol{A+B} & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{A+B} & 0 & 7.2 & 8.5 & 3.2 \\ \boldsymbol{C} & 7.2 & 0 & 2.2 & 4.2 \\ \boldsymbol{D} & 8.5 & 2.2 & 0 & 5.4 \\ \boldsymbol{E} & 3.2 & 4.2 & 5.4 & 0 \end{bmatrix}$$

Получаем три кластера: A + B, C + D, E. По той же схеме строим новые матрицы расстояний. Осталось объединить всё в один большой кластер. Финальный штрих – соединяем все ветви на расстоянии 4.2 (можете посчитать и проверить самостоятельно). Завершим построение дендрограммы!

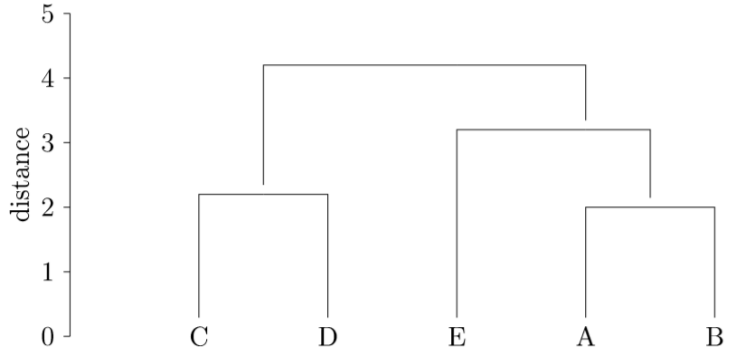


Рис. 5: Дендрограмма: шаг 5

Благодаря дендрограмме мы увидели все возможные варианты деления наших наблюдений на группы. *Но сколько кластеров выбрать?* Ответ на этот вопрос довольно простой, но при этом многозначный:

1. стоит взять столько кластеров, сколько можем содержательно проинтерпретировать;
2. стоит взять столько кластеров, сколько является разумным с точки зрения выраженности межгрупповых различий.

В то время как первый критерий выбора зависит исключительно от наших экспертных, во многом субъективных, знаний, выполнение второго мы можем проверить формально, используя известные статистические методы для сравнения двух и более

групп (критерий Стьюдента для двух выборок, критерий Уилкоксона, ANOVA, критерий Краскелла-Уоллиса и другие).

Наш пример с пятью наблюдениями, конечно, слишком игрушечный, чтобы всерьёз рассуждать, сколько кластеров здесь выбрать, но явно выбор будет стоять между двумя и тремя группами. Если мы хотим получить более общую классификацию и избежать слишком маленьких кластеров из одного человека, то стоит выбрать две группы: C + D и A + B + E. Если для нас выделяющиеся наблюдения играют важную роль, то деление на три кластера, где точка E обособлена, будет в приоритете.

16 **Визуализация инструментов Data Mining**

Лекция: ЛЕКЦИЯ_ВИЗУАЛЬНЫЙ_АНАЛИЗ_ДАННЫХ.pdf

Каждый из алгоритмов Data Mining использует определенный подход к визуализации. В ходе использования каждого из методов, а точнее, его программной реализации, мы получали некие визуализаторы, при помощи которых нам удавалось интерпретировать результаты, полученные в результате работы соответствующих методов и алгоритмов.

• Для деревьев решений это визуализатор дерева решений, список правил, таблица сопряженности.

• Для нейронных сетей в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.

• Для карт Кохонена: карты входов, выходов, другие специфические карты.

• Для линейной регрессии в качестве визуализатора выступает линия регрессии.

• Для кластеризации: дендрограммы, диаграммы рассеивания.

Диаграммы и графики рассеивания часто используются для оценки качества работы того или иного метода. Все эти способы визуального представления или отображения данных могут выполнять одну из функций:

• являются иллюстрацией построения модели (например, представление структуры (графа) нейронной сети);

• помогают интерпретировать полученный результат;

• являются средством оценки качества построенной модели;

• сочетают перечисленные выше функции (дерево решений, дендрограмма).

Визуализация Data Mining моделей

Первая функция (иллюстрация построения модели), по сути, является визуализацией Data Mining модели. Существует много различных способов представления моделей, но графическое ее представление дает пользователю максимальную "ценность".

Таким образом, доступность является одной из основных характеристик модели Data Mining. Несмотря на это, существует и такой распространенный и наиболее простой способ представления модели, как *"черный ящик"*. В этом случае пользователь не понимает поведения той модели, которой пользуется. Однако, несмотря на непонимание, он получает результат - выявленные закономерности. Классическим примером такой модели является *модель нейронной сети*.

Другой способ представления модели - *представление ее в интуитивном, понятном виде.* В этом случае пользователь действительно может понимать то, что происходит "внутри" модели. Такие модели обеспечивают пользователю возможность обсуждать ее логику с коллегами, клиентами и другими пользователями, или объяснять ее.

Понимание модели ведет к пониманию ее содержания. В результате понимания возрастает доверие к модели. Классическим примером является *дерево решений*.

Кроме понимания, такие модели обеспечивают пользователя возможностью взаимодействовать с моделью, задавать ей вопросы и получать ответы. Примером такого взаимодействия является средство "что, если". При помощи диалога "система-пользователь" пользователь может получить понимание модели.

17 **При помощи каких средств визуализации можно оценить качество модели? Какие помогают интерпретировать результат**

Лекция: ЛЕКЦИЯ_ВИЗУАЛЬНЫЙ_АНАЛИЗ_ДАННЫХ.pdf

Теперь перейдем к функциям, которые помогают интерпретировать и оценить результаты построения Data Mining моделей. *Это всевозможные графики, диаграммы, таблицы, списки и т.д.*

Примерами средств визуализации, при помощи которых можно оценить качество модели, являются *диаграмма рассеивания, таблица сопряженности, график изменения величины ошибки.*

Диаграмма рассеивания представляет собой график отклонения значений, прогнозируемых при помощи модели, от реальных. Эти диаграммы используют для непрерывных величин. Визуальная оценка качества построенной модели возможна только по окончанию процесса построения модели.

Таблица сопряженности используется для оценки результатов классификации. Такие таблицы применяются для различных методов классификации. Они уже использовались нами в предыдущих лекциях. Оценка качества построенной модели возможно только по окончанию процесса построения модели.

График изменения величины ошибки. График демонстрирует изменение величины ошибки в процессе работы модели. Например, в процессе работы нейронных сетей пользователь может наблюдать за изменением ошибки на обучающем и тестовом множествах и остановить обучение для недопущения "переобучения" сети. Здесь оценка качества модели и его изменения может оцениваться непосредственно в процессе построения модели.

Примерами средств визуализации, которые помогают интерпретировать результат, являются: *линия тренда в линейной регрессии, карты Кохонена, диаграмма рассеивания в кластерном анализе.*

Из GPT:

Линия тренда в линейной регрессии - это прямая линия, которая лучше всего описывает зависимость между двумя переменными в наборе данных. В линейной регрессии линия тренда используется для прогнозирования значений зависимой переменной на основе значений независимой переменной.

Карты Кохонена (самоорганизующиеся карты) - это нейронные сети, используемые для визуализации и анализа многомерных данных. Карты Кохонена организуют данные в двумерное пространство, что позволяет выявлять скрытые структуры и паттерны. Они часто применяются в задачах кластеризации и снижения размерности данных.

Диаграмма рассеивания в кластерном анализе - это графическое представление данных, где каждая точка соответствует объекту из набора данных, а координаты точки определяются значениями двух переменных. Диаграмма рассеивания используется для визуализации распределения данных и выявления потенциальных кластеров или групп объектов. Она помогает понять, как объекты группируются в пространстве признаков и может быть полезна для оценки результатов кластерного анализа.

18 **Традиционные методы визуального анализа данных**

Лекция: ЛЕКЦИЯ_ВИЗУАЛЬНЫЙ_АНАЛИЗ_ДАННЫХ.pdf

Методы визуализации Методы визуализации, в зависимости от количества используемых измерений, принято классифицировать на две группы:

• представление данных в одном, двух и трех измерениях;

• представление данных в четырех и более измерениях.

Представление данных в одном, двух и трех измерениях

К этой группе методов относятся хорошо известные способы отображения информации, которые доступны для восприятия человеческим воображением. Практически любой современный инструмент Data Mining включает способы визуального представления из этой группы. В соответствии с количеством измерений представления это могут быть следующие способы:

• одномерное (univariate) измерение, или 1-D ;

• двумерное (bivariate) измерение, или 2-D ;

• трехмерное или проекционное (projection) измерение, или 3-D.

Наиболее естественно человеческий глаз воспринимает двухмерные представления информации. При использовании двух- и трехмерного представления информации пользователь имеет возможность увидеть закономерности набора данных:

• его кластерную структуру и распределение объектов на классы (например, на диаграмме рассеивания);

• топологические особенности;

• наличие трендов;

• информацию о взаимном расположении данных;

• существование других зависимостей, присущих исследуемому набору данных.

Если набор данных имеет более трех измерений, то возможны такие варианты:

• использование многомерных методов представления информации (они рассмотрены ниже);

• снижение размерности до одно-, двух- или трехмерного представления.

Существуют различные способы снижения размерности, один из них - факторный анализ (это статистический метод, используемый для выявления скрытых переменных (факторов), которые объясняют наблюдаемые корреляции между множеством исходных переменных).

Для снижения размерности и одновременного визуального представления информации на двумерной карте используются самоорганизующиеся карты Кохонена.

Представление данных в 4 + измерениях

Представления информации в четырехмерном и более измерениях недоступны для человеческого восприятия. Однако разработаны специальные методы для возможности отображения и восприятия человеком такой информации. Наиболее известные способы многомерного представления информации:

• параллельные координаты;

• " лица Чернова ";

• лепестковые диаграммы.

В *параллельных координатах* переменные кодируются по горизонтали, вертикальная линия определяет значение переменной. Пример набора данных, представленного в декартовых координатах и параллельных координатах.

