

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

ЛЕКЦИЯ 1

Введение в курс
Методы анализа данных
Data Mining

История возникновения

Предпосылки:

- законы больших чисел для конечных выборок не выполняются;
- характеристики центральной тенденции (средняя арифметическая, мода, медиана) часто не являются характеристиками совокупности и приводят к операциям над фиктивными величинами (типа средней температуры больных по больнице, среднего дохода рабочих и миллионеров);
- закон распределения нельзя достоверно определить по выборочным данным;
- вероятность как характеристика неопределенности часто вводится необоснованно;
- сумма воздействия ненаблюдаемых и неконтролируемых факторов может привести к структурным изменениям в наблюдаемой системе, которые приведут к изменению априорных условий моделирования и т. д.
- «проклятие размерности» при анализе сложных систем, предполагающем исследование всей системы

Дж.Тьюки в 60-е годы предложил разведочный анализ данных (РАД; Exploratory data analysis), основанный на использовании методов многомерной статистики.

- РАД предполагает изучение не только вероятностной, но и геометрической природы данных.

Разведочный анализ данных (РАД, Exploratory data analysis (EDA))

— анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей.

Цели РАД:

- максимальное "проникновение" в данные
- выявление основных структур
- выбор наиболее важных переменных
- обнаружение отклонений и аномалий
- проверка основных гипотез
- разработка начальных моделей

Основные инструменты РАД:

- анализ вероятностных распределений переменных
- построение и анализ корреляционных матриц
- факторный анализ
- дискриминантный анализ
- многомерное шкалирование и др.

Дальнейшее развитие

1994 г. известный математик Лотфи Заде сформулировал принцип «мягких вычислений» - Soft Computing (терпимость к нечёткости и частичной истинности используемых данных для достижения интерпретируемости, гибкости и низкой стоимости решений)



Появление в середине 90-х годов XX века нового направления в науке - Data Mining (добыча данных), или иначе: интеллектуальный анализ данных.

- Идеология Data Mining появилась на стыке **прикладной статистики, искусственного интеллекта, баз данных** и т. д.

Фактически рождению нового направления в анализе данных способствовало **появление компьютеров и совершенствование технологий записи и хранения данных**.

Добыча данных - Data Mining

Data Mining - исследование и обнаружение "машиной" (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком.

- Знания должны быть новые, ранее неизвестные.
- Знания должны быть нетривиальны.
- Знания должны быть практически полезны.
- Знания должны быть доступны для понимания человеку.

Термин введён Григорием Пятецким-Шапиро в 1989 году

Важное отличие процедуры добычи данных от классического РАД: системы добычи данных в большей степени ориентированы на практическое приложение полученных результатов, чем на выяснение природы явления.

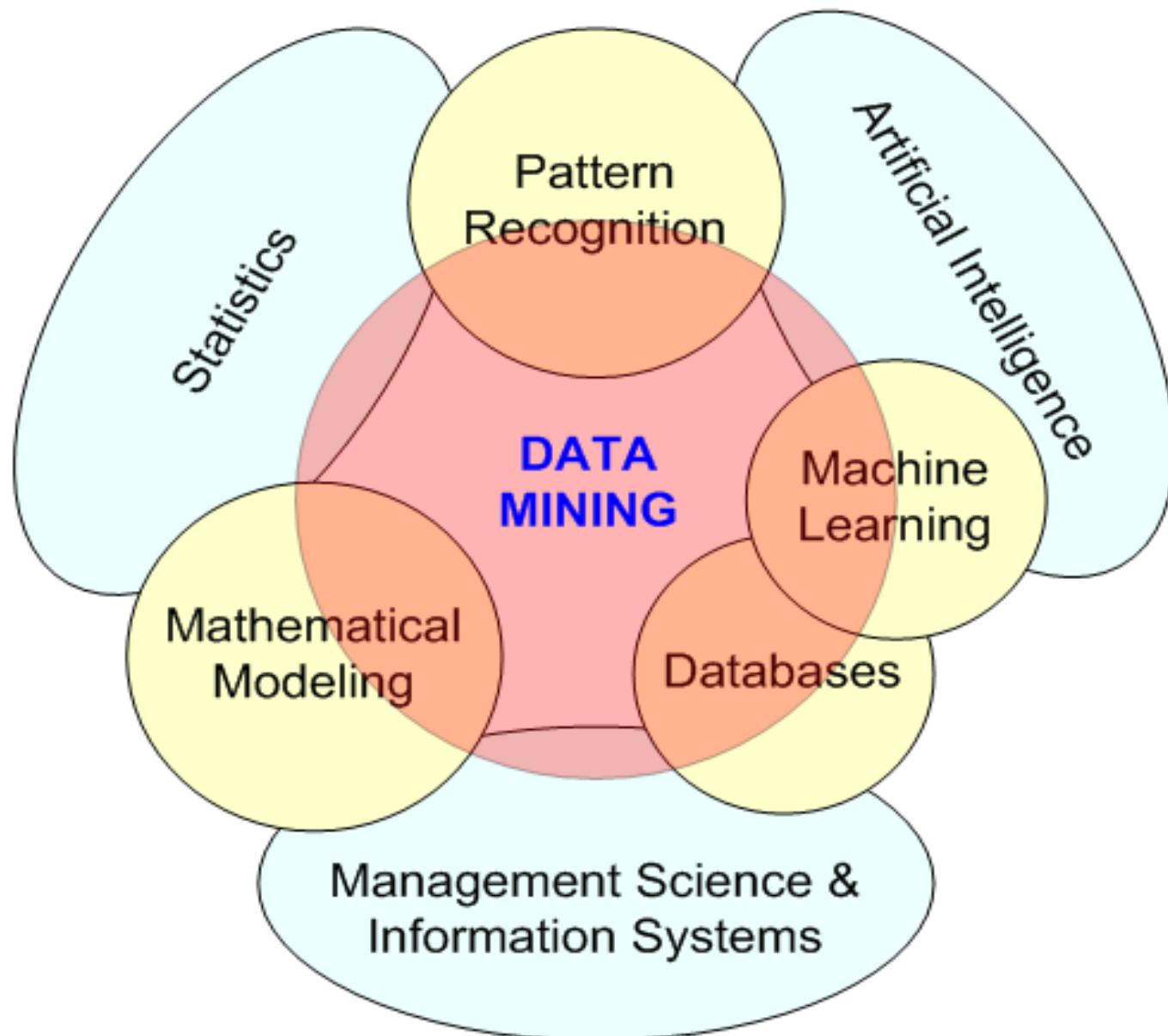
Knowledge Discovery in Database (KDD)

- Извлечение знаний из баз данных
- Описывает последовательность действий, которую необходимо выполнить для обнаружения полезного знания
- Не зависит от предметной области



Григорий Пятецкий-Шапиро
президент и главный редактор одного из
первых сайтов (1994 г.) по Анализу
данных «KDnuggets»
<http://www.kdnuggets.com/>

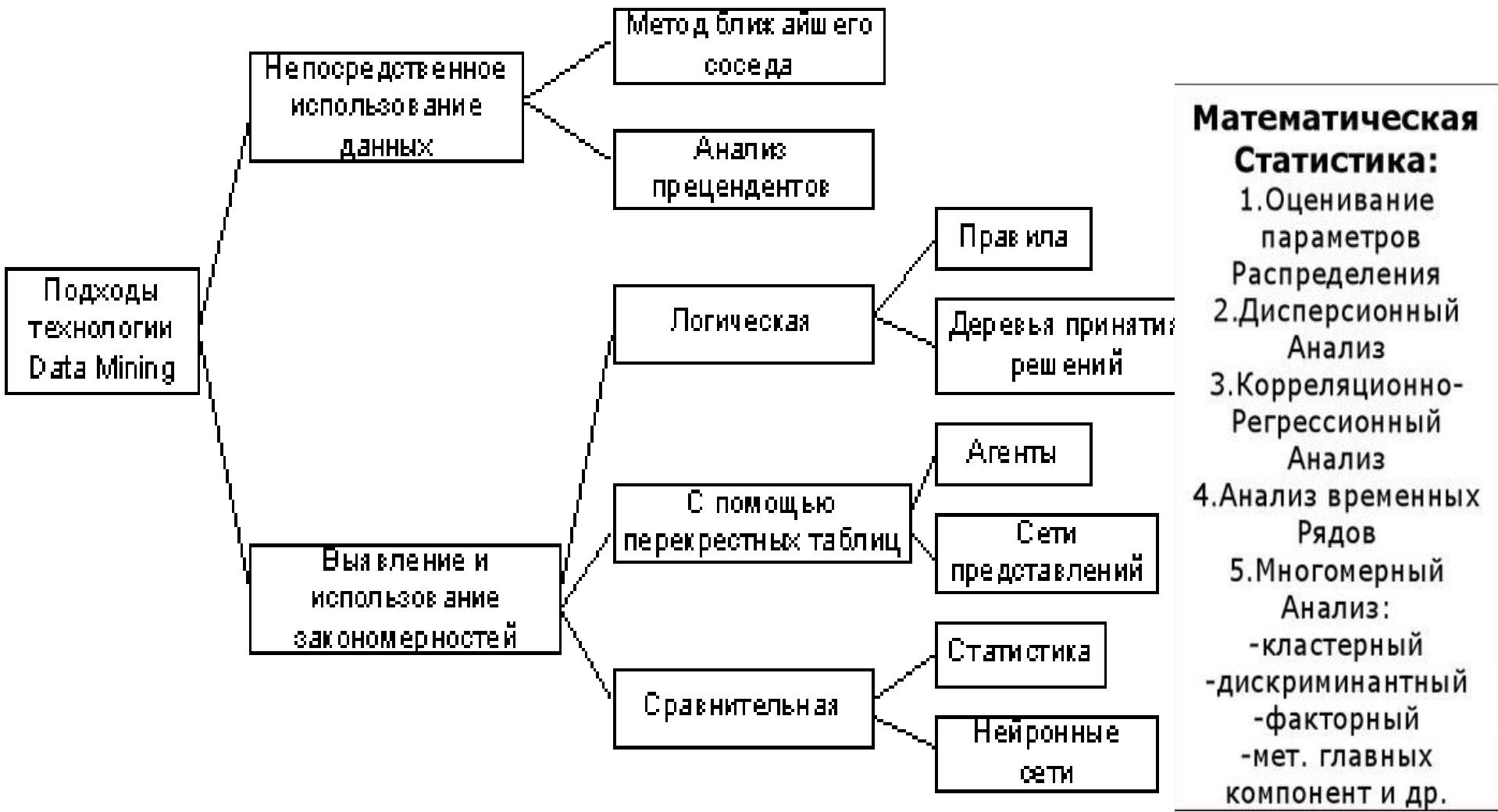
Связь с другими дисциплинами



Классификация уровней информации

Уровень информации	Описание
Сырые данные (raw data)	Необработанные данные, получаемые в результате наблюдения за объектами и отображающие их состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год, данные о ценах на рынке жилья, данные об абитуриентах, зачисленных на 1 курс)
Информация	Это либо: <ul style="list-style-type: none">- сырье данные, но систематизированные, представленные в более компактном виде (например, результаты поиска – сведения об абитуриентах, поступивших в ИИТиУТС СевГУ в этом году);- обработанные данные, имеющие информационную ценность для пользователя (например, сводные статистические характеристики – средний балл абитуриентов, поступивших в ИИТиУТС СевГУ в этом году – его абсолютная величина и % по отношению к тому же показателю за предыдущий год)
Знания	Понятие «знания» включает: <ul style="list-style-type: none">- скрытые взаимосвязи между объектами (признаками объектов);- некоторое ноу-хау, алгоритмы, методы решения задач. Знания обладают практической ценностью.

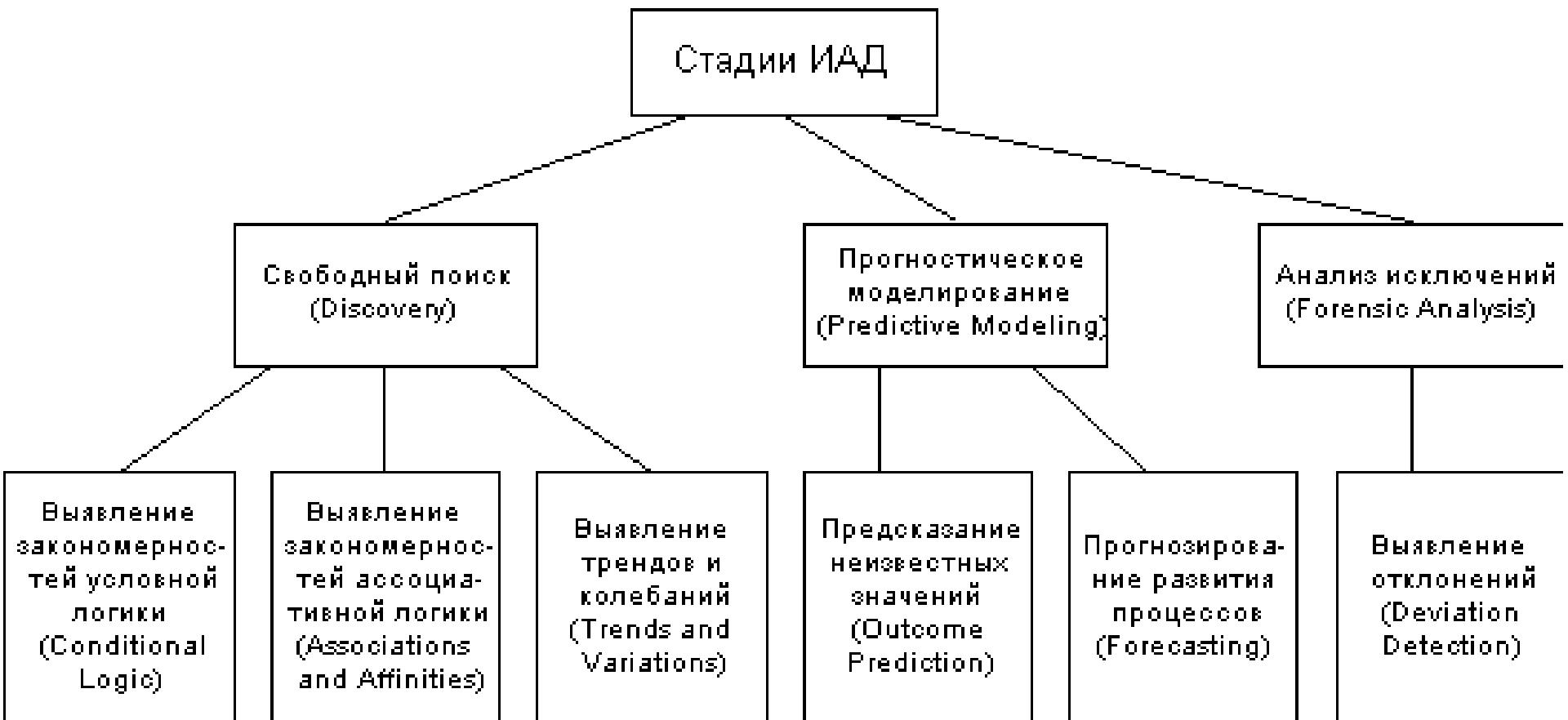
Подходы технологии ИАД



Методы и алгоритмы анализа данных

- 1) искусственные нейронные сети
- 2) деревья решений, символные правила
- 3) методы ближайшего соседа и k-ближайшего соседа
- 4) метод опорных векторов
- 5) байесовские сети
- 6) линейная регрессия
- 7) корреляционно-регрессионный анализ
- 8) иерархические методы кластерного анализа
- 9) неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы
- 10) методы поиска ассоциативных правил, в том числе алгоритм Apriori
- 11) метод ограниченного перебора
- 12) эволюционное программирование и генетические алгоритмы
- 13) разнообразные методы визуализации данных и др.