Основная идея представления информации в *"лицах Чернова"* состоит в кодировании значений различных переменных в характеристиках или чертах человеческого лица.

Для каждого наблюдения рисуется отдельное "лицо". На каждом "лице" относительные значения переменных представлены как формы и размеры отдельных черт лица (например, длина и ширина носа, размер глаз, размер зрачка, угол между бровями).

Анализ информации при помощи такого способа отображения основан на способности человека интуитивно находить сходства и различия в чертах лица.

Лепестковые диаграммы - это графическое представление многомерных данных, где каждая ось соответствует одной из переменных, и оси исходят из одной точки, образуя радиальную сетку. Значения переменных соединены линией, создавая форму (лепесток). Диаграммы позволяют сравнивать профиль данных по нескольким параметрам одновременно.

Перед использованием методов визуализации необходимо:

• Проанализировать, следует ли изображать все данные или же какую-то их часть.

• Выбрать размеры, пропорции и масштаб изображения.

• Выбрать метод, который может наиболее ярко отобразить закономерности, присущие набору данных.

Наличие большого количества средств визуализации, представленных в инструменте, который применяет пользователь, может даже вызвать растерянность.

Среди двухмерных и трехмерных средств наиболее широко известны *линейные графики, линейные, столбиковые, круговые секторные и векторные диаграммы.*

При помощи линейного графика можно отобразить тенденцию, передать изменения какого-либо признака во времени. Для сравнения нескольких рядов чисел такие графики наносятся на одни и те же оси координат.

Гистограмму применяют для сравнения значений в течение некоторого периода или же соотношения величин.

Круговые диаграммы используют, если необходимо отобразить соотношение частей и целого, т.е. для анализа состава или структуры явлений. Составные части целого изображаются секторами окружности. Секторы рекомендуют размещать по их величине: вверху - самый крупный, остальные - по движению часовой стрелки в порядке уменьшения их величины. Круговые диаграммы также применяют для отображения результатов факторного анализа, если действия всех факторов являются однонаправленными. При этом каждый фактор отображается в виде одного из секторов круга.

19 Основные тенденции в области визуализации

Лекция: ЛЕКЦИЯ_ВИЗУАЛЬНЫЙ_АНАЛИЗ_ДАННЫХ.pdf

Как уже отмечалось, при помощи средств визуализации поддерживаются важные задачи бизнеса, среди которых - процесс принятия решений. В связи с этим возникает необходимость перехода средств визуализации на более качественный уровень, который характеризуется появлением абсолютно новых средств визуализации и взглядов на ее функции, а также развитием ряда тенденций в этой области.

Среди основных тенденций в области визуализации Филип Рассом (Philip Russom) выделяет:

• Разработка сложных видов диаграмм.

• Повышение уровня взаимодействия с визуализацией пользователя.

• Увеличение размеров и сложности структур данных, представляемых визуализацией.

1. Разработка сложных видов диаграмм.

Большинство визуализаций данных построено на основе диаграмм стандартного типа (секторные диаграммы, графики рассеяния и.т.д.). Эти способы являются одновременно старейшими, наиболее элементарными и распространенными. В последние годы перечень видов диаграмм, поддерживаемых инструментальными средствами визуализации, существенно расширился.

2. Повышение уровня взаимодействия с визуализацией пользователя.

Еще совсем недавно большая часть средств визуализации представляла собой статичные диаграммы, предназначенные исключительно для просмотра. Сейчас широко используются динамические диаграммы, уже сами по себе являющиеся пользовательским интерфейсом, в котором пользователь может напрямую и интерактивно манипулировать визуализацией, подбирая новое представление информации.

Сложное взаимодействие позволяет пользователю изменять визуализацию для нахождения альтернативных интерпретаций данных. Взаимодействие с визуализацией подразумевает минимальный по своей сложности пользовательский интерфейс, в котором пользователь может управлять представлением данных, просто "кликая" на элементы визуализации, перетаскивая и помещая представления объектов данных или выбирая пункты меню. Инструменты OLAP или Data Mining превращают непосредственное взаимодействие с визуализацией в один из этапов итерационного анализа данных. Средства Text Mining или управления документами придают такому непосредственному взаимодействию характер навигационного механизма, помогающего пользователю исследовать библиотеки документов.

Визуальный запрос является наиболее современной формой *сложного взаимодействия* пользователя с данными. Пользователь использует информационные точки графика рассеяния, выбирать их мышкой и получать новые визуализации, представляющие именно эти точки. Приложение визуализации данных генерирует соответствующий язык запроса, управляет принятием запроса базой данных и визуально представляет результирующее множество. Пользователь может сфокусироваться на анализе, не отвлекаясь на составление запроса. Увеличение размеров и сложности структур данных, представляемых визуализацией.

Визуализация поддерживает обработку структурированных данных, она также является ключевым средством представления схем так называемых неструктурированных данных, например текстовых документов, т.е. Text Mining. В частности, средства Text Mining могут осуществлять парсинг больших пакетов документов и формировать предметные указатели понятий и тем, освещенных в этих документах. Когда предметные указатели созданы с помощью нейросетевой технологии, пользователю непросто продемонстрировать их без некоторой формы визуализации данных. Визуализация в таком случае преследует две цели:

- визуальное представление контента библиотеки документов;
- навигационный механизм, который пользователь может применять при исследовании документов и их тем.

Визуальный анализ данных обычно выполняется в три этапа:

• беглый анализ - позволяет идентифицировать интересные шаблоны и сфокусироваться на одном или нескольких из них;

• увеличение и фильтрация - идентифицированные на предыдущем этапе шаблоны отфильтровываются и рассматриваются в большем масштабе;

• детализация по необходимости - если пользователю нужно получить дополнительную информацию, он может визуализировать более детальные данные.

20. Что представляет из себя логит? С какими понятиями он связан.

Логит тесно связан с понятиями «вероятность» и «шанс». Вероятность – это объективная мера появления некоторого события, из-
меряемая от 0 до 1. На практике оценкой вероятности служит относительная частота появления события. Значение вероятности 0 означает невозмож- ность появления события. Значение вероятности 1 означает, что событие непременно произойдет.

Вероятность (probability): Y_i

Шансы – это отношение вероятности того, что событие произойдет, к вероят- ности того, что событие не произойдет. Можно еще сказать так: шансы – это отношение вероятности наступления события к вероятности ненаступления события. Вероятность наступления события часто называют просто веро- ятностью события, и когда вы встречаете фразы «вычислить вероятность», «оценить влияние предикторов на вероятности» в контексте логистической регрессии, то речь идет именно о вероятности события. С ростом вероятно-сти растут шансы, и наоборот. Значение шансов 1 соответствует ситуации, когда вероятности наступления события и ненаступления события равны.

Шансы (odds): $\frac{P_i}{1-P_i}$

Наконец, логит – это натуральный логарифм шансов.

Логит (logit), логарифм шансов (log odds), прологарифмированные шансы (logged odds):

$$\ln\left(\frac{P_i}{1-P_i}\right)$$

Поупражняемся вычислять шансы и логиты. Например, если P_i для первого наблюдения равно 0,2, то шансы равны 0,25,

или 0,2/0,8, а логит равен –1,386, т. е. натуральному логарифму шансов. Если P_i для второго наблюдения равно 0,7, то шансы равны 2,33, или 0,7/0,3, а логит равен 0,847. Если P_i для третьего наблюдения равно 0,9, то шансы равны 9, или 0,9/0,1, а логит равен 2,197.

Хотя формула преобразования вероятностей в логиты проста, требуется некоторое объяснение, чтобы проиллюстрировать ее полезность. Оказыва- ется, она прекрасно описывает зависимость между предикторами и распре- делением вероятностей, определяемым бинарной зависимой переменной. Формула включает два шага: на первом шаге мы берем отношение вероят- ности, что событие произойдет, к вероятности, что событие не произойдет, $\frac{P_i}{1-P_i}$, и получаем шансы возникновения события; на втором шаге берем натуральный логарифм шансов и получаем логит.

21. Смысл и свойства логита.

Использование натурального логарифма шансов исключает минимальное значение 0 («пол») так же, как преобразование вероятностей в шансы ис- ключает максимальное значение 1 («потолок»). Когда мы берем:

- натуральный логарифм шансов выше 0, но ниже 1, мы получаем от- рицательные числа;
- натуральный логарифм шансов, равный 1, мы получаем 0;
- натуральный логарифм шансов выше 1, мы получаем положительные числа.

Напомним, что логарифм 0 и отрицательных чисел не существует.

Таким образом, первое свойство логита состоит в том, что, в отличие от вероятности, он не имеет верхней или нижней границы. Шансы устраняют верхнюю границу вероятностей, а прологарифмированные шансы устраняют нижнюю границу вероятностей. Давайте убедимся в этом. Если P_i = 1, логит не определен, потому что шансы 1/0 не существуют. По мере того как вероят- ность приближается к 1, логит движется к +¥. Если P_i = 0, логит не опреде- лен, потому что логарифм шансов 0/1 или 0 не существует. По мере того как вероятность приближается к 0, логит движется к –¥. Таким образом, логит варьирует от –¥ до +¥. Проблема нижней и верхней границ для вероятно- стей (или нижней границы для шансов) отпадает.

Второе свойство логита заключается в том, что логит-преобразование симметрично относительно вероятности 0,5. Когда P_i = 0,5, логит равен 0 (0,5/0,5 = 1, а логарифм 1 равно 0). Вероятности ниже 0,5 (P_i меньше 1 – P_i) приведут к отрицательным логитам, потому что шансы падают ниже 1, но выше 0. А из курса школьной математики мы помним, что логарифм чисел выше 0, но ниже 1 дает отрицательное число. Вероятности выше 0,5 (P_i выше 1 – P_i) приведут к положительным логитам, потому что шансы превышают 1. Опять же из курса школьной математики помним, что логарифм чисел выше 1 дает положительное число.