Стадии интеллектуального анализа данных





Задачи Data Mining



Задачи Data Mining



Задачи Data Mining

- **Задача классификации** сводится к определению класса объекта по его характеристикам. Множество классов известно заранее.
- **Задача регрессии** подобно задаче классификации позволяет определить по известным характеристикам объекта значение некоторого параметра из множества действительных чисел.
- При **поиске ассоциативных правил** целью является нахождение частых зависимостей (или ассоциаций)
- **Задача кластеризации** заключается в поиске независимых групп (кластеров) и их характеристик во всем множестве анализируемых данных.

Описательные и предсказательные задачи

Описательные (descriptive) задачи предназначены для улучшения понимания анализируемых данных.

К такому виду задач относятся кластеризация и поиск ассоциативных правил

Предсказательные (predictive) задачи. Решение разбивается на два этапа:

- 1) на основании набора данных с известными результатами строится модель;
- 2) полученная модель используется для предсказания результатов на основании новых наборов данных (требование максимальной точности).

К данному виду задач относят задачи классификации и регрессии, задача поиска ассоциативных правил, если результаты ее решения могут быть использованы для предсказания появления некоторых событий.

Supervised и unsupervised learning

Supervised learning - обучение с учителем –

задача анализа данных решается в несколько этапов:

- строится модель анализируемых данных – классификатор;
- классификатор подвергается обучению (проверяется качество его работы, и, если оно неудовлетворительное, происходит дополнительное обучение классификатора)
- продолжается пока не будет достигнут требуемый уровень качества или не станет ясно, что выбранный алгоритм не работает корректно с данными, либо же сами данные не имеют структуры, которую можно выявить.

К этому типу задач относят задачи классификации и регрессии.

Unsupervised learning - обучение без учителя –

объединяет задачи, выявляющие описательные модели.

Достоинство таких задач - возможность их решения без каких либо предварительных знаний об анализируемых данных. К этим задачам относятся кластеризация и поиск ассоциативных правил.

Задача классификации и регрессии

Постановка: требуется определить, к какому из известных классов относятся исследуемые объекты, т. е. классифицировать их.

Клиент банка: «кредитоспособен» и «некредитоспособен».

Фильтр электронной почты: «спам», «не спам»

Распознавание цифр: от 0 до 9.

В Data Mining задачу классификации рассматривают как задачу определения значения одного из параметров анализируемого объекта на основании значений других параметров.

Задача классификации и регрессии решается в два этапа:

1) выделяется обучающая выборка, в нееходят объекты, для которых известны значения как независимых, так и зависимых переменных.

Задача классификации и регрессии

На основании обучающей выборки строится модель определения значения зависимой переменной - функция классификации или регрессии.

Для получения максимально точной функции к обучающей выборке предъявляются следующие основные требования:

- количество объектов, входящих в выборку, должно быть достаточно большим. Чем больше объектов, тем точнее будет построенная на ее основе функция классификации или регрессии;
 - в выборку должны входить объекты, представляющие все возможные классы в случае задачи классификации или всю область значений в случае задачи регрессии;
 - для каждого класса в задаче классификации или для каждого интервала области значений в задаче регрессии выборка должна содержать достаточное количество объектов.
- 2) построенную модель применяют к анализируемым объектам (к объектам с неопределенным значением зависимой переменной).

Задача поиска ассоциативных правил

Первоначально она решалась при анализе тенденций в поведении покупателей в супермаркетах (анализ рыночных корзин - Basket Analysis). При анализе этих данных интерес прежде всего представляет информация о том, какие товары покупаются вместе, в какой последовательности, какие категории потребителей какие товары предпочитают, в какие периоды времени и т. п.

В сфере обслуживания интерес представляет информация о том, какими услугами клиенты предпочитают пользоваться в совокупности.

В медицине - анализ сочетания симптомов и болезней.

Сиквенциальный анализ учитывает последовательность происходящих событий (телеkomмуникационные компании, анализ аварий).

Задача кластеризации

Задача кластеризации состоит в разделении исследуемого множества объектов на группы "похожих" объектов, называемых кластерами (cluster).

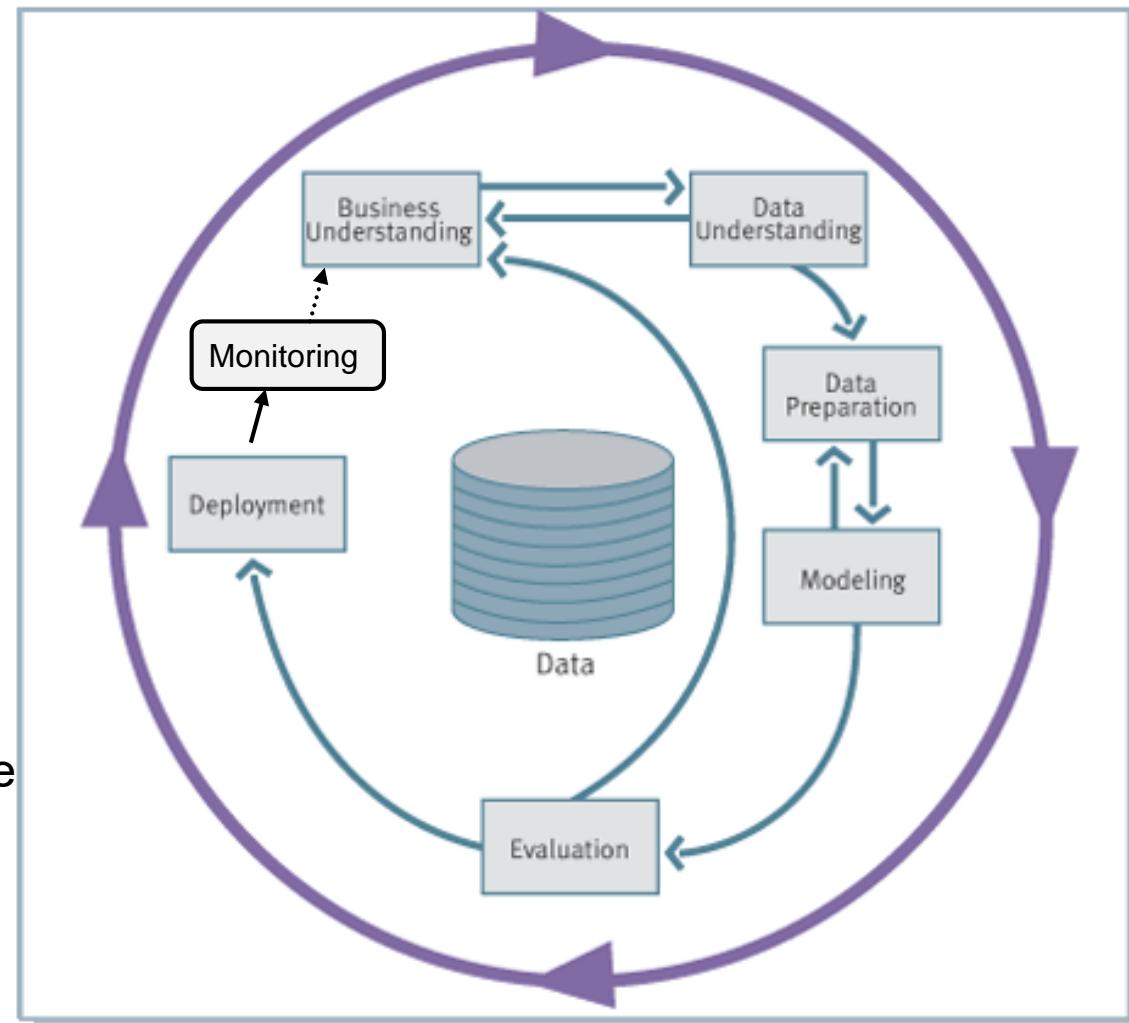
Периодическая система элементов Д.И. Менделеева.

Сегментация в маркетинге. Критериями сегментации являются: географическое местоположение, социально-демографические характеристики, мотивы совершения покупки и т. п.

На основании результатов сегментации маркетолог может определить, например, такие характеристики сегментов рынка, как реальная и потенциальная емкость сегмента, группы потребителей, чьи потребности не удовлетворяются в полной мере ни одним производителем, работающим на данном сегменте рынка, и т. п.

Главное задание Data Mining: найти истинные закономерности и избежать переобучения

Переобучение (overfitting) в машинном обучении и статистике - явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительно плохо работает на примерах, не участвовавших в обучении (на примерах из тестовой выборки).
Это связано с тем, что при построении модели («в процессе обучения») в обучающей выборке обнаруживаются некоторые случайные закономерности, которые отсутствуют в генеральной совокупности.



Практическое применение Data Mining

Интернет-технологии

- персонализация посетителей Web-сайтов
- поиск случаев мошенничества с кредитными картами
- Web Mining: Web content mining и Web usage mining

Торговля

- анализ рыночных корзин и сиквенциональный анализ

Телекоммуникации

- анализ доходности и риска потери клиентов
- защита от мошенничества,
- выявление категорий клиентов с похожими стереотипами пользования услугами и разработка привлекательных наборов цен и услуг

Практическое применение Data Mining

Промышленное производство

прогнозирование качества изделия в зависимости от замеряемых параметров технологического процесса.

Медицина и биология

построение диагностической системы

исследование эффективности хирургического вмешательства

Биоинформатика – изучение генов, разработка новых лекарств

Банковское дело

оценка кредитоспособности заемщика

Модели Data Mining

Предсказательные модели

модели классификации

модели последовательностей

Описательные модели

регрессионные модели

модели кластеров

модели исключений

итоговые модели

ассоциативные модели

Предсказательные модели

модели классификации описывают правила или набор правил, в соответствии с которыми можно отнести описание любого нового объекта к одному из классов

Такие правила строятся на основании информации о существующих объектах путем разбиения их на классы;

модели последовательностей описывают функции, позволяющие прогнозировать изменение непрерывных числовых параметров.

Они строятся на основании данных об изменении некоторого параметра за прошедший период времени.

Описательные модели

регрессионные модели описывают функциональные зависимости между зависимыми и независимыми показателями и переменными в понятной человеку форме.

описывают функциональную зависимость не только между непрерывными числовыми параметрами, но и между категориальными параметрами;

модели кластеров описывают группы (кластеры), на которые можно разделить объекты, данные о которых подвергаются анализу. Группируются объекты (наблюдения, события) на основе данных (свойств), описывающих сущность объектов.

объекты внутри кластера должны быть "похожими" друг на друга и отличаться от объектов, вошедших в другие кластеры.

Чем сильнее "похожи" объекты внутри кластера и чем больше отличий между кластерами, тем точнее кластеризация;

Описательные модели

Модели исключений описывают исключительные ситуации в записях (например, отдельных пациентов), которые резко отличаются чем либо от основного множества записей (группы больных).

Знание исключений может быть использовано двояким образом. Возможно, эти записи представляют собой случайный сбой, например ошибки операторов, вводивших данные в компьютер.

С другой стороны, отдельные исключительные записи могут представлять самостоятельный интерес для исследования, т. к. они могут указывать на некоторые редкие, но важные аномальные заболевания.

Описательные модели

Итоговые модели - выявление ограничений на данные анализируемого массива.

Например, при изучении выборки данных по пациентам не старше 30 лет, перенесшим инфаркт миокарда, обнаруживается, что все пациенты, описанные в этой выборке, либо курят более 5 пачек сигарет в день, либо имеют вес не ниже 95 кг

Построение итоговых моделей заключается в нахождении каких либо фактов, которые верны для всех или почти всех записей в изучаемой выборке данных, но которые достаточно редко встречались бы во всем мыслимом многообразии записей;

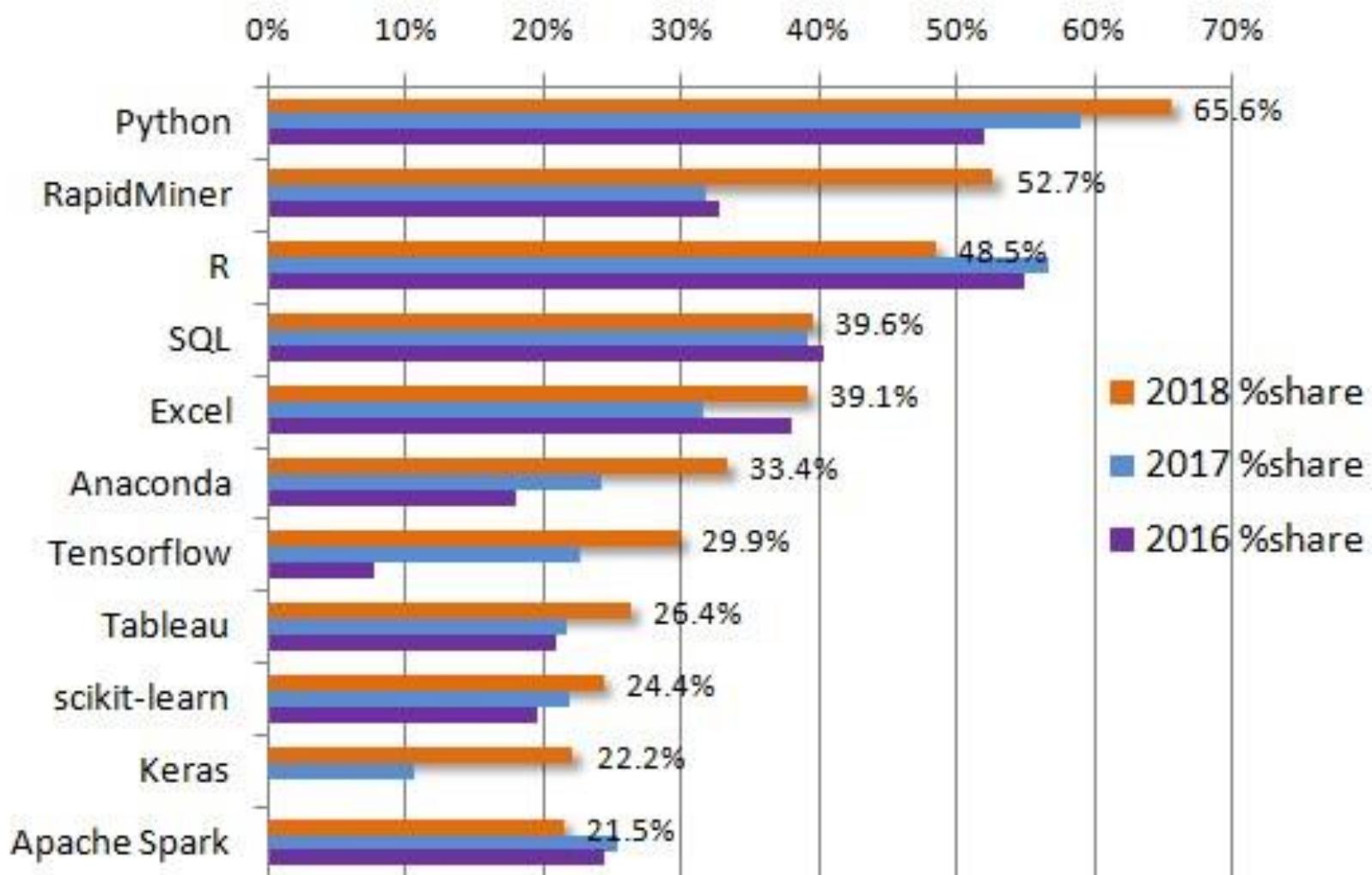
ассоциативные модели - выявление закономерностей между связанными событиями.

Количественный анализ:

это манипуляции с измеренными
характеристиками изучаемого объекта

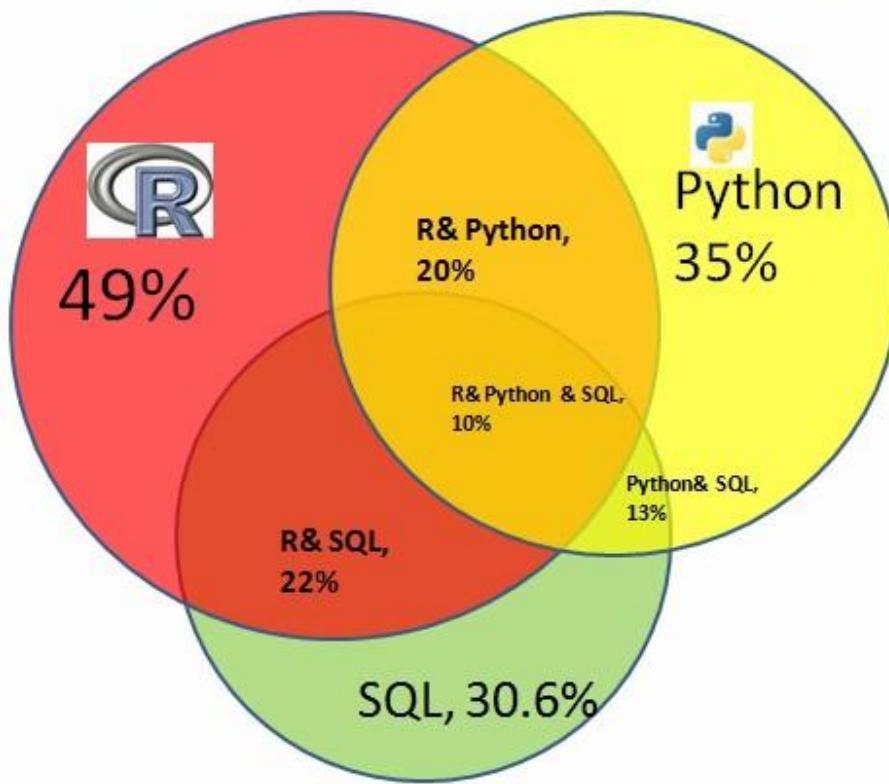
- направлен в основном на формализованное, «внешнее» изучение объекта, анализ его измеряемых признаков;
- основным итогом является упорядоченная совокупность «внешних», измеряемых признаков объекта;
- реализуется при помощи математико-статистических методов;
- доминирует аналитическая составляющая познания.

KDnuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018

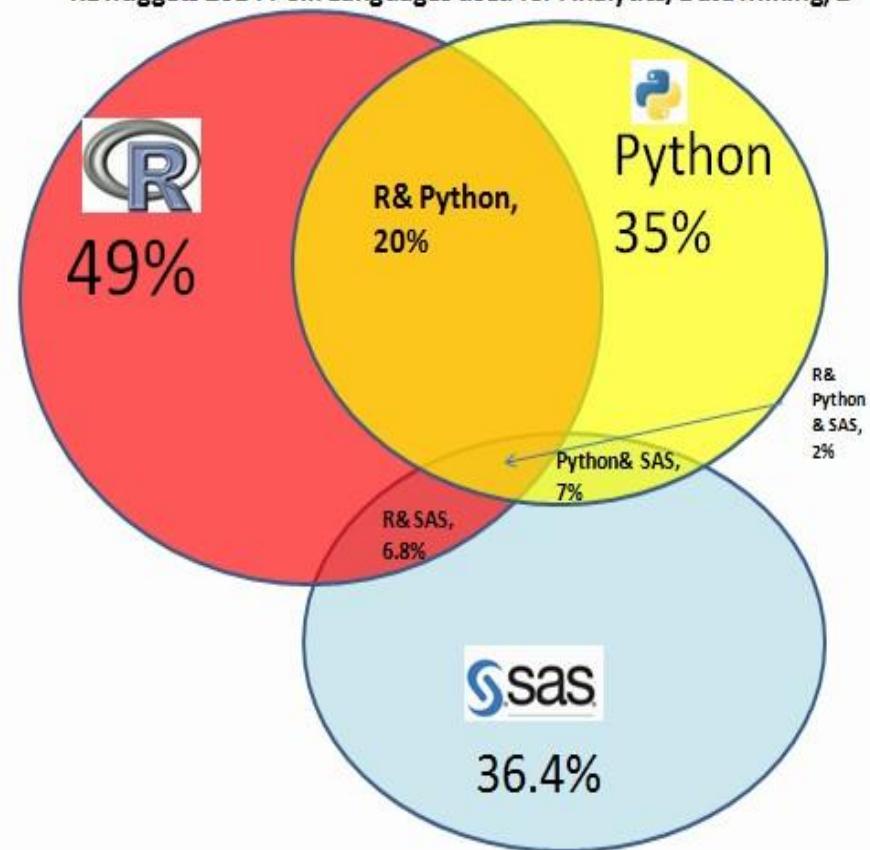


Консолидация среди топ-4 языков

KDnuggets 2014 Poll: Languages used for Analytics/Data Mining



KDnuggets 2014 Poll: Languages used for Analytics/Data Mining, 2



Основные понятия и сведения из истории создания пакета R

Термин R используется в двух значениях:

- (интерпретируемый) язык программирования для статистической обработки данных
- программная среда вычислений.

Год разработки **1993**.

Разработчики - сотрудники Оклендского университета (Новая Зеландия)

Росс Айхэка ([Ross Ihaka](#))

Роберт Джентлмен ([Robert Gentleman](#)).

Название языка и среды вычислений - первая буква имён разработчиков.

Язык и среда R широко используются как статистическое программное обеспечение для анализа данных - *фактический стандарт программного обеспечения для статистической обработки информации*.

В 2010 году R вошёл в список победителей конкурса журнала **InfoWorld** в номинации на *лучшее открытое программное обеспечение для разработки приложений*.

Полезные ссылки:

1. «KDnuggets»: <https://www.kdnuggets.com/>

2. Ссылки для скачивания дистрибутива (R):

The Comprehensive R Archive Network - <https://cran.r-project.org/>

The R Project for Statistical Computing - <https://www.r-project.org/>

 R Studio - <https://www.rstudio.com/products/rstudio/download/#download>

3. Роберт И. Кабаков. R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. - М.: ДМК Пресс, 2014. - 588 с. -

<http://kek.ksu.ru/eos/WM/Kabacoff2014ru.pdf>

4. DATA MINING FOR BUSINESS ANALYTICS. Concepts, Techniques, and Applications in R (Galit Shmueli et. al)

https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR_Brett_Lantz.pdf

5. Data Science Central - <https://www.datasciencecentral.com/>

Лекция 3-4

дисперсионный анализ

Дисперсионный анализ. Постановка задачи

Дисперсионный анализ как метод исследования появился в работах Р. Фишера (1918-1935 гг.) в связи с исследованиями в сельском хозяйстве для выявления условий, при которых испытываемый сорт сельскохозяйственной культуры даёт максимальный урожай. (В агрономических исследованиях первый фактор - сорт почвы, второй фактор - способ обработки.)

Дальнейшее развитие дисперсионный анализ получил в работах Йетса.

Сейчас теорию дисперсионного анализа можно считать в достаточной мере сформировавшейся, но способы организации эксперимента и вычислительные схемы продолжают совершенствоваться.

Дисперсионный анализ

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящий от разных, одновременно действующих факторов, выбора наиболее важных факторов и оценки их влияния.

Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия.

Последующее сравнение этих дисперсий позволяет оценить существенность влияния факторов на исследуемую величину.

Дисперсионный анализ. Постановка задачи

*В любом ряде испытаний имеется несколько факторов, вызывающих изменчивость средних значений наблюдаемых случайных величин - **результативных признаков**.*

Эти факторы могут принадлежать одному или нескольким источникам изменчивости (например, расположение торговых заведений в центре и на окраине города, изменения в законодательстве, разные климатические условия, разные уровни образования и т. п.).

Даже при самом тщательном исследовании не удается выявить все источники изменчивости, а иногда в этом нет необходимости или смысла.

Но при наличии опыта у эксперта и в зависимости от цели исследования всегда можно выдвинуть гипотезу о существовании влияния тех или иных факторов на результативный признак.

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе используются следующие термины:

фактор (X) – то, что, как мы считаем, должно оказывать влияние на результат (результативный признак) Y ;

уровень фактора (или способ обработки, иногда буквально, например - способ обработки почвы) - значения (X_i , $i = 1, 2, \dots, I$), которые может принимать фактор;

отклик – значение измеряемого признака (величина результата Y_i).

Техника дисперсионного анализа меняется в зависимости от числа изучаемых независимых факторов.

Если факторы, вызывающие изменчивость среднего значения признака, принадлежат одному источнику, то мы имеем простую группировку, или однофакторный дисперсионный анализ, и далее, соответственно, двойная группировка - двухфакторный дисперсионный анализ, трехфакторный дисперсионный анализ, m -факторный.

Факторы в многофакторном анализе принято обозначать латинскими буквами: A, B, C и т. д.

Дисперсионный анализ. Постановка задачи

Задача дисперсионного анализа - исследование влияния тех или иных факторов (или уровней факторов) на изменчивость средних значений наблюдаемых случайных величин.

Сущность дисперсионного анализа. Дисперсионный анализ состоит в выделении и оценке отдельных факторов, вызывающих изменчивость.

С этой целью производят разложение общей дисперсии σ^2 наблюданной частичной совокупности (общей дисперсии признака), вызванной всеми источниками изменчивости, на составляющие дисперсий, порожденные независимыми факторами. Каждая из этих составляющих дает оценку дисперсии $\sigma^2_A, \sigma^2_B, \dots$, вызванную конкретным источником изменчивости, в общей совокупности.

Для проверки значимости этих составляющих оценок дисперсии их сравнивают с общей дисперсией в общей совокупности (по критерию Фишера).

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе рассматривается гипотеза:

H_0 – ни один из рассматриваемых факторов не оказывает влияния на изменчивость признака.

Значимость каждой из оценок дисперсии проверяется по величине её отношения к оценке случайной дисперсии и сравнивается с соответствующим критическим значением, при уровне значимости α , с помощью таблиц критических значений F-распределения Фишера - Снедекора (табулировано).

Гипотеза H_0 относительно того или иного источника изменчивости отвергается, если $F_{\text{расч.}} > F_{\text{кр.}}$

Дисперсионный анализ. Постановка задачи

В дисперсионном анализе рассматриваются эксперименты 3-х видов:

- а) эксперименты, в которых *все факторы имеют систематические (фиксированные) уровни;*
- б) эксперименты, в которых *все факторы имеют случайные уровни;*
- в) эксперименты, в которых *есть факторы, имеющие случайные уровни, а также факторы, имеющие фиксированные уровни.*

Случай а), б), в) соответствуют трем моделям, которые рассматриваются в дисперсионном анализе.

Применение дисперсионного анализа предполагает, что:

$$M(\varepsilon_{ij})=0, \quad D(\varepsilon_{ij})=\sigma^2=\text{const}, \quad \varepsilon_{ij} \rightarrow N(0, \sigma^2) \text{ или } x_{ij} \rightarrow N(a, \sigma^2).$$

ε_{ij} - вариация результатов внутри отдельного уровня фактора.

Однофакторный дисперсионный анализ

Предположим, что совокупности случайных величин имеют нормальное распределение и равные дисперсии.

Пусть имеется m таких совокупностей, из которых произведены выборки объемом n_1, n_2, \dots, n_m . Обозначим выборку из i -ой совокупности $(x_{i1}, x_{i2}, \dots, x_{in})$.

Тогда все выборки можно записать в виде следующей таблице, которая называется *матрицей наблюдений*.

Матрица наблюдений

Количество совокупностей (n)	Количество элементов совокупности						
	1	2	...	j	...	n	
Количество совокупностей (m)							
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1n1}	
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2n2}	
...	
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ini}	
...	
m	X_{m1}	X_{m2}	...	X_{mj}	...	X_{mm}	

Средние выборок обозначим через $\beta_1, \beta_2, \dots, \beta_m$.

Проверим нулевую гипотезу о равенстве средних:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_m.$$

Гипотеза H_0 проверяется сравнением внутригрупповых и межгрупповых дисперсий по F-критерию Фишера.

Если расхождение между ними значительно, то нулевая гипотеза принимается.

В противном случае гипотеза о равенстве средних отвергается и делается заключение о том, что различие в средних обусловлено не только случайностями выборок, но и действием исследуемого фактора.

Рассмотрим структуру межгрупповой и внутригрупповой дисперсии.

Для этого найдем сначала средние арифметические членов каждой совокупности:

$$\bar{x}_{i^*} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}; \dots \bar{x}_{ij} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}; \dots \bar{x}_{mj} = \frac{\sum_{j=1}^{n_m} x_{mj}}{n_m}.$$

Общую среднюю арифметическую всех m совокупностей рассчитываем по формуле:

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m \bar{x}_{i^*}.$$

Найдем сумму квадратов отклонений x_{ij} от \bar{x} :

$$\underbrace{\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2}_{Q} = \underbrace{n \sum_{i=1}^m (\bar{x}_{i*} - \bar{x})^2}_{Q_1} + \underbrace{\sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i*})^2}_{Q_2},$$

Слагаемое Q_1 является суммой квадратов разностей между средними отдельных совокупностей и общей средней всей совокупности наблюдений.

Эта сумма называется *суммой квадратов отклонений между группами* и характеризует систематическое расхождение между совокупностями наблюдений.

Величину Q_1 называют иногда *рассеиванием по факторам* (т.е. за счет исследуемого фактора).

Слагаемое Q_2 представляет собой сумму квадратов разностей между отдельными наблюдениями и средней соответствующей совокупности.