Кроме того, вероятности, которые находятся на одинаковом расстоянии от 0,5 выше или ниже (например, 0,6 и 0,4, 0,7 и 0,3, 0,8 и 0,2), имеют одина- ковые логиты, но с разными знаками (например, логиты для вероятностей, перечисленных выше, равны 0,405 и –0,405, –0,847 и –0,847, 1,386 и –1,386). Удаленность логита от 0 отражает удаленность вероятности от 0,5 (опять же отметим, что логиты не имеют границ).

Третье свойство логита заключается в том, что одно и то же изменение вероятности приводит к различным изменениям в логитах. Простой прин- цип заключается в том, что по мере приближения P_i к 0 и 1 одно и то же изменение вероятности приводит к большему изменению логита. Вы можете увидеть это на примере:

P_i	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$1 - P_i$	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1
Шансы	0,111	0,25	0,429	0,667	1	1,5	2,33	4	9
Логит	−2,2	−1,39	−0,847	−0,405	0	0,405	0,847	1,39	2,2
Δ логит	0,81	0,543	0,442	0,405	0,405	0,442	0,543	0,81	

Изменение вероятности на 0,1 с 0,5 до 0,6 (или с 0,5 до 0,4) приводит к из- менению логита на 0,405, тогда как такое же изменение вероятности на 0,1 с 0,8 до 0,9 (или с 0,2 до 0,1) приводит к изменению логита на 0,81.

Для одного и того же изменения вероятности изменение логита в край- них значениях вероятности будет в два раза больше изменения логита для среднего значения вероятности. Повторим, что общий принцип заключается в том, что небольшое изменение вероятности приводит к большому изменению логита, когда вероятности находятся вблизи границ 0 и 1.

22. Линейная модель множественной регрессии, спецификация модели. Требования к факторам. Коэффициент корреляции.

Парная регрессия может дать хороший результат при моделировании, если влиянием других факторов, воздействующих на объект исследования, можно пренебречь. Если же этим влиянием пренебречь нельзя, то следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии:

$$y = f\left(x_1, x_2, \ldots, x_m\right),$$

где y – зависимая переменная (результативный признак), i x – независимые, или объясняющие, переменные (признаки-факторы). Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и целом ряде других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

В линейной множественной регрессии

$$\tilde{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_p \cdot x_p \qquad (1)$$

параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего параметра на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Он включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

Включение в уравнение множественной регрессии того или иного набора факторов связано прежде всего с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.

2. Факторы не должны быть интеркоррелированы и тем более находиться в точной функциональной связи. Включение в модель факторов с высокой интеркорреляцией, может привести к нежелательным последствиям – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии. Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

Корреляция – вероятностная связь, взаимозависимость случайных величин.
• **Коэффициент корреляции** — это показатель степени связи между двумя переменными или измерениями.

23. Показатели качества множественной регрессии (оценка значимости коэффициентов регрессии, F-критерий, Коэффициент детерминации, Частные коэффициенты корреляции.

Коэффициент детерминации – квадрат коэффициента или индекса корреляции.

Для оценки качества построенной модели регрессии можно использовать показатель (коэффициент, индекс) детерминации R^2 либо среднюю ошибку аппроксимации.

Чем выше показатель детерминации или чем ниже средняя ошибка аппроксимации, тем лучше модель описывает исходные данные.

Средняя ошибка аппроксимации – среднее относительное отклонение расчетных значений от фактических

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\,\%.$$

Построенное уравнение регрессии считается удовлетворительным, если значение A не превышает 10–12 %.

Оценка значимости всего уравнения регрессии в целом осуществляется с помощью *F*-критерия Фишера.

***F*-критерий Фишера** заключается в проверке гипотезы H_o о статистической незначимости уравнения регрессии. Для этого выполняется сравнение фактического $F_{факт}$ и критического (табличного) $F_{табл.}$ значений *F*-критерия Фишера.

$F_{факт}$ определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы

$$F_{факт} = \frac{\sum \frac{(\hat{y} - \bar{y})^2}{m}}{\sum \frac{(y - \hat{y})^2}{n - m - 1}} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot \frac{n - m - 1}{m}, \qquad (1.5)$$

где n – число единиц совокупности; m – число параметров при переменных. Для линейной регрессии $m = 1$.

Для нелинейной регрессии вместо r^2_{xy} используется R^2.

$F_{табл}$ – максимально возможное значение критерия под влиянием случайных факторов при степенях свободы $k_1 = m$, $k_2 = n - m - 1$ (для линейной регрессии $m = 1$) и уровне значимости α.

Оценка значимости коэффициентов регрессии позволяет определить, насколько каждая независимая переменная влияет на зависимую переменную. Это важно для понимания структуры модели и выявления наиболее значимых факторов. Основным инструментом для оценки значимости коэффициентов регрессии является t-критерий.

Частный коэффициент корреляции – это показатель, измеряющий степень сопряженности двух признаков при постоянном значении третьего.

Математическая статистика позволяет установить корреляцию между двумя признаками при постоянном значении третьего, не ставя специального эксперимента, а используя парные коэффициенты корреляции r_{xy}, r_{xz} и r_{yz}. Частные коэффициенты корреляции рассчитывают по формулам:

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xy}^2) \cdot (1 - r_{yz}^2)}} \text{ ; (12.1)}$$

$$r_{xz\cdot y} = \frac{r_{xz} - r_{xy} \cdot r_{zy}}{\sqrt{(1 - r_{xy}^2) \cdot (1 - r_{zy}^2)}} \text{ ; (12.2)}$$

$$r_{yz\cdot x} = \frac{r_{yz} - r_{xy} \cdot r_{xz}}{\sqrt{(1 - r_{xy}^2) \cdot (1 - r_{xz}^2)}} \text{ . (12.3)}$$

Здесь в индексах буквы перед тире указывают, между какими признаками изучается зависимость, а буква после тире – влияние какого признака исключается (элиминируется). Ошибку и критерий значимости частной корреляции определяют по тем же формулам, что и парной корреляции (11.8):

$$s_{r_{xy\cdot z}} = \sqrt{\frac{1 - r_{xy\cdot z}^2}{n - 2}} \text{ ; (12.4)}$$

$$t = \frac{r}{s_r} \text{ (12.5)}$$

Теоретические значения t берут из таблицы приложения критерия Стьюдента для принятого уровня значимости и *n–3* степеней свободы.

Подобно парным коэффициентам корреляции частные коэффициенты могут принимать значения, заключенные между –1 и +1. Частные коэффициенты детерминации находят путем возведения в квадрат частных коэффициентов корреляции.

24. Этапы кластерного анализа. Многомерные данные.

Кластерный анализ предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

2.1. Этапы кластерного анализа

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

– иерархические;
– неиерархические.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Многомерные данные представляют собой набор данных, в котором каждый объект описывается несколькими признаками (переменными). Важно учитывать корреляции между признаками и возможную избыточность данных.

25. Цели и целевые функции. Значимость переменных объекта.

Цели и целевые функции — это ключевые элементы при построении и оценке моделей в анализе данных. Они определяют, что модель должна достигнуть и каким образом будет измеряться её эффективность.

Цели анализа данных варьируются в зависимости от конкретной задачи и области применения, но могут включать:

- Предсказание

Построение модели для прогнозирования будущих значений зависимой переменной на основе новых данных.

- Классификация

Разделение данных на категории или классы.

- Кластеризация

Группировка объектов в кластеры на основе их сходства.

Целевая функция (или функция потерь) — это математическое выражение, которое оценивает качество модели. Она используется для обучения модели и минимизации ошибок. Выбор целевой функции зависит от типа задачи:

Регрессия:

Основные целевые функции включают среднеквадратичную ошибку (MSE), среднюю абсолютную ошибку (MAE).

Классификация:

Используются логистическая регрессия, перекрестная энтропия и AUC-ROC.

Значимость переменных (факторов) — это оценка вклада каждой независимой переменной в объяснение вариации зависимой переменной. Определение значимости переменных позволяет выявить ключевые факторы, влияющие на целевую переменную, и упростить модель, исключив незначимые переменные.

26. Этапы кластерного анализа. Данные о близости.

2.1. Этапы кластерного анализа

Данные о близости (или данные о расстояниях) являются ключевым компонентом кластерного анализа. Они отражают степень сходства или различия между объектами в многомерном пространстве. Рассмотрим основные аспекты данных о близости:
Метрики расстояний используются для количественной оценки близости между объектами. Выбор метрики зависит от типа данных и специфики задачи.

1. Евклидово расстояние, также обозначается L2, одно из самых распространённых расстояний[3]:
$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{jp})^2}$$
2. Квадрат евклидова расстояния, также обозначается L2 squared, необходимо для некоторых методов агломерации, в частности, для метода Уорда (Варда):
$$d(x_i, x_j) = \sum_{p=1}^{P} (x_{ip} - x_{jp})^2.$$
3. Манхэттенское расстояние, оно же блочное расстояние, также обозначается L1:
$$d(x_i, x_j) = \sum_{p=1}^{P} |x_{ip} - x_{jp}|.$$
4. Расстояние Чебышёва:
$$d(x_i, x_j) = \max\{|x_{ip} - x_{jp}|\}.$$

Для того, чтобы реализовать иерархический кластерный анализ, нам необходимо определиться с метрикой и получить матрицу расстояний – матрицу, состоящую из расстояний между всеми парами точек. Как можно догадаться, эта матрица будет квадратной, симметричной, а на главной диагонали будут находиться 0 (вспомним свойства метрики).

27 Переменные объекта и факторы.

Наблюдаемая переменная — переменная, которую зафиксировали в явном виде.

Скрытая переменная — переменная, которую вывели через математические модели с использованием наблюдаемых переменных.

Факторный анализ – процедура, с помощью которой большое число переменных, относящихся к имеющимся наблюдениям, сводят к меньшему количеству независимых влияющих величин, называемых факторами:

- в один фактор объединяются переменные, сильно коррелирующие между собой
- переменные из разных факторов слабо коррелируют между собой.

Факторный анализ классифицирует признаки (переменные), описывающие наблюдения.