Эта сумма называется *суммой квадратов отклонений внутри группы*.

Она характеризует *остаточное рассеивание* случайной погрешности совокупностей.

Наконец, Q называется *общей* или *полной суммой квадратов отклонений отдельных наблюдений от общей средней*.

Оценим дисперсии:

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{x}_{i*} - \bar{x})^2 = \frac{Q_1}{m-1},$$

$$s_2^2 = \frac{1}{m(n-1)} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i*})^2 = \frac{Q_2}{m(n-1)},$$

$$s^2 = \frac{1}{mn-1} \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x})^2.$$

Произведем оценку различия между дисперсиями по F-критерию:

$$F = \frac{\frac{Q_1}{(m-1)}}{\frac{Q_2}{m(n-1)}}.$$

Если полученное значение критерия больше табличного с заданным уровнем значимости α и числом степенями свободы $(m-1; m(n-1))$, то нулевую гипотезу отвергаем, т.е. фактор влияет на исследуемую величину.

Таблица однофакторного дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат	Оценка дисперсий
Межгрупповая	$m \sum_i (\bar{x}_{i*} - \bar{x})^2$	$m-1$	$\frac{1}{m-1} \sum_i (\bar{x}_{i*} - \bar{x})^2$	s_1^2
Внутригрупповая	$\sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	$m(n-1)$	$\frac{1}{m(n-1)} \sum_{ij} (x_{ij} - \bar{x}_{i*})^2$	s_2^2
Полная (общая)	$\sum_{ij} (x_{ij} - \bar{x})^2$	$mn-1$	$\frac{1}{mn-1} \sum_{ij} (x_{ij} - \bar{x})^2$	s^2

Далее необходимо рассчитать доли влияния учтенного и неучтенного факторов как отношения соответствующих сумм квадратов отклонений:

$$\eta_1^2 = \frac{Q_1}{Q}; \eta_2^2 = \frac{Q_2}{Q}; \eta_1^2 + \eta_2^2 = 1(100%),$$

где η_1^2 - доля влияния учтенных факторов;

η_2^2 - доля влияния неучтенных факторов.

Пример. В трех различных местах обитания были собраны жуки-скакуны. У каждого жука измерялась ширина головки. Требуется с помощью дисперсионного анализа выяснить, влияет ли место обитания на ширину головки жука (данные приведены в таблице).

Вначале подсчитаем средние значения:

$$\bar{x}_{1*} = 3,76; \bar{x}_{2*} = 3,6137; \bar{x}_{3*} = 3,5635; \bar{x} = 3,6457.$$

$$m=3, n=10.$$

Таблица. Значения ширины головки жука в месте обитания

Мест- ность	Измерения ширины головки жука, мм									
	1	2	3	4	5	6	7	8	9	10
A1	3,712	3,732	3,752	3,762	3,769	3,775	3,782	3,787	3,792	3,737
A2	3,602	3,605	3,607	3,61	3,612	3,615	3,617	3,62	3,622	3,627
A3	3,532	3,537	3,542	3,549	3,555	3,562	3,572	3,582	3,592	3,612

Рассчитываем суммы квадратов:

$$Q_1 = n \sum_{i=1}^m (\bar{x}_{i*} - \bar{x})^2 = 0,20845; k_1 = m - 1 = 2;$$

$$Q_2 = \sum_{i=1}^3 \sum_{j=1}^{10} (x_{ij} - \bar{x}_{i*})^2 = 0,01282; k_2 = m(n - 1) = 27;$$

$$Q = Q_1 + Q_2 = 0,22127; k = mn - 1 = 29.$$

Оценим дисперсии:

$$s_1^2 = \frac{Q_1}{k_1} = 0,104225; s_2^2 = \frac{Q_2}{k_2} = 0,000475;$$

$$s^2 = Q/k = 0,22127/29 = 0,00763.$$

Значение критерия Фишера равно

$$F = \frac{s_1^2}{s_2^2} = 219,421.$$

Табличное значение критерия при уровне значимости $\alpha=0,05$ равно 3,354.

Так как полученное значение критерия Фишера больше табличного ($219,421 > 3,354$), то гипотезу о том, что место обитания не влияет на размеры головки жука, отвергаем.

Рассчитаем доли влияния учтенного и неучтенных факторов:

$$\eta_1^2 = \frac{Q_1}{Q} = \frac{0,208453}{0,221274} = 0,942;$$

$$\eta_2^2 = \frac{Q_2}{Q} = \frac{0,012821}{0,221274} = 0,0579.$$

На долю учтенного фактора – место обитания приходится 94,2% изменчивости, а 5,79% составляют неучтенные факторы.

Таким образом, место обитания оказывает существенное влияет на размеры головки жуков-скакунов.

Многофакторный дисперсионный анализ

Если исследуют действие двух, трех и т.д. факторов, то структура дисперсионного анализа та же, что и при однофакторном анализе, усложняются лишь вычисления. Рассмотрим задачу оценки действия двух одновременно действующих факторов.

Введем некоторые ограничения:

- включаемые в анализ факторы должны быть независимы друг от друга, корреляция между ними не допустима;
- число наблюдений по совокупностям должно быть одинаковым или хотя бы пропорциональным.

Пусть имеется несколько разнотипных участков земли и несколько типов удобрения.

Требуется выяснить, значимо ли влияние качества различных участков земли и качества удобрений на урожайность зерновой культуры.

Пусть фактор А – влияние земли, фактор В – влияние качества удобрения, x_{ij} – урожайность.

Рассмотрим случай, когда для каждого участка земли и для каждого вида удобрения сделано одно наблюдение.

Тогда матрица наблюдений будет следующей:

Участки земли (i)	Вид удобрения(j)		...	B_v	...
	B_1	B_2			
A_1	x_{11}	x_{12}	...	x_{1v}	x_{1*}
A_2	x_{21}	x_{22}	...	x_{2v}	x_{2*}
...
A_r	x_{r1}	x_{r2}	...	x_{rv}	x_{r*}
\bar{x}_{*j}	x_{*1}	x_{*1}		$x_{*\nu}$	\bar{x}

По каждому столбцу и строке вычислим среднее значение, а также общее среднее.

$$\bar{x}_{i*} = \frac{1}{v} \sum_{j=1}^v x_{ij}, \quad \bar{x}_{*j} = \frac{1}{r} \sum_{i=1}^r x_{ij}, \quad \bar{x} = \frac{1}{rv} \sum_{i=1}^r \sum_{j=1}^v x_{ij}.$$

Основное тождество однофакторного анализа в данном случае принимает вид:

$$\sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x})^2 = v \sum_{i=1}^r (\bar{x}_{i*} - \bar{x})^2 + r \sum_{j=1}^v (\bar{x}_{*j} - \bar{x})^2 + \sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2,$$

$$Q = Q_1 + Q_2 + Q_3.$$

Слагаемое Q_1 представляет собой сумму квадратов разностей между средними по строкам и общим средним и характеризует изменение признака по фактору А.

Слагаемое Q_2 представляет собой сумму квадратов разностей между средними по столбцам и общим средним и характеризует изменение признака по фактору В.

Слагаемое Q_3 называется *остаточной суммой квадратов* и характеризует влияние неучтенных факторов.

Сумма Q называется *общей или полной суммой квадратов отклонений* отдельных наблюдений от общей средней.

Произведем оценку дисперсий:

$$s^2 = \frac{1}{rv-1} \sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x})^2 = \frac{Q}{rv-1},$$

$$s_1^2 = \frac{1}{r-1} v \sum_{i=1}^r (\bar{x}_{i*} - \bar{x})^2 = \frac{Q_1}{r-1},$$

$$s_2^2 = \frac{1}{v-1} r \sum_{j=1}^v (\bar{x}_{*j} - \bar{x})^2 = \frac{Q_2}{v-1},$$

$$s_3^2 = \frac{1}{(r-1)(v-1)} \sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2 = \frac{Q_3}{(r-1)(v-1)}.$$

В двухфакторном анализе для выяснения значимости влияния факторов А и В на исследуемый признак сравнивают дисперсии по факторам с остаточной дисперсией.

$$F_A = \frac{Q_1 / (r - 1)}{Q_3 / (r - 1)(v - 1)} = \frac{s_1^2}{s_3^2},$$

$$F_B = \frac{Q_2 / (v - 1)}{Q_3 / (r - 1)(v - 1)} = \frac{s_2^2}{s_3^2}.$$

Полученные значения F_A и F_B сравнивают с табличными значениями при выбранном уровне значимости α и соответствующих числах степеней свободы.

При $F_A < F_{\alpha}$ и $F_B < F_{\alpha}$ нулевые гипотезы о равенстве средних не отвергается, т.е. влияние факторов А и В на исследуемый признак незначимо.

Для расчета доли влияния учтенных факторов А, В и неучтенного фактора воспользуемся формулами:

$$\eta_A^2 = \frac{s_1^2}{S^2}; \quad \eta_B^2 = \frac{s_2^2}{S^2}; \quad \eta^2 = \frac{s_3^2}{S^2}.$$

Результаты анализа заносятся в таблицу дисперсионного анализа.

Таблица двухфакторного дисперсионного анализа

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсий
Между средними по строкам	$Q_1 = v \sum_{i=1}^r (\bar{x}_{i*} - \bar{x})^2$	$r-1$	s_1^2
Между средними по столбцам	$Q_2 = r \sum_{j=1}^v (\bar{x}_{*j} - \bar{x})^2$	$v-1$	s_2^2
Остаточная	$Q_3 = \sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2$	$(r-1)(v-1)$	s_3^2
Полная (общая)	$Q = \sum_{i=1}^r \sum_{j=1}^v (x_{ij} - \bar{x})^2$	$rv-1$	s^2

Алгоритм расчетов

1. Построение вспомогательной таблицы.
2. Вычисление средних.
3. Вычисление сумм квадратов.
4. Вычисление оценок дисперсий.
5. Проверка гипотезы H_0 . Если H_0 не отклоняется, то – проверка значимости уровней факторов.

Лекция 4-5

МЕТОДЫ КЛАСТЕРНОГО АНАЛИЗА

Классификация методов анализа данных



Выбор метода решения зависит прежде всего от того, являются качественными или количественными зависимые переменные.

Окончательно решение о выборе метода анализа данных принимается в зависимости от типа независимых переменных.

Кластерный анализ. Основные понятия

Для поиска качественных факторов применяется группа методов, известная под названием ***кластерный анализ***.

Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы «схожих» объектов, называемых кластерами.

Методы кластеризации довольно разнообразны, в них по-разному выбирается способ определения близости между кластерами (и между объектами), а также используются различные алгоритмы вычислений.

Поэтому результаты кластеризации зависят от выбранного метода.

Особенности кластерного анализа

В отличие от задач классификации, *кластерный анализ* не требует априорных предположений о наборе данных, не накладывает ограничения на *представление* исследуемых объектов, позволяет анализировать показатели различных типов данных (*интервальным данным, частотам, бинарным данным*).

При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной.

Особенности кластерного анализа

Кластерный анализ может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Кластерный анализ параллельно развивался в нескольких направлениях, таких как биология, психология, др., поэтому у большинства методов существует по два и более названий. Это существенно затрудняет работу при использовании кластерного анализа.

История возникновения

1. Концепция классификации и систематизации, предложенная французским ботаником **Огюстеном Декандолем (1778-1841)** в 1813 году с целью систематизации растений. Данная теория получила наименование таксономия.
2. Статья польского антрополога **Яна Чекановского**, которую он написал в 1911 году. В своей работе он показывает идею «структурной классификации», содержащую главную мысль кластерного анализа – выделение компактных групп близких объектов, а также некоторые методы выделения таких групп, которые лежат в основе более последних алгоритмов.
3. «Метод корреляционных плеяд», созданный советским гидробиологом **П.В. Терентьевым** в 1925 году. Однако издан он был лишь через много лет в 1959 г.
4. Сам термин «кластерный анализ» был впервые введен и использован только в 1939 году английским ученым **Р. Трионом**

Кластерный анализ.

Основные задачи

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи.

- **Упростить дальнейшую обработку данных**, разбить множество на группы схожих объектов чтобы работать с каждой группой в отдельности.
- **Сократить объём хранимых данных**, оставив по одному представителю от каждого кластера (задачи сжатия данных).
- **Выделить нетипичные объекты**, которые не подходят ни к одному из кластеров.
- **Построить иерархию множества объектов.**

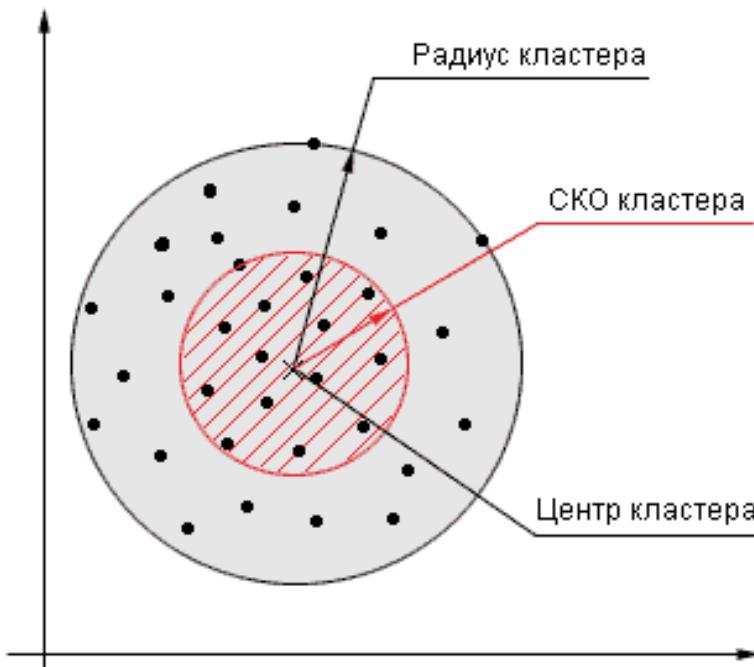
Кластерный анализ. Основные понятия

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.
4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Кластерный анализ. Основные понятия

Кластерный анализ (или **кластеризация**) – задача разбиения заданной выборки **объектов** (ситуаций) на непересекающиеся подмножества, называемые **кластерами**, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.



Кластер имеет следующие **математические характеристики**: **центр**, **радиус**, **среднеквадратическое отклонение**, **размер кластера**.

Центр кластера - это среднее геометрическое место точек в пространстве переменных.

Радиус кластера - максимальное расстояние точек от **центра кластера**.

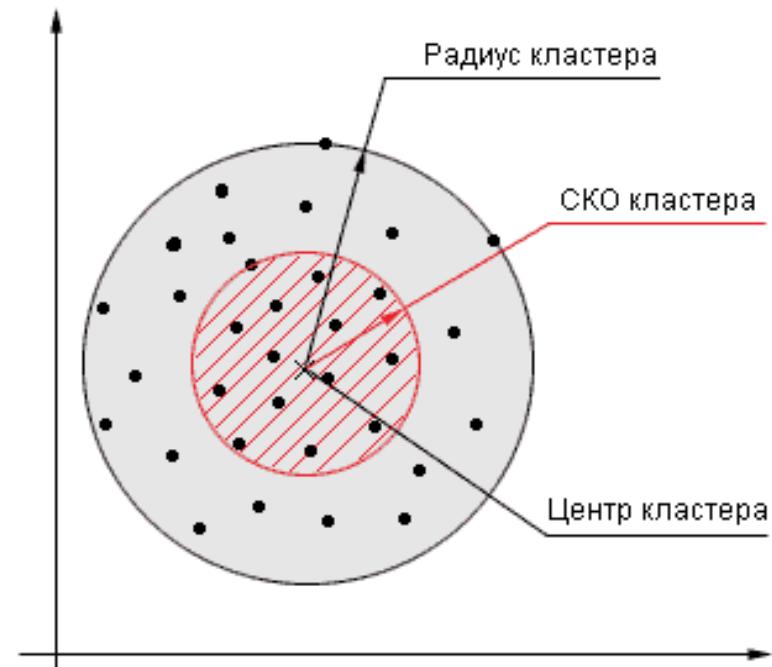
Размер кластера может быть определен либо по **радиусу кластера**, либо по **среднеквадратичному отклонению** объектов для этого кластера.

Кластерный анализ. Основные понятия

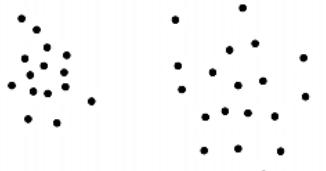
Для определения сходства («близости») объектов используются различные метрики.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

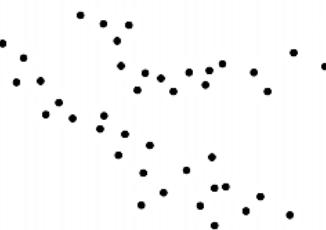
Объект относится к кластеру, если *расстояние от объекта до центра кластера* меньше *радиуса кластера*. Если условие выполняется для двух и более кластеров, объект является спорным.



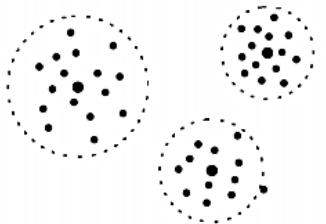
Типы кластерных структур*



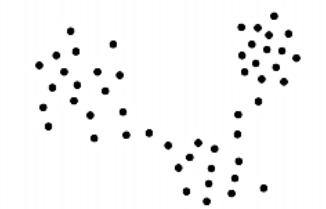
Сгущения: внутрикластерные расстояния, как правило, меньше межкластерных.



Ленты: для любого объекта найдётся близкий к нему объект того же кластера, в то же время существуют объекты одного кластера, которые не являются близкими.



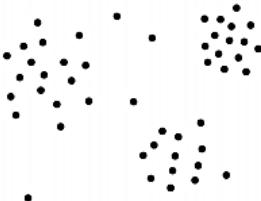
Кластеры с центром: в каждом кластере найдётся объект, такой, что почти все объекты кластера лежат внутри шара с центром в этом объекте.



Кластеры могут соединяться перемычками, что затрудняет работу многих алгоритмов кластеризации.

* Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: <http://www.ccas.ru/voron>

Типы кластерных структур*



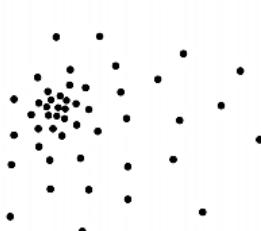
Кластеры могут накладываться на разреженный фон из редких нетипичных объектов.



Кластеры могут перекрываться.



Кластеры могут образовываться не по принципу сходства, а по каким-либо иным, заранее неизвестным, свойствам объектов. Стандартные методы кластеризации здесь бессильны.



Кластеры могут вообще отсутствовать. В этом случае надо применять не кластеризацию, а иные методы анализа данных.

*Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: <http://www.ccas.ru/voron>

Кластерный анализ. Основные понятия

Понятие «расстояние между объектами» является интегральной мерой сходства объектов между собой.

Расстоянием между объектами в пространстве признаков называется такая величина d_{ij} , которая удовлетворяет следующим аксиомам:

$d_{ij} > 0$ (неотрицательность расстояния)

$d_{ij} = d_{ji}$ (симметрия)

$d_{ij} + d_{jk} > d_{ik}$ (неравенство треугольника)

Если d_{ij} не равно 0, то i не равно j (различимость нетождественных объектов)

Если $d_{ij} = 0$, то $i = j$ (неразличимость тождественных объектов)

Меру близости (сходства) объектов удобно представить как обратную величину от расстояния между объектами.

Кроме термина "расстояние" в литературе часто встречается и другой термин - "метрика", который подразумевает метод вычисления того или иного конкретного расстояния.

Метрики кластеризации

- 1) евклидово расстояние или евклидова метрика $d_{ij} = \left(\sum_{k=1}^v (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$
- 2) квадрат евклидова расстояния $d_{ij}^2 = \sum_{k=1}^v (x_{ik} - x_{jk})^2$
- 3) степенное расстояние
(обобщенная метрика Минковского)
$$d_{ij} = (\sum_{k=1}^v |x_{ik} - x_{jk}|^p)^{\frac{1}{p}}$$
- 4) расстояние городских кварталов
(манхэттенское расстояние, city-block)
$$d_{ij} = \sum_{k=1}^v |x_{ik} - x_{jk}|$$
- 5) метрика доминирования
(Sup-метрика,
расстояние Чебышева)
$$d_{ij} = \max |x_{ik} - x_{jk}|.$$

Метрики кластеризации

6) расстояние Махalanобиса $d_{ij} = (X_i - X_j)^T S^{-1} (X_i - X_j)$

В отличие от метрики Минковского и евклидовой метрики, расстояние Махalanобиса через матрицу дисперсий-ковариаций S связано с корреляциями переменных.

Когда корреляции между переменными равны нулю, расстояние Махalanобиса эквивалентно квадрату евклидового расстояния.

7) расстояние Хемминга $d_{ij} = \sum_{k=1}^v |x_{ik} - x_{jk}|$

Применяется в случае использования дихотомических (имеющих всего два значения) качественных признаков, равно числу несовпадений значений соответствующих признаков для рассматриваемых i-того и j-того объектов.

От выбора метрики во многом зависит результат кластеризации.
Выбор осуществляется в зависимости от пространства, в котором расположены объекты и неявных характеристик кластеров.

Общая методология кластеризации



[Jain, Dubes. Algorithms for clustering data (Prentice-Hall, 1988)]

Общая методология кластеризации

Сбор данных: получение «сырых» данных из различных источников.

Первоначальный отбор: подготовка данных к анализу, нормализация. Выявление данных, которые будут мешать дальнейшему анализу, например, незначащие характеристики, дубликаты, противоречия.

Представление: перевод данных в форму, пригодную для дальнейшего анализа.

Тенденция кластеризации: выявление неслучайной структуры в данных. Если данные не имеют тенденцию к кластеризации, то выбирается другая техника анализа данных.

Стратегия кластеризации: выбор соответствующего метода (иерархический\неиерархический) и затем алгоритма. Внимание должно быть уделено соответствуанию алгоритма конкретным данным.

Валидация: сравнение с данными, полученными «извне»; сравнение с данными, полученными при работе других алгоритмов.

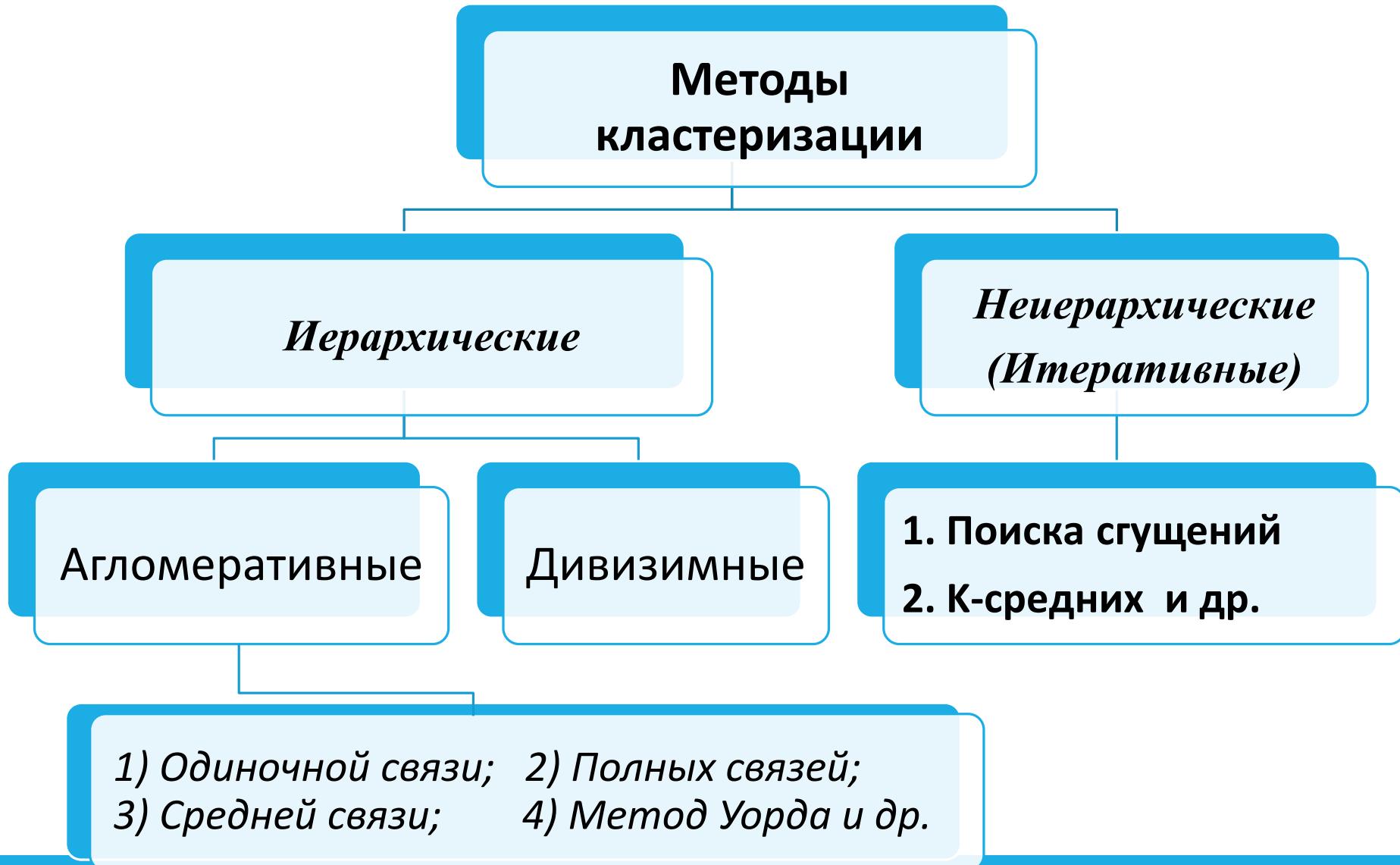
Интерпретация: графическое представление результатов кластерного анализа.

Этапы кластеризации



- **Выявление вектора характеристик:** выбор наиболее эффективных подмножеств характеристик или создание новых характеристик путем трансформации существующих.
- **Выбор метрики:** выбор меры расстояний для определения «близости» объектов. Выбор осуществляется в зависимости от пространства, в котором расположены объекты и неявных характеристик кластеров.
- **Разбиение объектов на кластеры:** выполняется в соответствии с выбранным алгоритмом. Производится изменение метрики или вектора характеристик при неудовлетворительном результате разбиения.

Кластерный анализ. Классификация методов



Кластерный анализ. Классификация методов

Иерархические (Hierarchical) — построение дендограммы (дерево вложенных кластеров)

Агglomerативные (Agglomerative) — в начале работы алгоритма количество кластеров равно количеству объектов, далее итерационно «снизу-вверх» ближайшие два объединяются

Дивизимные (Divisive) — в начале работы алгоритма все объекты относятся к одному кластеру, далее итерационно «сверху-вниз» каждый кластер делится на два

Дендрограмма (dendrogram)

Иерархические алгоритмы связаны с построением *дендрограмм*, которые являются результатом *иерархического кластерного анализа*.

Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (*dendrogram*) - древовидная диаграмма, содержащая *n* уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Дендрограмма (dendrogram)

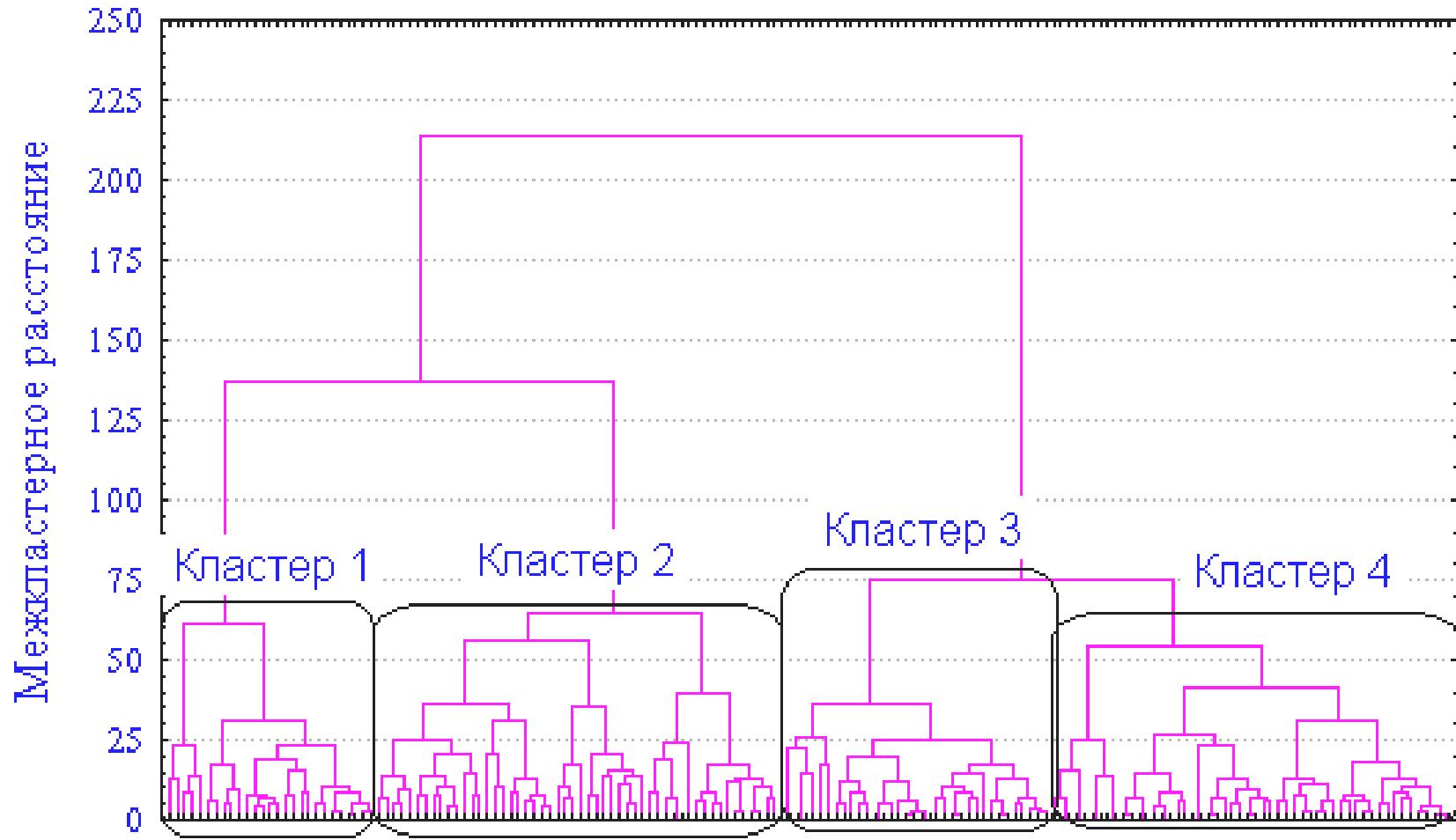
Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