Фактор – латентная (скрытая) переменная, конструируемая таким образом, чтобы можно было объяснить корреляцию между набором имеющихся переменных. Концепция факторного анализа заключается в сжатии информации.

28 Проверка достоверности результатов.

Проверить достоверность результатов можно с помощью метрик recall, presision, accuracy, матрицы ошибок, критерия Фишера, критерия Стьюдента, ROC-кривой.

recall – полнота модели - доля истинно положительных, среди всех действительно положительных

presision – точность, доля истинно положительных, среди всех, которые модель классифицировала как положительные

accuracy - насколько правильно модель классифицировала данные. Это отношение числа правильно классифицированных экземпляров к общему числу экземпляров.

ROC-кривая – используется для оценки качества классификационной модели. Она показывает, насколько хорошо модель различает два класса, и помогает выбрать оптимальный порог вероятности для классификации.

Матрица ошибок – нужна для оценки точности в задачах классификации. Это таблица с 4 различными комбинациями прогнозируемых и фактических значений. Прогнозируемые значения описываются как положительные и отрицательные, а фактические – как истинные и ложные.

Критерий Фишера — статистический критерий для оценки значимости различия дисперсий двух случайных выборок. С помощью Критерия Фишера оценивают качество регрессионной модели. Оценивание качества уравнения регрессии - состоит в проверке гипотезы Но о статистической незначимости уравнения регрессии и показателя тесноты связи. Для этого выполняется сравнение фактического Fфакт и критического (табличного) Fтабл значений F-критерия Фишера. Fфакт определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы. Fтабл - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости а. Уровень значимости а - вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно а принимается равной 0,05 или 0,01. Если Fтабл < Fфакт, то Но - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если Fтабл > Fфакт, то гипотеза Но не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

Критерий Стьюдента – Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются t-критерий Стьюдента и доверительные интервалы каждого из показателей. Выдвигается гипотеза Но о случайной природе показателей, т.е. о незначимом их отличии от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью t-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки. Сравнивая фактическое и критическое (табличное) значения t-статистики - $t_{табл}$ и $t_{факт}$ - принимаем или отвергаем гипотезу H_o.

29 Регрессионный анализ. Основная идея. Линейность и нелинейность по параметрам и факторам. Линеаризация.

Регрессионный анализ - это статистический метод выявления и количественной оценки связи между зависимой переменной и одной или несколькими независимыми переменными. Уравнение регрессии показывает, как в среднем изменяется y при изменении любого из xi.

Если независимая переменная одна - это простой регрессионный анализ. Если же их несколько, то такой анализ называется многофакторным.

В ходе регрессионного анализа решаются две основные задачи:

- построение уравнения регрессии, т.е. нахождение вида зависимости между результатным показателем и независимыми факторами x1, x2, …, xn.
- оценка значимости полученного уравнения, т.е. определение того, насколько выбранные факторные признаки объясняют вариацию признака y.

Нелинейная по переменным модель - линейная модель y=f(x), в которой возможна замена переменной z=g(x), приводящая получившуюся модель y = F(z) - к линейной.

Нелинейная по параметрам модель - модель, которую нельзя привести заменами переменных к линейной.

Линейная регрессия: $y = a + bx + \varepsilon$

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам.

Регрессии, нелинейные по объясняющим переменным:

- полиномы разных степеней $y=a+b_1 \cdot x+b_2 \cdot x^2+b_3 \cdot x^3+\varepsilon$
- равносторонняя гипербола $y = a + \frac{b}{x} + \varepsilon$.

Регрессии, нелинейные по оцениваемым параметрам:

- степенная $y=a \cdot x^b \cdot \varepsilon$
- показательная $y=a \cdot b^x \cdot \varepsilon$
- экспоненциальная $y=e^{a+b \cdot x} \cdot \varepsilon$

Линеаризация – подбор преобразований к анализируемым переменным, которые позволили бы представить искомую зависимость в виде линейного соотношения между преобразованными переменными.

30 Переменные объекта, их значения и взаимосвязи.

Переменные объекта в анализе данных - это характеристики или атрибуты, которые мы изучаем. Они могут быть различными и зависят от конкретного исследования. Например, если мы анализируем данные о погоде, переменными объекта могут быть температура, влажность, скорость ветра и так далее.

Значения переменных - это конкретные измерения или наблюдения, которые мы собираем для каждой переменной. Например, если температура является одной из наших переменных, то конкретные температурные показатели, которые мы записываем, будут являться значениями этой переменной.

Взаимосвязи между переменными - это ключевой аспект анализа данных. Мы хотим понять, как одна переменная влияет на другую. Например, мы можем исследовать, как температура влияет на влажность. Это может быть выражено с помощью статистических методов, таких как корреляционный анализ или регрессионный анализ.

В контексте многомерного анализа, мы можем изучать взаимосвязи между несколькими переменными одновременно. Это может помочь нам понять сложные взаимодействия и закономерности в данных.

Важно отметить, что анализ данных - это итеративный процесс. Мы можем начать с определенного набора переменных и гипотез, а затем по мере получения новых данных и информации мы можем обновлять наши переменные и гипотезы. Это помогает нам постоянно улучшать наше понимание данных и делать более точные прогнозы и выводы.

31 Метод DBSCAN.

DBSCAN – алгоритм кластеризации, развивает идею кластеризации с помощью выделения связных компонент.

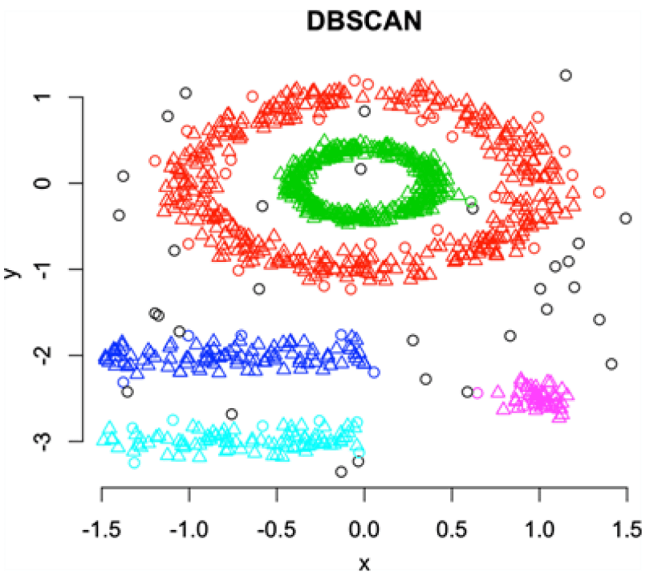
Плотность в DBSCAN определяется в окрестности каждого объекта выборки xi как количество других точек выборки в шаре $B(\varepsilon,xi)$. Кроме радиуса ε окрестности в качестве гиперпараметра алгоритма задается порог $N0$ по количеству точек в окрестности.

Далее все объекты выборки делятся на три типа: **внутренние / основные точки (core points)**, **граничные (border points)** и **шумовые точки (noise points)**. К основным относятся точки, в окрестности которых больше $N0$ объектов выборки. К граничным — точки, в окрестности которых есть основные, но общее количество точек в окрестности меньше $N0$. Шумовыми называют точки, в окрестности которых нет основных точек и в целом содержится менее $N0$ объектов выборки.

Алгоритм кластеризации с помощью DBSCAN выглядит следующим образом:

1. Шумовые точки убираются из рассмотрения и не приписываются ни к какому кластеру.
2. Основные точки, у которых есть общая окрестность, соединяются ребром.
3. В полученном графе выделяются компоненты связности.
4. Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней основная точка.

Удобство DBSCAN заключается в том, что он сам определяет количество кластеров (по модулю задания других гиперпараметров — ε и $N0$), а также в том, что метод успешно справляется даже с достаточно сложными формами кластеров. Кластеры могут иметь вид протяжённых лент или быть вложенными друг в друга как концентрические гиперсферы. На изображении ниже показан пример выделения кластеров достаточно сложной формы с помощью DBSCAN:



DBSCAN — один из самых сильных алгоритмов кластеризации, но работает он, как правило, заметно долго, к тому же весьма чувствителен к размерности пространства признаков, поэтому используется на практике DBSCAN только тогда, когда успевает отрабатывать за приемлемое время.

32 Общие предположения о выборке и процедуре проведения эксперимента

Последовательность n значений $x_1, x_2, ..., x_n$, полученных в результате наблюдения некоторого процесса, мы будем рассматривать как совокупность значений одинаково распределенных независимых случайных величин $\xi_1, \xi_2, ..., \xi_n$, представляющих собой n экземпляров одной и той же случайной величины ξ. Эта последовательность значений называется ***выборкой.*** В этом случае говорят, что выборка взята из ***генеральной совокупности*** случайной величины ξ. Если величина ξ следует закону распределения $F(x)$, то мы будем говорить, что генеральная совокупность распределена по закону $F(x)$.

Пусть x1,x2,..,xn − выборка объема n из некоторой генеральной совокупности. По этой выборке можно оценить основные числовые характеристики генеральной совокупности. Различные элементы выборки i x называются вариантами.
Выборка называется случайной, если из генеральной совокупности элементы берутся наугад и в выборку каждый из них может попасть с одинаковой вероятностью.
Если случайная выборка такова, что по её распределению по некоторому признаку можно судить о распределении поэтому же признаку неизвестной генеральной совокупности, то такая выборка называется репрезентативной, т.е. хорошо представляющей генеральную совокупность.
Статистическая гипотеза — это определённое предположение о распределении вероятностей, лежащем в основе наблюдаемой выборки данных.
Проверка статистической гипотезы — это процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза наблюдаемой выборке данных.
В анализе больших данных эксперимент проводится, чтобы подтвердить или опровергнуть гипотезы.
Процедура проведения эксперимента:

Пусть в (статистическом) эксперименте доступна наблюдению случайная величина X, распределение которой $\mathbb{P}$ полностью или частично неизвестно. Тогда любое утверждение относительно $\mathbb{P}$ называется **статистической гипотезой**. Гипотезы различают по виду предположений, содержащихся в них:

- Статистическая гипотеза, однозначно определяющая распределение $\mathbb{P}$, то есть $H: \{\mathbb{P} = \mathbb{P}_0\}$, где $\mathbb{P}_0$ — какой-то конкретный закон, называется **простой**.
- Статистическая гипотеза, утверждающая принадлежность распределения $\mathbb{P}$ к некоторому семейству распределений, то есть вида $H: \{\mathbb{P} \in \mathcal{P}\}$, где $\mathcal{P}$ — семейство распределений, называется **сложной**.