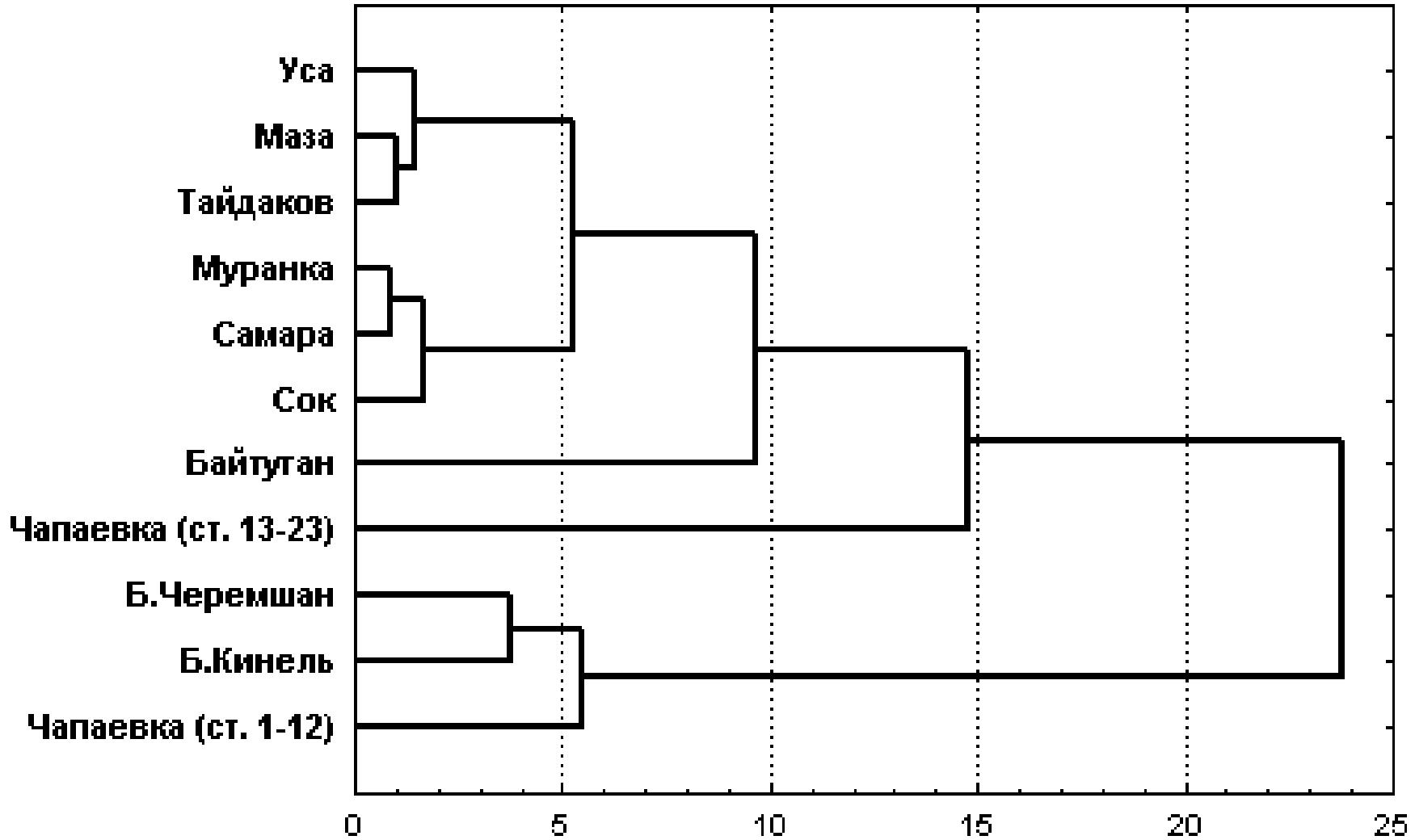
Существует много способов построения *дендрограмм*. В *дендрограмме* объекты могут располагаться вертикально или горизонтально.

Пример горизонтальной дендрограммы

Дендрограмма наблюдений за 158 крысами



Пример вертикальной дендрограммы (Кластеризация рек Самарской области)



Методы связи:

включают методы «ближайшего соседа», «далекого соседа» и «среднего расстояния»

Метод «ближайшего соседа» (одиночная связь) первыми объединяют два объекта, расстояние между которыми минимально.

Далее определяют следующее по величине самое короткое расстояние, и в кластер с первыми двумя объектами вводят третий объект.

Расстояние между кластерами – расстояние между их ближайшими точками.

В методе «далекого соседа» (полная связь) расстояния между кластерами вычисляют как расстояния между их самыми удаленными точками

В методе «среднего расстояния» расстояние между кластерами определяют, как среднее значение всех расстояний между объектами двух кластеров.

Метод Ворда

В этом методе *в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений*, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект.

На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов.

Этот метод направлен на объединение близко расположенных кластеров.

MST (Minimum Spanning Trees)

- **иерархический дивизимный** алгоритм, основанный на теории графов.

Основная идея алгоритма: строится минимальное оставное дерево на основе множества всех исходных объектов, далее дерево делится на кластеры.

Минимальное оставное дерево (или минимальное покрывающее дерево) в связном, взвешенном, неориентированном графе — это оставное дерево этого графа, имеющее минимальный возможный вес, где под весом дерева понимается сумма весов входящих в него рёбер.

Достоинства: алгоритм выделяет кластеры произвольной формы.

Алгоритм:

1. Построение минимального оставного дерева (алгоритмы Борувки, Крускала, Прима).
2. Разделение на кластеры. Дуги с наибольшими весами разделяют кластеры.

Кластерный анализ. Классификация методов

Неиерархические (Partitional)

Вероятностные

k-средних (k-means)

k-медиан и k-метоидов (k-medians, k-medoids)

Плотностные (метод поиска сгущений)

Алгоритм K-means (k внутригрупповых средних) Мак-Куина (McQueen)

Простой и широко используемый *неиерархический* алгоритм кластеризации.

Основная идея алгоритма: разбиение всего множества объектов на k кластеров с последующим пересчетом их центров и перераспределением объектов по кластерам.

Значение k - количество кластеров задается заранее и является основным из входных данных алгоритма.

Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма.

Недостатки: качество результата сильно зависит от выбора начального разбиения; медленная работа на больших объемах исходных данных; необходимо задавать количество кластеров; алгоритм чувствителен к выбросам.

Алгоритм k -средних

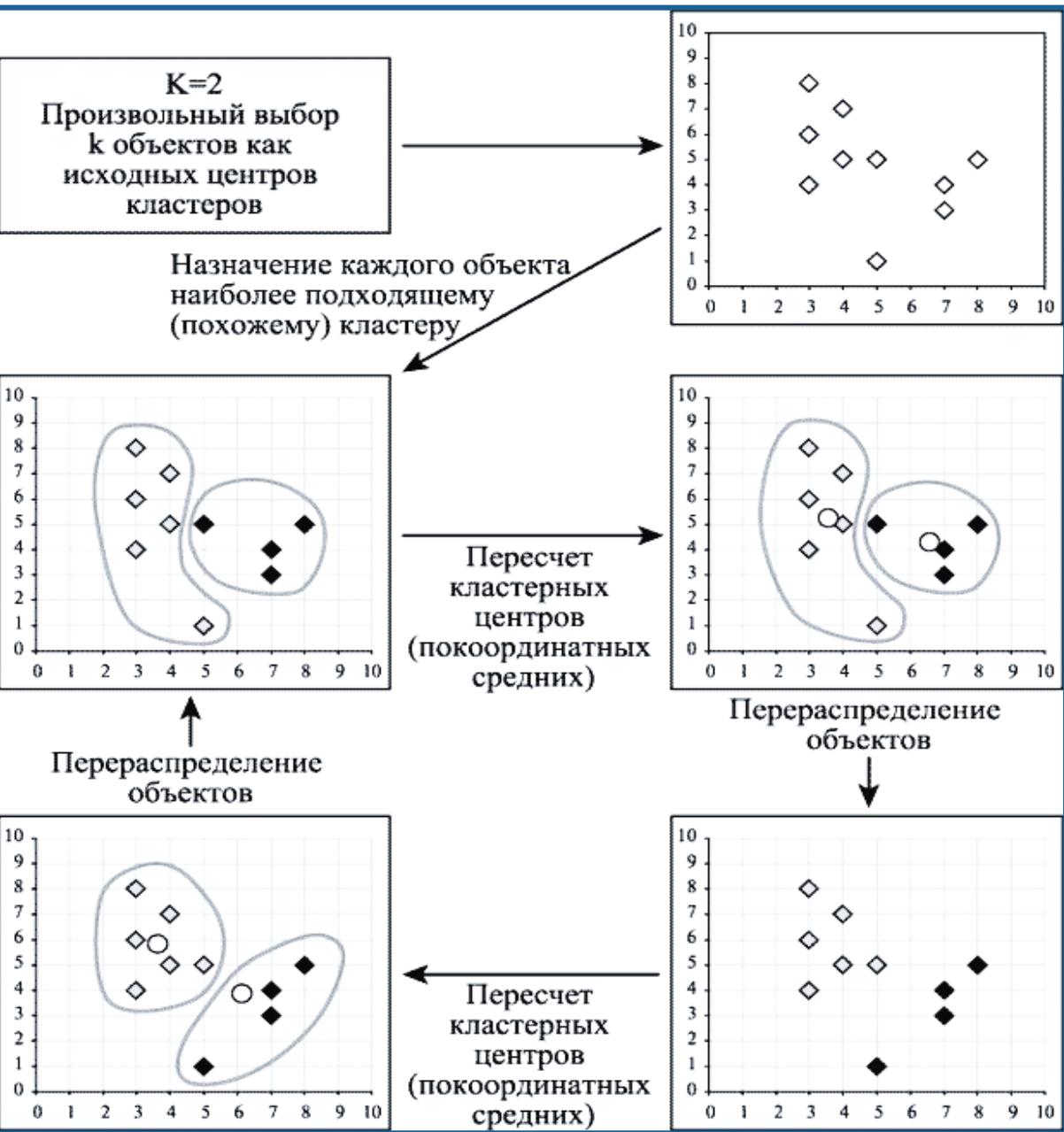
1. Выбирается k объектов (по числу кластеров) и на первом шаге они считаются в качестве центров.
2. В соответствии с выбранной метрикой каждый объект исходного множества присваивается определенному кластеру, исходя из близости его к центру кластера.
3. Пересчитываются центры кластеров в соответствии с влиянием новых объектов, попавших в кластер.
4. Далее Пункт 2.
5. Алгоритм заканчивается когда кластерные центры стабилизировались или число итераций стало равно максимальному числу итераций.

ИЛИ:

- Каждый объект x из $X = \mathbb{R}^n$ описывается n числовыми признаками: $x \equiv f_1(x), \dots, f_n(x)$. Каждый кластер $y \in Y$ описывается n -мерной гауссовой плотностью $p_y(x)$ с центром $\mu_y = (\mu_{y1}, \dots, \mu_{yn})$ и диагональной матрицей ковариаций Σ_y .
- 1) сформировать начальное приближение центров всех кластеров $y \in Y$:
 μ_y – наиболее удалённые друг от друга объекты выборки;
 - 2) **повторять**
 - отнести каждый объект к ближайшему центру:
 $y_i := \arg \min_{y \in Y} \rho(x_i; \mu_y), i = 1 \dots l;$
 - вычислить новое положение центров:
$$\mu_{yj} := \frac{\sum_{i=1}^l [y_i=y] f_j(x_i)}{\sum_{i=1}^l [y_i=y]}, y \in Y, j = 1, \dots, n;$$
- пока** y_i не перестанут изменяться *.

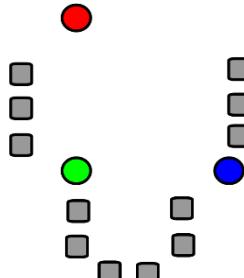
* Источник: Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: <http://www.ccas.ru/voron>

Пример работы алгоритма k-средних ($k=2$)

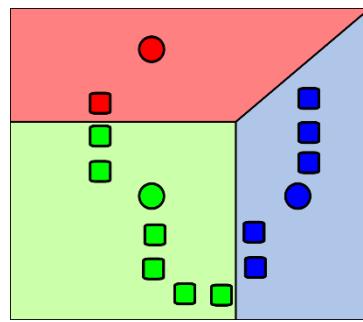


Пример работы алгоритма k-средних ($k=3$)

1. Исходные точки и случайно выбранные начальные точки.

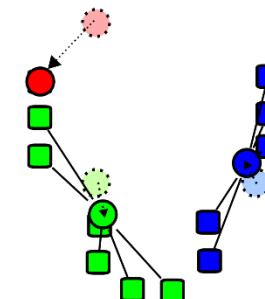


2. Точки, отнесённые к начальным центрам.

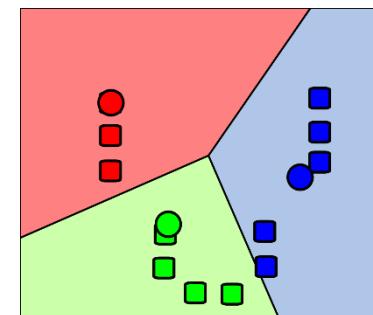


Разбиение на плоскости – диаграмма Вороного относительно начальных центров.

3. Вычисление новых центров кластеров



4. Предыдущие шаги, за исключением первого, повторяются, пока алгоритм не сойдётся.



Другие неиерархические методы

Метод k-медиан - вариация метода k-средних для задач кластеризации, где для определения центроида кластера вместо среднего вычисляется *медиана*.

- Задача определения k-медиан состоит в поиске таких k-центров, что сформированные по ним кластеры будут наиболее «компактными».
- Формально, при заданных точках данных x_i , k центров должны быть выбраны так, чтобы минимизировать сумму расстояний от каждой x_i до ближайшего j -го центроида.
- Метод иногда работает лучше, чем метод k-средних, где минимизируется сумма квадратов расстояний.