На практике обычно требуется проверить какую-то конкретную и, как правило, простую гипотезу H_0. Такую гипотезу принято называть *нулевой*. При этом параллельно рассматривается противоречащая ей гипотеза H_1, называемая *конкурирующей* или *альтернативной*.

Выдвинутая гипотеза нуждается в проверке, которая осуществляется статистическими методами, поэтому гипотезу называют статистической. Для проверки гипотезы используют критерии, позволяющие принять или опровергнуть гипотезу.

Этапы проверки статистических гипотез [править | править код]

1. Формулировка основной гипотезы H_0 и конкурирующей гипотезы H_1.
2. Задание уровня значимости α, на котором в дальнейшем и будет сделан вывод о справедливости гипотезы. Он равен вероятности допустить ошибку первого рода.
3. Расчёт статистики ϕ критерия такой, что:
   - её величина зависит от исходной выборки $\mathbf{X} = (X_1, \ldots, X_n)$: $\phi = \phi(X_1, \ldots, X_n)$;
   - по её значению можно делать выводы об истинности гипотезы H_0;
   - статистика ϕ, как функция случайной величины $\mathbf{X}$, также является случайной величиной и подчиняется какому-то закону распределения.
4. Построение критической области. Из области значений ϕ выделяется подмножество $\mathcal{C}$ таких значений, по которым можно судить о существенных расхождениях с предположением. Его размер выбирается таким образом, чтобы выполнялось равенство $P(\phi \in \mathcal{C}) = \alpha$. Это множество $\mathcal{C}$ и называется **критической областью**.
5. Вывод об истинности гипотезы. Наблюдаемые значения выборки подставляются в статистику ϕ и по попаданию (или непопаданию) в критическую область $\mathcal{C}$ выносится решение об отвержении (или не отвержении) выдвинутой гипотезы H_0.

Более простое объяснение:
Основные этапы проверки статистических гипотез

1. Исходя из задач исследования, формулируются статистические гипотезы
2. Выбирается уровень значимости, на котором будут проверяться гипотезы
3. На основе выборки, полученной из результатов измерения, определяется статистическая характеристика гипотезы
4. Выбирается критерий для проверки статистической гипотезы
5. Вычисляется наблюдаемое (фактическое) значение статистического критерия
6. Определяется критическое значение статистического критерия по соответствующей таблице на основании выбранного уровня значимости и объема выборки
7. На основе сравнения наблюдаемого и критического значения критерия в зависимости от результатов проверки нулевая гипотеза либо принимается, либо отклоняется в пользу альтернативной.

33. Типы классификаций. Кластер. Типы кластерных структур.

Классификация - системное распределение изучаемых предметов, явлений, процессов по родам, видам, типам, по каким-либо существенным признакам для удобства их исследования; группировка исходных понятий и расположение их в определенном порядке, отражающем степень этого сходства.

Классификация - упорядоченное по некоторому принципу множество объектов, которые имеют сходные классификационные признаки (одно или несколько свойств), выбранных для определения сходства или различия между этими объектами.

Классификация относится к **стратегии обучения с учителем** (supervised learning), которое также именуют контролируемым или управляемым обучением.

Классификация может быть **одномерной** (по одному признаку) и **многомерной** (по двум и более признакам).

Различают:

- **вспомогательную** (искусственную) классификацию, которая производится по внешнему признаку и служит для придания множеству предметов (процессов, явлений) нужного порядка;
- **естественную** классификацию, которая производится по существенным признакам, характеризующим внутреннюю общностьпредметов и явлений. Она является результатом и важным средством научного исследования, т.к. предполагает и закрепляет результаты изучения закономерностей классифицируемых объектов.

В зависимости от выбранных признаков, их сочетания и процедуры деления понятий **классификация может быть**:

- **простой** - деление родового понятия только по признаку и только один раз до раскрытия всех видов. Примером такой классификации является дихотомия, при которой членами деления бывают только два понятия, каждое из которых является противоречащим другому (т.е. соблюдается принцип: "А и не А");
- **сложной** - применяется для деления одного понятия по разным основаниям и

синтеза таких простых делений в единое целое. Примером такой классификации является периодическая система химических элементов.

Под классификацией будем понимать **отнесение объектов (наблюдений, событий) к одному из заранее известных классов.**

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

Само понятие "**кластер**" определено неоднозначно: в каждом исследовании свои "кластеры". Переводится понятие кластер (cluster) как **"скопление", "гроздь".**

Кластер можно охарактеризовать как **группу объектов, имеющих общие свойства**. Характеристиками кластера можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

Приведем краткую характеристику подходов к кластеризации.

- **Алгоритмы, основанные на разделении данных** (Partitioning algorithms), в т.ч. итеративные:
  - разделение объектов на k кластеров;
  - итеративное перераспределение объектов для улучшения кластеризации.
- **Иерархические алгоритмы** (Hierarchy algorithms):
  - агломерация: каждый объект первоначально является кластером, кластеры, соединяясь друг с другом, формируют больший кластер и т.д.
- **Методы, основанные на концентрации объектов** (Density-based methods):
  - основаны на возможности соединения объектов;
  - игнорируют шумы, нахождение кластеров произвольной формы.
- **Грид-методы** (Grid-based methods):
  - квантование объектов в грид-структуры.
- **Модельные методы** (Model-based):
  - использование модели для нахождения кластеров, наиболее соответствующих данным.

34. Представление исходных данных в регрессионном анализе.

1. Сбор данных, которые должны быть релевантны исследуемому вопросу. Данные могут быть собраны из различных источников, включая эксперименты, опросы.

2. Исходные данные состоят из двух основных типов переменных:

- **Зависимая переменная (целевая переменная) –** переменная, значение которой мы пытаемся предсказать или объяснить через модель;
- **Независимые переменные (предикторы) –** переменные, которые используются для предсказания значения зависимой переменной.

3. Формат данных

Данные должны быть организованы в формате, пригодном для анализа. Обычно это таблица, где строки представляют наблюдения, а столбцы — переменные.

4. Предварительная обработка данных

Перед проведением анализа данные часто требуют предварительной обработки:

- **Очистка данных. У**даление или корректировка неправильных, пропущенных или аномальных значений;
- **Преобразование данных**. Нормализация данных для улучшения их распределения или масштабирования.

5. Набор исходных данных (или выборку данных) разбивают на два множества:

- Обучающее множество (training set) - множество, которое включает данные, использующиеся для обучения (конструирования) модели. Такое множество содержит входные и выходные (целевые) значения примеров. Выходныезначения предназначены для обучения модели.
- Тестовое (test set) множество также содержит входные и выходные значения примеров. Здесь выходные значения используются для проверки работоспособности модели.

35. Задачи классификации. Общая постановка.

Задачей классификации часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Например, можно предсказать, кто из клиентов фирмы является потенциальным покупателем определенного товара, а кто - нет, кто воспользуется услугой фирмы, а кто - нет, и т.д. Этот тип задач относится к задачам бинарной классификации, в них зависимая переменная может принимать только два значения (например, да или нет, 0 или 1).

Другой вариант классификации возникает, если зависимая переменная может принимать значения из некоторого множества предопределенных классов. **Например**, когда необходимо предсказать, какую марку автомобиля захочет купить клиент. В этих случаях рассматривается множество классов для зависимой переменной.

36. Причины неадекватности уравнения регрессии. Классическая схема линейного регрессионного анализа.

Причины неадекватности уравнения регрессии

Метод наименьших квадратов (МНК) — это стандартный статистический метод для оценки параметров в линейной регрессии. Этот метод стремится минимизировать сумму квадратов разностей между наблюдаемыми значениями зависимой переменной и значениями, предсказанными моделью регрессии. Для адекватной работы этого метода и обеспечения точности, надежности и статистической значимости результатов регрессионного анализа должны выполняться определенные предпосылки.

Предпосылки МНК.

1. Математическое ожидание случайного отклонения ε_i равно 0 для всех наблюдений ($M(\varepsilon_i) = 0$).
2. Гомоскедастичность (постоянство дисперсий отклонений). Дисперсия случайных отклонений ε_i постоянна: $D(\varepsilon_i) = D(\varepsilon_j) = S^2$ для любых i и j.
3. отсутствие автокорреляции.
4. Случайное отклонение должно быть независимо от объясняющих переменных: $Y_{eixi} = 0$.
5. Модель является линейное относительно параметров.
6. отсутствие мультиколлинеарности. Между объясняющими переменными отсутствует строгая (сильная) линейная зависимость.
7. Ошибки ε_i имеют нормальное распределение. Выполнимость данной предпосылки важна для проверки статистических гипотез и построения доверительных интервалов.

Теорема Гаусса - Маркова

Теорема. Если предпосылки 1 – 5 выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными. Это говорит об отсутствии систематической ошибки при определении положения линии регрессии.

2. Оценки состоятельны. Это означает, что с ростомnнадежность оценок возрастает.

3. Оценки эффективны, т.е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин yi.

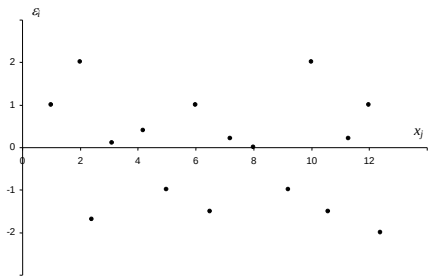


Рис.- Зависимость случайных остатков от величины фактора x_j

Если расположение остатков на графике не имеет направленности, то они независимы от значений x_j (см. рис). Если же график показывает наличие зависимости ε_i и x_j, то модель **неадекватна**. Причины неадекватности могут быть разные. Возможно, нарушена третья предпосылка МНК и дисперсия остатков непостоянна для каждого значения фактора x_j. Может быть неправильной спецификация модели, и в нее необходимо ввести дополнительные члены от x_j, например, x_j^2, или преобразовать значения y. Скопление точек в определенных участках значений фактора x_i говорит о наличии систематической погрешности модели.