Другая альтернатива - **метод k-медоидов**, в котором ищут оптимальный *medoid*, а не медиану кластера (медоид является одной из точек данных, в то время как медианы таковыми быть не обязаны).

Другие неиерархические методы

Метод поиска сгущений. Суть итеративного алгоритма данного метода – в применении гиперсферы заданного радиуса, которая перемещается в пространстве классификационных признаков с целью поиска локальных сгущений объектов.

Метод поиска сгущений требует вычисления матрицы расстояний (или матрицы мер сходства) между объектами и выбора первоначального центра сферы.

- Обычно на первом шаге центром сферы служит объект (точка), в ближайшей окрестности которого расположено наибольшее число соседей.
- На основе заданного радиуса сферы (R) определяется совокупность точек, попавших внутрь этой сферы, и для них вычисляются координаты центра (вектор средних значений признаков).

Другие неиерархические методы

- Когда очередной пересчет координат центра сферы приводит к такому же результату, как и на предыдущем шаге, перемещение сферы прекращается, а точки, попавшие в нее, образуют кластер, и из дальнейшего процесса кластеризации исключаются.
- Перечисленные процедуры повторяются для всех оставшихся точек.
- Работа алгоритма завершается за конечное число шагов, и все точки оказываются распределенными по кластерам.
- Число образовавшихся кластеров заранее неизвестно и сильно зависит от радиуса сферы.
- Для оценки устойчивости полученного разбиения целесообразно повторить процесс кластеризации несколько раз для различных значений радиуса сферы, изменяя каждый раз радиус на небольшую величину.

Качество кластеризации

— степень приближения результата кластеризации к идеальному решению. Для большинства задач идеальное решение неизвестно.

Оценка качества кластеризации может быть произведена двумя способами:

- **Формальный способ.** Формальный способ основан на определении формальных критериев. Наилучшим считается решение, для которого значение формального критерия максимально.
- **Экспертный способ.** Решение оценивается специалистами заданной предметной области.

Качество кластеризации

Критерии качества:

- Показатели четкости: коэффициент разбиения, модифицированный коэффициент разбиения, индекс четкости.

- Энтропийные критерии: энтропия разбиения, модифицированная энтропия.

- Показатель компактности и изолированности

- Индекс эффективности



Основные этапы оценки качества кластеризации:

1. Алгоритм кластеризации, построение модели данных.
2. Вычисление критерия качества кластеризации. Критерии вычисляются на основе получившейся в ходе работы алгоритма кластеризации матрицы принадлежности и/или множества кластерных центров.
3. Определение параметров настройки алгоритма.

Кластерный анализ. Сравнение методов

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации.

- Недостаток - аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации.

*Если нет предположений относительно числа кластеров, рекомендуют использовать **иерархические алгоритмы**.*

Кластерный анализ. Сравнение методов

Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты.

За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

Иерархические методы, в отличие от неиерархических, отказываются от определения числа кластеров, а строят полное дерево вложенных кластеров.

Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций.

Кластерный анализ. Сравнение методов

- Преимущество этой группы методов в сравнении с неиерархическими методами - их наглядность и возможность получить детальное представление о структуре данных.
- При использовании *иерархических методов* существует возможность достаточно легко идентифицировать выбросы в наборе данных и, в результате, повысить качество данных.
- Эта процедура лежит в основе **двухшагового алгоритма** кластеризации. Такой набор данных в дальнейшем может быть использован для проведения *неиерархической кластеризации*.

Лекция 6

ФАКТОРНЫЙ, КОМПОНЕНТНЫЙ И
ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Факторный анализ. Основные принципы

В отличие от кластерного анализа, методы **факторного анализа** применяются, когда неизвестные факторы ищут в форме количественных переменных.

Факторный анализ - это совокупность методов, которые на основе реально существующих связей объектов (признаков) позволяют выявить латентные (неявные) обобщающие характеристики организационной структуры.

Факторный анализ. Основные принципы

При этом предполагается, что наблюдаемые переменные являются линейной комбинацией факторов.

Под *фактором* понимается гипотетическая непосредственно не измеряемая, скрытая (латентная) переменная в той или иной мере связанная с исходными наблюдаемыми переменными.

К факторному анализу в широком смысле относятся:

- метод главных компонент,
- методы многомерного шкалирования, применяемые для формирования факторного пространства по информации о близости объектов,
- методы кластерного анализа, применяемые для описания неколичественных факторов.

Основные цели факторного анализа:

- 1) сокращение числа переменных (редукция данных);
- 2) определение структуры взаимосвязей между переменными (классификация переменных);
- 3) косвенные оценки признаков, неподдающихся непосредственному измерению;
- 4) преобразование исходных переменных к более удобному для интерпретации виду.

Факторный и компонентный анализ – для решения задачи снижения размерности

- **Факторный и компонентный анализ** в большинстве случаев проводятся совместно.
- **Компонентный анализ** является методом определения структурной зависимости между случайными переменными. В результате его использования получается сжатое описание малого объема, несущее почти всю информацию, содержащуюся в исходных данных.

Факторный и компонентный анализ – для решения задачи снижения размерности

- **Факторный анализ** является более общим методом преобразования исходных переменных по сравнению с компонентным анализом.
- *Факторный анализ предназначен* для выявления действия различных факторов и их комбинаций на величину результативного признака. При этом сокращается число переменных и определяется структура взаимосвязей между переменными.

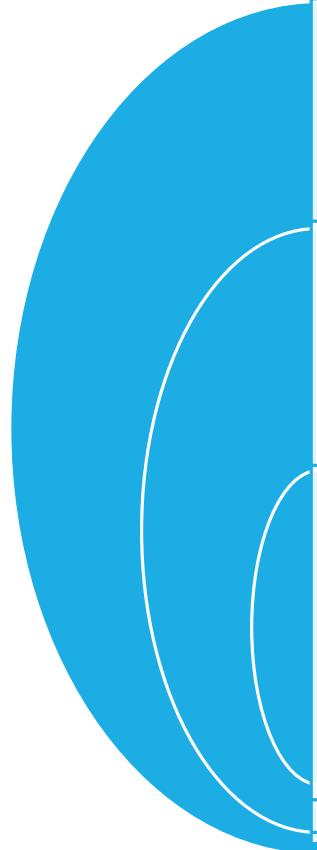
Факторный анализ. Особенности

- 1) *факторный анализ*, в противоположность контролируемому эксперименту, опирается в основном на наблюдения над естественным варьированием переменных;
- 2) при использовании *факторного анализа* совокупность переменных, изучаемых с точки зрения связей между ними, не выбирается произвольно: сам метод позволяет выявить основные факторы, оказывающие существенное влияние в данной области;
- .

Факторный анализ. Особенности

- 3) *факторный анализ* не требует предварительных гипотез, наоборот, он сам может служить методом выдвижения гипотез, а также выступать критерием гипотез, опирающихся на данные, полученные другими методами;
- 4) *факторный анализ* не требует априорных предположений относительно того, какие переменные независимы, а какие зависимы, метод не преувеличивает причинно-следственные связи и решает вопрос об их мере в процессе дальнейших исследований.

Применение методов факторного анализа включает три этапа:

- 
- 1) выделение первоначальных факторов;**
 - 2) вращение выделенных факторов с целью облегчения их интерпретации в терминах исходных переменных (в частности, для исключения отрицательных значений);**
 - 3) содержательная интерпретация новых факторов в предметных терминах, что является творческой задачей исследователя, выходящей за рамки предлагаемого формального метода.**

Наиболее часто **факторный анализ** используется для выявления в наблюдаемых признаках x_1, x_2, \dots, x_k некоторых латентных (скрытых) переменных f_m , называемых *факторами*.

Гипотеза о наличии этих факторов основана на предположении о существовании чего-то общего в наблюдаемых признаках.

Гипотетические факторы обладают следующими свойствами:

1. Они образуют линейно независимый набор переменных, т.е. ни один из факторов (компонент) не выводится как линейная комбинация остальных.
2. Переменные, являющиеся гипотетическими факторами, можно разделить на два основных вида – общие и характерные факторы. Они отличаются структурой весов в линейном уравнении, которое выводит значение наблюдаемой переменной из гипотетических факторов.

Общий фактор имеет несколько переменных с ненулевым весом или факторной нагрузкой, соответствующей этому фактору.

При этом фактор называется *общим*, если хотя бы две его нагрузки значительно отличаются от нуля.

Гипотетические факторы обладают следующими свойствами:

Характерный фактор имеет только одну переменную с ненулевым весом (т.е. только одна переменная от него зависит).

3. Всегда предполагается, что общие факторы не коррелируют с характерным фактором, также характерные факторы не коррелированы между собой.

4. Обычно предполагается, что число общих факторов меньше, чем число наблюдаемых переменных, однако число характерных факторов принимают равным числу наблюдаемых переменных.

Факторный анализ. Основные принципы

Набор «новых» признаков объясняет большую часть общей изменчивости наблюдаемых данных, а поэтому передают большую часть информации, заключенной в первоначальных наблюдениях.



Особенностью такого преобразования признаков, осуществляемого при помощи процедуры, называемой «вращением факторов» и приводящей к определению «нагрузок» признаков на агрегированные факторы, является то, что оно осуществляется без существенной потери информации.

Метод главных компонент

- Преобразование (вращение) факторов приводит к получению бесконечного множества решений, среди которых нужно выбрать те, которые облегчают интерпретацию вновь полученных факторов.
- Для того, чтобы выразить большое число откликов через малое число факторов, наиболее часто используется **метод главных компонент**.
- Это метод основан на ортогональном проектировании исходного многомерного пространства в пространство меньшей размерности, в котором точки-наблюдения имеют наибольший разброс.

Метод главных компонент

Главные компоненты получаются из исходных переменных путем целенаправленного *вращения*, т.е. как линейные комбинации исходных переменных.

Вращение производится таким образом, чтобы *главные компоненты* были ортогональны и имели максимальную дисперсию среди возможных линейных комбинаций исходных переменных X.

При этом переменные не коррелированы между собой и упорядочены по убыванию дисперсии (первая компонента имеет наибольшую дисперсию).

Общая дисперсия после преобразования остается без изменений.

Метод главных компонент

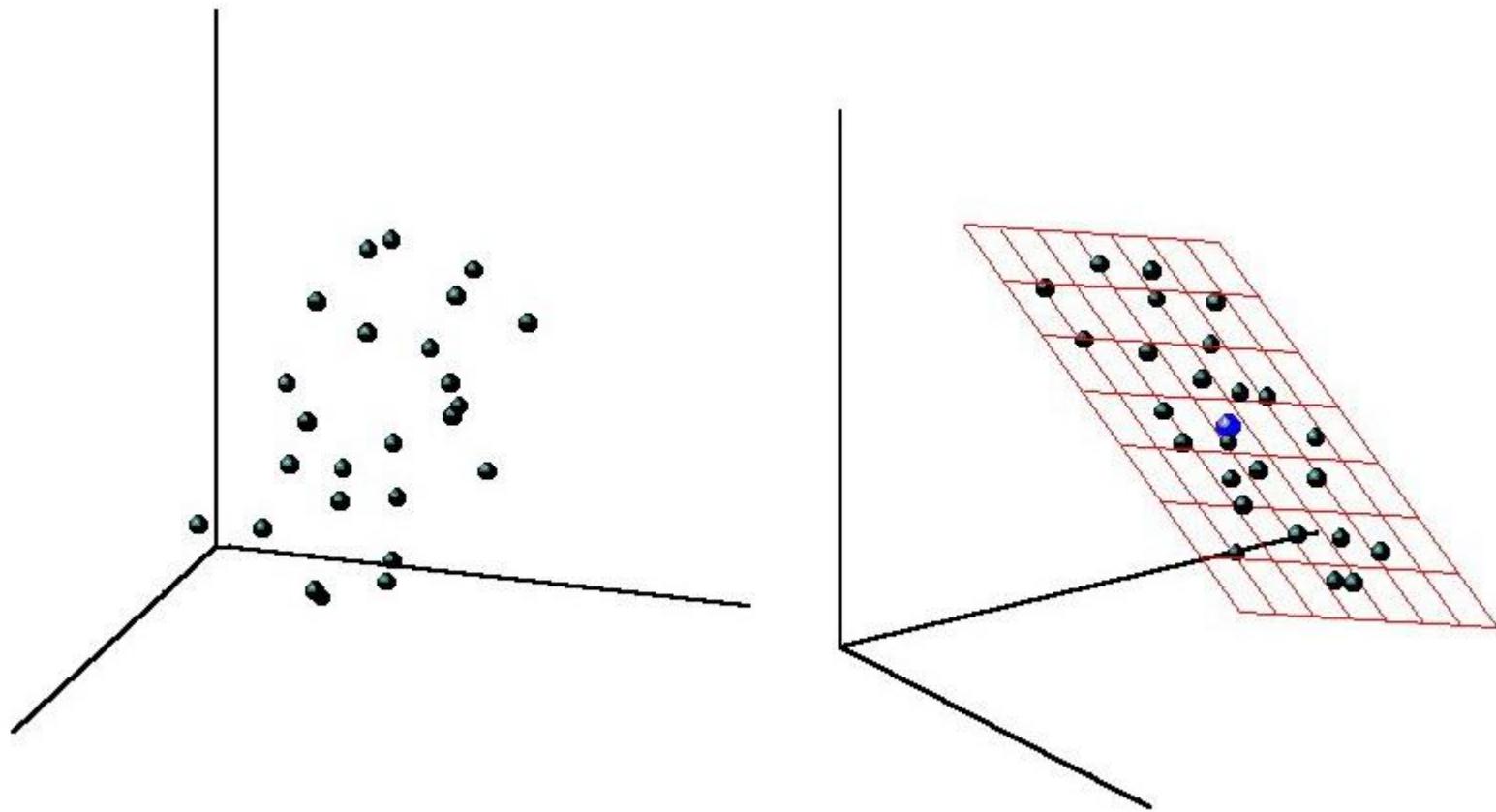


Рис. Графическое представление метода главных компонент

Метод главных компонент

Вспомогательные переменные можно спроектировать на подпространство факторов, чтобы сделать выводы об этих переменных, даже если они не участвовали непосредственно в вычислениях. То есть, вспомогательные переменные используются только для интерпретации результатов.

Метод главных компонент

- Аналогично наблюдения можно разделить на *вспомогательные* и *активные* наблюдения для анализа.
- Только основные наблюдения будут участвовать в вычислениях главных компонент.
- Вспомогательные наблюдения позже проектируются на векторное подпространство, образованное факторами, которые были вычислены на основе переменных анализа и основных наблюдений.
- Выводы на основе вычисленных факторов применимы и к вспомогательным наблюдениям.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Дискриминантный анализ

Дискриминантный анализ является разделом многомерного статистического анализа, который включает в себя методы классификации многомерных наблюдений по принципу максимального сходства при наличии обучающих признаков.

В кластерном анализе рассматриваются методы многомерной классификации без обучения.