Классическая схема линейного регрессионного анализа

1. Определение модели

Первый шаг — это формулировка статистической модели. В линейной регрессии это обычно выражается уравнением:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

где:

- Y — зависимая переменная;
- $X_1, X_2, \ldots, X_k$ — независимые переменные;
- $\beta_0, \beta_1, \ldots, \beta_k$ — параметры модели, которые нужно оценить;
- ϵ — случайная ошибка модели.

2. Сбор данных

3. Оценка параметров, используя метод наименьших квадратов (МНК)

4. Проверка адекватности модели

После оценки параметров проводится анализ остатков (разниц между наблюдаемыми и модельными значениями), чтобы проверить, выполняются ли предпосылки МНК (гомоскедастичность, нормальность распределения ошибок, отсутствие автокорреляции). Также проводятся различные статистические тесты, включая тесты на значимость коэффициентов (t-тесты) и на адекватность модели в целом (F-тест).

5. Интерпретация результатов

На этом этапе анализируются и интерпретируются полученные коэффициенты. Оценивается, как изменение независимых переменных влияет на зависимую переменную.

6. Применение модели

Финальный шаг — использование регрессионной модели для прогнозирования значений зависимой переменной или для понимания взаимосвязей между переменными.

37. МНК. Основная идея. Диаграмма рассеяния.

Метод наименьших квадратов (МНК) — математический метод, применяемый для решения различных задач. Он основан на минимизации суммы квадратов отклонений некоторых функций от экспериментальных входных данных.

Примеры применения МНК:

Решение переопределённых систем уравнений (когда количество уравнений превышает количество неизвестных).

Поиск решения в случае обычных (не переопределённых) нелинейных систем уравнений.

Аппроксимация точечных значений некоторой функции.

МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

Из лекции:

Диаграмма рассеяния, представляющая собой зависимость между непрерывной независимой переменной и бинарной зависимой переменной

Из интернета: **Диаграмма рассеяния** (также точечная диаграмма, англ. scatter plot) — математическая диаграмма, изображающая значения двух переменных в виде точек на декартовой плоскости. Могут использоваться и полярные координаты, особенно в случаях, когда одна из переменных представляет собой физическое направление или имеет циклический характер.

Для наглядности построим **диаграмму рассеивания** на основе имеющихся данных.

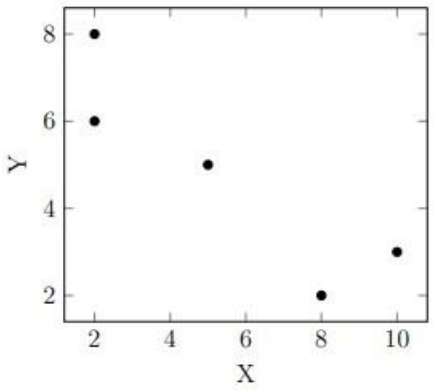


Рис. 1: Диаграмма рассеивания X vs Y

38. Меры типа расстояния.

Из лекции:

1. Евклидово расстояние, также обозначается L2, одно из самых распространённых расстояний[3]:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{P}(x_{ip} - x_{jp})^2}$$

2. Квадрат евклидова расстояния, также обозначается L2 squared, необходимо для некоторых методов агломерации, в частности, для метода Уорда (Варда):

$$d(x_i, x_j) = \sum_{p=1}^{P}(x_{ip} - x_{jp})^2.$$

3. Манхэттенское расстояние, оно же блочное расстояние, также обозначается L1:

$$d(x_i, x_j) = \sum_{p=1}^{P}|x_{ip} - x_{jp}|.$$

4. Расстояние Чебышёва:

$$d(x_i, x_j) = \max\{|x_{ip} - x_{jp}|\}.$$

Конечно, перечисленными выше расстояниями список метрик не исчерпывается, их гораздо больше. Но далеко не все метрики активно используются в кластерном анализе. Тем не менее, хотелось бы выделить группу метрик, которые помимо «естественных» геометрических координат используют информацию о распределении данных. Так, например, расстояние Махаланобиса использует ковариационную матрицу, содержащую информацию о связи между случайными векторами и их дисперсии.

Из гпт:

Евклидово расстояние (L2):

- **Описание:** Измеряет прямое расстояние между точками в пространстве, вычисляя корень из суммы квадратов различий между соответствующими элементами двух точек.
- **Использование:** Используется в алгоритмах, где важны реальные физические расстояния, например в алгоритмах кластеризации, таких как K-means, или в задачах классификации с использованием методов ближайших соседей.

Квадрат Евклидова расстояния (L2 squared):

- **Описание:** Вместо извлечения корня считается просто сумма квадратов разностей между соответствующими элементами. Это делает вычисления проще и быстрее.
- **Использование:** Полезно в оптимизационных алгоритмах, где порядок расстояний важнее точных значений, например в алгоритмах кластеризации или при вычислении ошибок в методах обучения машин.

Манхэттенское расстояние (L1):

- **Описание:** Суммирует абсолютные значения различий между элементами двух точек. Это расстояние аналогично тому, как бы вы перемещались по городским улицам, которые устроены по принципу сетки.
- **Использование:** Хорошо работает в ситуациях, где данные имеют аномалии (выбросы), которые могут существенно искажать Евклидово расстояние. Также используется в задачах компрессии и анализа изображений.

Расстояние Чебышева:

- **Описание:** Измеряет максимальное из абсолютных различий между элементами двух точек. Это как бы максимальное расстояние, которое нужно пройти в одном измерении.
- **Использование:** Используется в бесконечно-мерных пространствах и в играх на сетке, где вы можете двигаться в любом направлении (например, король в шахматах), а также в многомерных анализах, где важно наибольшее отклонение среди всех измерений.

Эти меры расстояния выбираются в зависимости от природы данных и специфических требований задачи, что позволяет более гибко адаптироваться к различным условиям и целям анализа.

39. Анализ остатков. Построение графиков

Из лекции:

Проверка на наличие гетероскедастичности.

1) **Методом графического анализа остатков.** В этом случае по оси абсцисс откладываются значения объясняющей переменной Xi, а по оси ординат квадраты отклонения εi 2 . Если имеется определенная связь между отклонениями, то гетероскедастичность имеет место. Отсутствие зависимости скоре всего будет свидетельствовать об отсутствии гетероскедастичности

2) При помощи теста ранговой корреляции Спирмена.

https://math.semestr.ru/corel/spirmen.php

Задаются два параметра: xi, ε.

y	y(x)	ε = y-y(x)	ε^2
0.9	1.94	-1.04	1.08
1.7	1.3	0.4	0.16
0.7	0.8	-0.0954	0.0091
1.7	1.39	0.31	0.0961
2.6	1.38	1.22	1.49
1.3	1.58	-0.28	0.0781
4.1	4.41	-0.31	0.0977
1.6	1.33	0.27	0.0721
6.9	6.63	0.27	0.0703
0.4	1.13	-0.73	0.54

Из интернета:

Анализ остатков— это инструмент для оценки соответствия статистической модели. Он позволяет изучить разницу между прогнозируемыми значениями и фактическими значениями переменной отклика.

Анализ остатков помогает определить, правильно ли модель определяет взаимосвязь между предикторами и переменной ответа.

Основные характеристики остатков:

1. **Независимость.** Если остатки не являются независимыми, это указывает на то, что в данных есть шаблон, который модель не захватила.

2. **Распределённость.** Если остатки обычно не распределены, это может указывать на то, что модель не подходит для данных.

3. **Постоянная дисперсия.** Если дисперсия остатков изменяется со значениями прогнозируемой переменной, это указывает на то, что модель не подходит для данных.

40. Общиˇподход к оцениванию факторов. Общая постановка задачи.

***Из лекции*:**

При **оценке мультиколлинеарности факторов** следует учитывать, что чем ближе к 0 определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии.

Для отбора наиболее значимых факторов xi учитываются следующие условия:

- связь между результативным признаком и факторным должна быть выше межфакторной связи;
- связь между факторами должна быть не более 0.7;
- при высокой межфакторной связи признака отбираются факторы с меньшим коэффициентом корреляции между ними.

Более объективную характеристику тесноты связи дают частные коэффициенты корреляции, измеряющие влияние на результат фактора xi при неизменном уровне других факторов.

Из гпт:

1. Постановка задачи:

Определение цели анализа: Необходимо четко сформулировать, какую информацию вы хотите извлечь из данных и для каких целей. Это может быть предсказание будущих событий, выявление закономерностей, улучшение бизнес-процессов и т.д.

Определение факторов и переменных: Необходимо определить, какие именно факторы (переменные) будут учитываться в анализе. Факторы могут быть внутренними (например, данные о продажах) и внешними (например, экономические индикаторы).

2. Сбор данных:

Источники данных: Определение источников данных (базы данных, веб- сайты, сенсоры и т.д.) и методы их получения.

Очистка данных: Обработка и очистка данных для удаления ошибок, дубликатов и пропусков.

3. Предварительный анализ данных:

Описание данных: Исследование структуры данных, анализ распределения значений переменных.

Корреляционный анализ: Выявление взаимосвязей между различными переменными.

4. Выбор модели и методов:

Методы статистического анализа: Линейная регрессия, логистическая регрессия, ANOVA и т.д.

Методы машинного обучения: Решающие деревья, случайные леса, градиентный бустинг, нейронные сети и т.д.

5. Обучение модели:

Разделение данных: Разделение данных на тренировочную и тестовую выборки.

Обучение и валидация: Обучение модели на тренировочной выборке и оценка ее качества на валидационной выборке.

6. Оценка значимости факторов:
Статистические методы: t-тесты, p-значения, коэффициенты корреляции и другие методы для оценки значимости переменных.

Методы машинного обучения: Оценка важности признаков с помощью методов вроде Gini Importance в решающих деревьях или методов регуляризации в линейных моделях.

7. Интерпретация и визуализация результатов:

Интерпретация: Оценка полученных результатов с точки зрения их практической значимости.

Визуализация: Использование графиков, диаграмм и других методов визуализации для наглядного представления результатов.

8. Внедрение и мониторинг:

Внедрение модели: Интеграция модели в бизнес-процессы или ИТ-системы.
Мониторинг и обновление: Постоянный мониторинг работы модели и обновление ее по мере необходимости.

41. МНК. Функция потерь. Система нормальных уравнений.

Из гпт:

МНК (Метод Наименьших Квадратов)

1. Основные концепции:

Метод наименьших квадратов (МНК) используется для оценки параметров линейной регрессии, минимизируя сумму квадратов разностей между наблюдаемыми значениями и предсказанными значениями, полученными моделью.

1. **Модель**: Определение линейной регрессионной модели: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$.
2. **Функция потерь**: Минимизация суммы квадратов разностей между наблюдаемыми и предсказанными значениями: $S = \sum (y_i - \hat{y}_i)^2$.
3. **Оптимизация**: Решение системы нормальных уравнений для нахождения оценок параметров β_j.

Эти шаги формируют основу метода наименьших квадратов для аппроксимации данных линейной функцией и определения параметров этой функции.

2. **ФУНКЦИЯ ПОТЕРЬ:**

Функция потерь для МНК определяется как сумма квадратов ошибок (разностей между наблюдаемыми значениями и предсказанными значениями):

$$L(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2$$

где:

- y_i - наблюдаемое значение зависимой переменной для i-го наблюдения,
- $\hat{y}_i$ - предсказанное значение зависимой переменной для i-го наблюдения,
- $\beta_0, \beta_1, \ldots, \beta_p$ - коэффициенты модели,
- $x_{i1}, x_{i2}, \ldots, x_{ip}$ - значения независимых переменных для i-го наблюдения.

3. **СИСТЕМА НОРМАЛЬНЫХ УРАВНЕНИЙ:**

Для минимизации функции потерь используется частичное дифференцирование по каждому параметру βj и приравнивание производных к нулю. Это приводит к системе нормальных уравнений:

$$X^T X\beta = X^T y$$

где:

- X - матрица независимых переменных (размером $n \times (p+1)$, где n - количество наблюдений, а p - количество независимых переменных),
- y - вектор зависимых переменных (размером $n \times 1$),
- β - вектор коэффициентов модели (размером $(p+1) \times 1$).

4. Решение системы нормальных уравнений:

Для получения оценки параметров β необходимо решить систему нормальных уравнений. Это можно сделать путем умножения обеих частей уравнения на (X ^T X) ^−1:

$$\beta = (X^T X)^{-1} X^T y$$

где $(X^T X)^{-1}$ - обратная матрица к $X^T X$ (при условии, что эта матрица существует и не вырождена).

42. Вычисление оценки дисперсии ошибки измерения целевой функции

Лекция: Уравнение_множественной_регрессии.pdf

Гомоскедастичность (постоянство дисперсий отклонений). Дисперсия случайных отклонений εi постоянна: D(εi) = D(εj) = S2 для любых i и j.

Матрица A^TA.

10	21.9	416.6	264.6
21.9	82.67	1875.41	982.39
416.6	1875.41	48382.48	25388.13
264.6	982.39	25388.13	15298.4

Полученная матрица имеет следующее соответствие:

$\sum n$	$\sum y$	$\sum x_1$	$\sum x_2$
$\sum y$	$\sum y^2$	$\sum x_1 y$	$\sum x_2 y$
$\sum x_1$	$\sum y x_1$	$\sum x_1{}^2$	$\sum x_2 x_1$
$\sum x_2$	$\sum y x_2$	$\sum x_1 x_2$	$\sum x_2{}^2$

Найдем парные коэффициенты корреляции.

Для y и x_1

Средние значения

$$\overline{x} = \frac{\sum x_i}{n} = \frac{416.6}{10} = 41.66$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{21.9}{10} = 2.19$$

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{1875.41}{10} = 187.54$$

Дисперсия

$$D(x) = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{48382.48}{10} - 41.66^2 = 3102.69$$

$$D(y) = \frac{\sum y_i^2}{n} - \overline{y}^2 = \frac{82.67}{10} - 2.19^2 = 3.47$$

Среднеквадратическое отклонение

$$s(x) = \sqrt{D(x)} = \sqrt{3102.69} = 55.7$$

$$s(y) = \sqrt{D(y)} = \sqrt{3.47} = 1.86$$

Коэффициент корреляции

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{s(x) \cdot s(y)} = \frac{187.54 - 41.66 \cdot 2.19}{55.7 \cdot 1.86} = 0.93$$

Для y и x_2

Средние значения

$$\overline{x} = \frac{\sum x_i}{n} = \frac{264.6}{10} = 26.46$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{21.9}{10} = 2.19$$

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{982.39}{10} = 98.24$$

Дисперсия

$$D(x) = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{15298.4}{10} - 26.46^2 = 829.71$$

$$D(y) = \frac{\sum y_i^2}{n} - \overline{y}^2 = \frac{82.67}{10} - 2.19^2 = 3.47$$

Среднеквадратическое отклонение

$$s(x) = \sqrt{D(x)} = \sqrt{829.71} = 28.8$$

$$s(y) = \sqrt{D(y)} = \sqrt{3.47} = 1.86$$

Коэффициент корреляции

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{s(x) \cdot s(y)} = \frac{98.24 - 26.46 \cdot 2.19}{28.8 \cdot 1.86} = 0.75$$

Для x_1 и x_2

Средние значения

$$\overline{x} = \frac{\sum x_i}{n} = \frac{264.6}{10} = 26.46$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{416.6}{10} = 41.66$$

$$\overline{xy} = \frac{\sum x_i y_i}{n} = \frac{25388.13}{10} = 2538.81$$

Дисперсия

$$D(x) = \frac{\sum x^2_i}{n} - \overline{x}^2 = \frac{15298.4}{10} - 26.46^2 = 829.71$$

$$D(y) = \frac{\sum y^2_i}{n} - \overline{y}^2 = \frac{48382.48}{10} - 41.66^2 = 3102.69$$

Среднеквадратическое отклонение

$$s(x) = \sqrt{D(x)} = \sqrt{829.71} = 28.8$$

$$s(y) = \sqrt{D(y)} = \sqrt{3102.69} = 55.7$$

Коэффициент корреляции

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{s(x) \cdot s(y)} = \frac{2538.81 - 26.46 \cdot 41.66}{28.8 \cdot 55.7} = 0.9$$

Для несмещенной оценки дисперсии проделаем следующие вычисления:

Несмещенная ошибка $\varepsilon = Y - Y(x) = Y - X^*s$ (абсолютная ошибка аппроксимации)

Y	Y(x)	ε	$(Y-Ycp)^2$
0.9	1.94	-1.04	1.66
1.7	1.3	0.4	0.24
0.7	0.8	-0.0954	2.22
1.7	1.39	0.31	0.24
2.6	1.38	1.22	0.17
1.3	1.58	-0.28	0.79
4.1	4.41	-0.31	3.65
1.6	1.33	0.27	0.35
6.9	6.63	0.27	22.18
0.4	1.13	-0.73	3.2
			34.71

$$s_\varepsilon^2 = (Y - X^*s)^T(Y - X^*s) = 3.7$$

Несмещенная оценка дисперсии равна:

$$s^2 = \frac{1}{n-m-1} s_\varepsilon^2 = \frac{1}{10-3} 3.7 = 0.53$$

Оценка среднеквадратичного отклонения равна (*стандартная ошибка для оценки Y*):

$$S = \sqrt{S^2} = \sqrt{0.53} = 0.73$$

Найдем оценку ковариационной матрицы вектора $k = S \cdot (X^TX)^{-1}$

$$k(x) = 0.728 \begin{vmatrix} 0,187 & 0,001 & -0,004 \\ 0,001 & 0 & -0 \\ -0,004 & -0 & 0,001 \end{vmatrix} = \begin{vmatrix} 0,136 & 0 & -0,003 \\ 0 & 0 & -0 \\ -0,003 & -0 & 0 \end{vmatrix}$$

Дисперсии параметров модели определяются соотношением S 2i = Kii, т.е. это элементы, лежащие на главной диагонали

GPT

Вычисление оценки дисперсии ошибки измерения целевой функции — важный этап анализа данных, особенно в задачах машинного обучения и статистики.

Шаг 1: Сбор данных

Соберите набор данных, содержащий измерения целевой функции y.
Пусть у вас будет n измерений $y_1, y_2, \ldots, y_n$.

Шаг 2: Вычисление средней величины

Найдите среднее значение измерений:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Шаг 3: Вычисление отклонений от среднего

Для каждого измерения y_i вычислите отклонение от среднего значения:

$$d_i = y_i - \bar{y}$$

Шаг 4: Вычисление квадратов отклонений

Вычислите квадрат каждого отклонения:

$$d_i^2 = (y_i - \bar{y})^2$$

Шаг 5: Суммирование квадратов отклонений

Суммируйте все квадраты отклонений:

$$S = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Шаг 6: Вычисление оценки дисперсии ошибки измерения

Дисперсия ошибки измерения оценивается как среднее значение квадратов отклонений:

$$\sigma^2 = \frac{S}{n-1}$$

43. Регрессионный анализ. Основная идея. Регрессионная, функциональная, аппроксимирующая зависимости. Зависимые и независимые переменные.

Уравнение множественной регрессии.
Уравнение множественной регрессии может быть представлено в виде:
$Y = f(\beta, X) + \varepsilon$
где $X = X(X_1, X_2, ..., X_m)$ - вектор независимых (объясняющих) переменных; β - вектор параметров (подлежащих определению); ε - случайная ошибка (отклонение); Y - зависимая (объясняемая) переменная.
теоретическое линейное уравнение множественной регрессии имеет вид:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon$
β_0 - свободный член, определяющий значение Y, в случае, когда все объясняющие переменные X_j равны 0.

GPT
Регрессионный анализ — набор статистических методов исследования влияния одной или

нескольких независимых переменных на зависимую переменную.
Цели регрессионного анализа:

- Определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными).
- Предсказание значения зависимой переменной с помощью независимой переменной.
- Определение вклада отдельных независимых переменных в вариацию зависимой.
- Наиболее распространённый вид регрессионного анализа — линейная регрессия, когда находят линейную функцию, которая, согласно определённым математическим критериям, наиболее соответствует данным.

  Регрессионная зависимость характеризует среднюю тенденцию изменения зависимой переменной при изменении независимых переменных. Эта зависимость выражается через регрессионную модель, обычно в виде уравнения. Например, в линейной регрессии эта зависимость выражается уравнением прямой линии.

  Функциональная зависимость — это строгая математическая зависимость, при которой каждому значению независимой переменной соответствует одно определенное значение зависимой переменной. В регрессионном анализе функциональная зависимость редко встречается, так как реальные данные часто содержат случайные колебания.

  Аппроксимирующая зависимость используется для описания реальных данных, которые могут быть шумными и содержать ошибки. Цель аппроксимации — найти гладкую функцию, которая наилучшим образом приближает данные. Регрессионные модели часто используются в качестве аппроксимирующих зависимостей.

  Зависимая переменная — это переменная, которую мы пытаемся предсказать или объяснить. Независимые переменные — это переменные, которые используются для объяснения вариации зависимой переменной

44 Метод k-средних.

Из лекции:

Задача классификации решается при помощи различных методов, наиболее простой - линейная регрессия. Выбор метода должен базироваться на исследовании исходного набора данных. **Наиболее распространенные методы решения задачи кластеризации: метод k-средних** (работает только с числовыми атрибутами), иерархический кластерный анализ (работает также с символьными атрибутами), метод SOM. Сложностью кластеризации является необходимость ее оценки.

Если речь идёт об иерархическом кластерном анализе, заранее знать количество кластеров, которое мы хотим получить, необязательно, однако если мы говорим о кластеризации методом k-средних, количество желаемых кластеров является необходимым параметром. О различиях этих видов кластерного анализа мы поговорим чуть позже, пока стоит отметить, что в любом случае априорная информация о количестве кластеров будет полезна.

Из гпт:

Метод k-средних — это алгоритм кластеризации, который стремится разделить набор данных на
k
k предварительно заданных групп так, чтобы каждая точка данных принадлежала кластеру с ближайшим средним значением. Он широко используется в анализе данных для выявления групп или кластеров схожих элементов в наборе данных. Вот подробное объяснение каждого шага алгоритма:

1. Инициализация

Выбираются k начальных центров (центроидов), обозначаемых как $\mu_1, \mu_2, \ldots, \mu_k$. Центроиды могут быть случайными точками из данных или выбраны с помощью улучшенного метода инициализации.

2. Шаг присваивания

Каждая точка данных x_i присваивается кластеру на основе минимального расстояния до центроидов. Расстояние часто вычисляется как Евклидово расстояние. Математически это можно выразить как:

$$\text{Присвоить } x_i \text{ к кластеру } j, \text{ так что } j = \arg\min_k \|\mu_k - x_i\|^2$$

Здесь $\|\mu_k - x_i\|^2$ представляет квадрат Евклидова расстояния между точкой x_i и центроидом μ_k.

3. Шаг обновления

Центроид каждого кластера обновляется, чтобы стать средним всех точек, отнесенных к этому кластеру. Если обозначить S_j множество всех точек данных в кластере j, то новый центроид μ_j вычисляется как:

$$\mu_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$$

Здесь $|S_j|$ — количество элементов в кластере j, а суммирование производится по всем точкам x_i, которые принадлежат кластеру j.

4. Повторение

Шаги 2 и 3 повторяются до тех пор, пока не будет достигнута сходимость, что часто определяется как отсутствие или минимальные изменения в положениях центроидов после последовательных итераций.

5. Результат

После завершения алгоритма, данные разделены на k кластеров, где каждый кластер характеризуется своим центроидом и состоит из точек, которые ближе всего к этому центроиду по сравнению с другими.

Эти формулы и шаги обеспечивают математическую основу для понимания и реализации алгоритма k-средних в различных приложениях.

45 **Расчет коэффициентов регрессии.**

Коэффициенты регрессии (β) — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой.

Смысл коэффициента регрессии

В общем случае коэффициент регрессии *k* показывает, как в среднем изменится *результативный признак* (*Y*), если *факторный признак* (*X*) увеличится на единицу .
Y = 87610 + 2984 *X* ; ***X*** *– число рабочих,* ***Y*** *– объем годового производства (руб.).*
Пример интерпретации коэффициента регрессии

- В уравнении *Y* = 87610 + **2984** *X* ; коэффициент регрессии равен +2984.

Что это означает?

•В данном случае смысл коэффициента регрессии состоит в том, что увеличение *числа рабочих* на 1 чел. приводит в среднем к увеличению объема годового *производства* на 2984 руб.

Свойства коэффициента регрессии

- Коэффициент регрессии может принимать любые значения.
- Коэффициент регрессии *не симметричен* , т.е. изменяется, если *X* и *Y* поменять местами.
- *Единицей измерения* коэффициента регрессии является отношение единицы измерения *Y* к единице измерения *X* : ([*Y*] / [*X*]).
- Коэффициент регрессии *изменяется при изменении единиц измерения* X и Y .
- Поскольку результативный признак *Y* измеряется в рублях, а факторный признак *X* в количестве рабочих (чел.), то коэффициент регрессии измеряется *в рублях на человека* (руб. / чел.)

Расчет коэффициентов регрессии для линейной и множественной регрессии

1. Линейная регрессия

Модель линейной регрессии:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

где Y — зависимая переменная, X — независимая переменная, β_0 и β_1 — коэффициенты регрессии, которые нам нужно оценить, и ϵ — случайная ошибка.

Метод наименьших квадратов (МНК) используется для нахождения коэффициентов β_0 и β_1 путем минимизации суммы квадратов остатков:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Решение для β_1 и β_0:

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

где $\overline{x}$ и $\overline{y}$ — средние значения X и Y соответственно.

2. Множественная регрессия

Модель множественной регрессии:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

где Y — зависимая переменная, $X_1, X_2, \ldots, X_k$ — независимые переменные, и $\beta_0, \beta_1, \ldots, \beta_k$ — коэффициенты регрессии.

Метод наименьших квадратов в матричной форме:

Матричная форма уравнения регрессии:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

где $\mathbf{Y}$ — вектор зависимых переменных, $\mathbf{X}$ — матрица наблюдений (включая столбец из единиц для коэффициента β_0), $\boldsymbol{\beta}$ — вектор коэффициентов, $\boldsymbol{\epsilon}$ — вектор ошибок.

Оценка коэффициентов:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Это выражение представляет собой решение нормальных уравнений регрессии и предполагает, что матрица $\mathbf{X}^T\mathbf{X}$ обратима.

Интерпретация коэффициентов:

- Коэффициент β_j показывает изменение в Y на одну единицу изменения в X_j, при условии постоянства всех других независимых переменных.
- Коэффициент β_0 интерпретируется как значение Y, когда все $X_j = 0$.

Заключение

В линейной и множественной регрессии метод наименьших квадратов предоставляет способ оценки коэффициентов, минимизируя сумму квадратов ошибок между наблюдаемыми и предсказанными значениями. Различия между линейной и множественной регрессией заключаются в количестве независимых переменных и сложности вычислений, особенно в матричных операциях для множественной регрессии.

46 **Информационная матрица. Ее свойства. Матрица ошибок.**

Информационная матрица Фишера используется для вычисления ковариационных матриц, связанных с оценками максимального правдоподобия.

В математической статистике информация Фишера (иногда называемая просто информацией) - это способ измерения количества информации, которую наблюдаемая случайная величина X несет в себе о неизвестном параметре θ распределения, моделирующего X. Формально это дисперсия оценки или ожидаемое значение наблюдаемой информации.

Определение

Информационная матрица $\mathbf{I}(\theta)$ для набора параметров θ модели определяется как матрица ожидаемых значений вторых производных логарифма функции правдоподобия по этим параметрам:

$$\mathbf{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log L(\theta;\mathbf{X})}{\partial\theta\partial\theta^T}\right]$$

где $L(\theta;\mathbf{X})$ — функция правдоподобия, а $\mathbf{X}$ — наблюдаемые данные.

Свойства:

- Симметричность: Информационная матрица симметрична
- Положительная полуопределенность: Информационная матрица является положительно полуопределённой, что означает, что все её собственные значения неотрицательны.

Связь с оценкой параметров:

Информационная матрица используется для оценки дисперсий и ковариаций максимального правдоподобия (MLE) оценок параметров. Если $\hat{\theta}$ — оценка параметра θ, то асимптотическая дисперсия этой оценки определяется как:

$$\mathrm{Var}(\hat{\theta}) \approx \mathbf{I}(\theta)^{-1}$$

-

Матрица ошибок (неточностей) — это инструмент, который позволяет определить, в чём модель ошибается.

Эта матрица сравнивает количество правильных и неправильных предсказаний для каждого класса.

- В матрицу ошибок записывают 4 вида результатов:
- Истинно положительные (ИП/TP): количество положительных наблюдений, которые модель правильно предсказала как положительные.

- Ложноположительные (ЛП/FP): количество отрицательных наблюдений, которые модель неверно предсказала как положительные.
- Истинно отрицательные (ИО/TN): количество отрицательных наблюдений, которые модель правильно предсказала как отрицательные.
- Ложноотрицательные (ЛО/FN): количество положительных наблюдений, которые

модель неверно предсказала как отрицательные.