В дискриминантном анализе новые кластеры не образуются, а формулируется правило, по которому объекты подмножества подлежащего классификации относятся к одному из уже существующих (обучающих) подмножеств (классов), на основе сравнения величины дискриминантной функции классифицируемого объекта, рассчитанной по дискриминантным переменным, с некоторой константой дискриминации.

Дискриминантный анализ

Дискриминантный анализ – это общий термин, относящийся к нескольким тесно связанным статистическим процедурам.

Эти процедуры можно разделить на методы *интерпретации межгрупповых различий – дискриминации* и методы *классификации наблюдений* по группам.

Дискриминантный анализ

Задачи дискриминантного анализа

Задачи первого типа – задачи дискrimинации (пример – в медицинской практике).

Второй тип задачи относится к ситуации, когда признаки принадлежности объекта к той или иной группе потеряны, и их нужно восстановить.

Задачи третьего типа связаны с предсказанием будущих событий на основании имеющихся данных.

Дискриминация

Основной целью дискриминации является нахождение такой линейной комбинации переменных (в дальнейшем эти переменные будем называть ***дискриминантными переменными***), которая бы оптимально разделила рассматриваемые группы.

Дискриминация

Линейная функция

$$d_{km} = \beta_0 + \beta_1 x_{1km} + \dots + \beta_p x_{pkm}, \quad m=1, \dots, n, \quad k=1, \dots, g$$

называется **канонической дискриминантной функцией** с неизвестными коэффициентами β_i

d_{km} — значение дискриминантной функции для m -го объекта в группе k

x_{ikm} — значение дискриминантной переменной X_i для m -го объекта в группе k .

С геометрической точки зрения дискриминантные функции определяют гиперповерхности в p -мерном пространстве.

Дискриминация

Коэффициенты β_i первой канонической дискриминантной функции d выбираются таким образом, чтобы центроиды различных групп как можно больше отличались друг от друга.

Коэффициенты второй группы выбираются также, но при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой.

Аналогично определяются и другие функции.

Отсюда следует, что *любая каноническая дискриминантная функция имеет нулевую внутригрупповую корреляцию с d_1, d_2, \dots, d_{g-1}*

Дискриминация

- Если число групп равно g , то число канонических дискриминантных функций будет на единицу меньше числа групп.
- Однако по многим причинам практического характера полезно иметь одну, две или же три дискриминантных функций.
- Тогда графическое изображение объектов будет представлено в одно-, двух- и трехмерных пространствах.

Коэффициенты канонической дискриминантной функции

Для получения коэффициентов β_i канонической дискриминантной функции нужен статистический критерий различия групп.

Классификация переменных будет осуществляться тем лучше, чем меньше рассеяние точек относительно центроида внутри группы и чем больше расстояние между центроидами групп.

Большая внутригрупповая вариация нежелательна, так как в этом случае любое заданное расстояние между двумя средними тем менее значимо в статистическом смысле, чем больше вариация распределений, соответствующих этим средним.

Коэффициенты канонической дискриминантной функции

Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d , которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$\lambda = \mathbf{B}(d)/\mathbf{W}(d) \quad (2)$$

где \mathbf{B} - межгрупповая и \mathbf{W} внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

В некоторых работах вместо \mathbf{W} используют матрицу рассеяния \mathbf{T} объединенных данных.

Коэффициенты канонической дискриминантной функции

Рассмотрим *максимизацию отношения* (2) *для произвольного числа классов.*

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в k -й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина переменной i для m -го наблюдения в k -й группе;

\bar{x}_{ik} – средняя величина переменной i в k -й группе;

\bar{x}_i – среднее значение переменной i по всем группам;

$T(u, v)$ – общая сумма перекрестных произведений для переменных u и v ;

$W(u, v)$ – внутригрупповая сумма перекрестных произведений для переменных u и v .

$t_{ij} = T(x_i, x_j); w_{ij} = W(x_i, x_j).$

В модели дискриминации должны соблюдаться следующие условия:

- 1) число групп: $g \geq 2$;
- 2) число объектов в каждой группе: $n_i \geq 2$;
- 3) число дискриминантных переменных: $0 < p < (n - 2)$;
- 4) дискриминантные переменные измеряются в интервальной
шкале;
- 5) дискриминантные переменные линейно независимы;
- 6) ковариационные матрицы групп примерно равны;
- 7) дискриминантные переменные в каждой группе подчиняются
многомерному нормальному закону распределения.

Коэффициенты канонической дискриминантной функции

Рассмотрим задачу максимизации отношения (2) когда имеются g групп.

Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп.

Для этого вычислим матрицу рассеяния \mathbf{T} , которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних \bar{x}_i , $i = 1, \dots, p$ по каждой переменной.

Элементы матрицы \mathbf{T} определяются выражением

$$t_{ij} = \sum_{k=1}^g \sum_{m=1}^n (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j)$$

Лекция 7

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Коэффициенты канонической дискриминантной функции

Один из методов поиска наилучшей дискриминации данных заключается в нахождении такой канонической дискриминантной функции d , которая бы максимизировала отношение межгрупповой вариации к внутригрупповой

$$\lambda = \mathbf{B}(d)/\mathbf{W}(d) \tag{2}$$

где \mathbf{B} - межгрупповая и \mathbf{W} внутригрупповая матрицы рассеяния наблюдаемых переменных от средних.

В некоторых работах вместо \mathbf{W} используют матрицу рассеяния \mathbf{T} объединенных данных.

Коэффициенты канонической дискриминантной функции

Рассмотрим *максимизацию отношения* (2) *для произвольного числа классов.*

Введем следующие обозначения:

g – число классов;

p – число дискриминантных переменных;

n_k – число наблюдений в k -й группе;

n – общее число наблюдений по всем группам;

x_{ikm} – величина переменной i для m -го наблюдения в k -й группе;

\bar{x}_{ik} – средняя величина переменной i в k -й группе;

\bar{x}_i – среднее значение переменной i по всем группам;

$T(u, v)$ – общая сумма перекрестных произведений для переменных u и v ;

$W(u, v)$ – внутригрупповая сумма перекрестных произведений для переменных u и v .

$t_{ij} = T(x_i, x_j); w_{ij} = W(x_i, x_j).$

В модели дискриминации должны соблюдаться следующие условия:

- 1) число групп: $g \geq 2$;**
- 2) число объектов в каждой группе: $n_i \geq 2$;**
- 3) число дискриминантных переменных: $0 < p < (n - 2)$;**
- 4) дискриминантные переменные измеряются в интервальной
шкале;**
- 5) дискриминантные переменные линейно независимы;**
- 6) ковариационные матрицы групп примерно равны;**
- 7) дискриминантные переменные в каждой группе подчиняются
многомерному нормальному закону распределения.**

Коэффициенты канонической дискриминантной функции

Рассмотрим задачу максимизации отношения (2) когда имеются g групп.

Оценим сначала информацию, характеризующую степень различия между объектами по всему пространству точек, определяемому переменными групп.

Для этого вычислим матрицу рассеяния \mathbf{T} , которая равна сумме квадратов отклонений и попарных произведений наблюдений от общих средних \bar{x}_i , $i = 1, \dots, p$ по каждой переменной.

Элементы матрицы \mathbf{T} определяются выражением

$$t_{ij} = \sum_{k=1}^g \sum_{m=1}^n (x_{ikm} - \bar{x}_i)(x_{jkm} - \bar{x}_j) \quad (3)$$

07.05.2020

Продолжение темы «Дискриминантный анализ»

Коэффициенты канонической дискриминантной функции

$$\text{В (3): } \bar{x}_i = (1/n) \sum_{k=1}^g n_i \bar{x}_{ik}, i = 1, \dots, p$$

$$\bar{x}_{ik} = (1/n_i) \sum_{m=1}^{n_k} x_{ikm}, i = 1, \dots, p; k = 1, \dots, g$$

Запишем это выражение в матричной форме. Обозначим p -мерную случайную векторную переменную k -ой группы следующим образом

$$X_k = \{x_{ikm}\} \quad i = 1, \dots, p, \quad k = 1, \dots, g, \quad m = 1, \dots, n_k$$

Тогда объединенная p -мерная случайная векторная переменная всех групп будет иметь вид

$$\mathbf{X} = [X_1 \ X_2 \ \dots \ X_g]$$

Коэффициенты канонической дискриминантной функции

Общее среднее этой p -мерной случайной векторной переменной будет равен вектору средних отдельных признаков

$$\bar{\mathbf{x}} = [\bar{x}_1 \bar{x}_2 \dots \bar{x}_p]$$

Матрица рассеяния от среднего при этом запишется в виде

$$\mathbf{T} = \sum_{k=1}^g (\mathbf{X}_k - \bar{\mathbf{x}})(\mathbf{X}_k - \bar{\mathbf{x}})'$$

Если использовать векторную переменную объединенных переменных \mathbf{X} , то матрица \mathbf{T} определится по формуле

$$\mathbf{T} = (\mathbf{X} - \bar{\mathbf{x}})(\mathbf{X} - \bar{\mathbf{x}})'$$

Коэффициенты канонической дискриминантной функции

- Матрица T содержит полную информацию о распределении точек по пространству переменных.
- Диагональные элементы представляют собой сумму квадратов отклонений от общего среднего и показывают как ведут себя наблюдения по отдельно взятой переменной.
- Внедиагональные элементы равны сумме произведений отклонений по одной переменной на отклонения по другой.

- Если разделить матрицу \mathbf{T} на $(n - 1)$, то получим ковариационную матрицу.
- Для проверки условия линейной независимости переменных полезно рассмотреть вместо \mathbf{T} нормированную корреляционную матрицу.
- Для измерения степени разброса объектов внутри групп рассмотрим матрицу \mathbf{W} , которая отличается от \mathbf{T} только тем, что ее элементы определяются векторами средних для отдельных групп, а не вектором средних для общих данных.
- Элементы внутригруппового рассеяния определяются выражением

$$w_{ij} = \sum_{k=1}^g \sum_{m=1}^{n_k} (x_{ikm} - \bar{x}_{ik})(x_{jkm} - \bar{x}_{jk})$$

Запишем это выражение в матричной форме.

Данным g групп будут соответствовать векторы средних

$$\bar{\mathbf{x}}_1 = [\bar{x}_{11} \bar{x}_{21} \dots \bar{x}_{p1}],$$

...

$$\bar{\mathbf{x}}_g = [\bar{x}_{1g} \bar{x}_{2g} \dots \bar{x}_{pg}].$$

Тогда матрица внутригрупповых вариаций запишется в виде

$$\mathbf{W} = \sum_{k=1}^g (\mathbf{X}_k - \bar{\mathbf{x}}_k)(\mathbf{X}_k - \bar{\mathbf{x}}_k)'$$

Если разделить каждый элемент матрицы \mathbf{W} на $(n-g)$, то получим оценку ковариационной матрицы внутригрупповых данных.

- Когда центроиды различных групп совпадают, то элементы матриц \mathbf{T} и \mathbf{W} будут равны.
- Если же центроиды групп различные, то разница

$$\mathbf{B} = \mathbf{T} - \mathbf{W} \quad (8)$$

будет определять межгрупповую сумму квадратов отклонений и попарных произведений.

- Если расположение групп в пространстве различается (т.е. их центроиды не совпадают), то степень разброса наблюдений внутри групп будет меньше межгруппового разброса.
- Элементы матрицы \mathbf{B} можно вычислить и по данным средних

$$b_{ij} = \sum_{k=1}^g n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j), \quad i, j = 1, \dots, p \quad (9)$$

- Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами.
- *Для лучшего разделения наблюдений на группы* нужно подобрать коэффициенты дискриминантной функции из условия максимизации отношения межгрупповой матрицы рассеяния к внутригрупповой матрице рассеяния при условии ортогональности дискриминантных плоскостей.
- Тогда нахождение коэффициентов дискриминантных функций сводится к решению задачи о собственных значениях и векторах.

- Это утверждение можно сформулировать так: если спроектировать g групп p -мерных выборок на $(g - 1)$ пространство, порожденное собственными векторами

$$(v_{1k}, \dots, v_{pk}), k = 1, \dots, g - 1,$$

то отношение (2) будет максимальным, т.е. *рассеивание между группами будет максимальным при заданном внутригрупповом рассеивании.*

- Если бы мы захотели спроектировать g выборок на прямую при условии максимизации наибольшего рассеивания между группами, то следовало бы использовать собственный вектор (v_{11}, \dots, v_{1k}) соответствующий максимальному собственному числу λ_1 .

При этом дискриминантные функции можно получать:
по **нестандартизованным и стандартизованным коэффициентам.**

Нестандартизованные коэффициенты

Пусть $\lambda_1 \geq \dots \geq \lambda_p$ и $\mathbf{v}_1, \dots, \mathbf{v}_p$ соответственно собственные значения и векторы. Тогда условие (2) в терминах собственных чисел и векторов запишется в виде

$$\lambda = \frac{\sum_k b_{jk} v_j v_k}{\sum_k w_{jk} v_j v_k}$$

что влечет равенство $\sum_k (b_{jk} - \lambda w_{jk}) v_k = 0$

или в матричной записи $(\mathbf{B} - \lambda_i \mathbf{W}) \mathbf{v}_i = 0, \quad \mathbf{v}_i' \mathbf{W} \mathbf{v}_j = \delta_{ij}$ (10)

где δ_{ij} символ Кронекера.

Таким образом, решение уравнения $|\mathbf{B} - \lambda \mathbf{W}| = 0$ позволяет нам определить компоненты собственных векторов, соответствующих дискриминантным функциям.

Нестандартизованные коэффициенты

Если \mathbf{B} и \mathbf{W} невырожденные матрицы, то собственные корни уравнения

$$|\mathbf{B} - \lambda \mathbf{W}| = 0 \quad \text{такие же, как и у} \quad |\mathbf{W}^{-1}\mathbf{B} - \lambda \mathbf{I}| = 0$$

Решение системы уравнений (10) можно получить путем использования *разложения Холецкого* $\mathbf{L}\mathbf{L}'$ матрицы \mathbf{W}^{-1} и решения задачи о собственных значениях

$$(\mathbf{L}'\mathbf{B}\mathbf{L} - \lambda_i \mathbf{I})\mathbf{v}_i = 0, \quad \mathbf{v}_i' \mathbf{v}_j = \delta_{ij}$$

- Каждое решение, которое имеет свое собственное значение λ_i и собственный вектор \mathbf{v}_i , соответствует одной дискриминантной функции.
- Компоненты собственного вектора \mathbf{v}_i можно использовать в качестве коэффициентов дискриминантной функции.
- Однако при таком подходе начало координат не будет совпадать с главным центроидом.

Нестандартизованные коэффициенты

Для того, чтобы начало координат совпало с главным центроидом нужно нормировать компоненты собственного вектора

$$\beta_i = v_i \sqrt{n - g}, \quad \beta_0 = - \sum_{i=1}^p \beta_i \bar{x}_i \quad (11)$$

- Нормированные коэффициенты (11) получены по нестандартизованным исходным данным, поэтому они называются *нестандартизованными*.
- Нормированные коэффициенты приводят к таким дискриминантным значениям, единицей измерения которых является стандартное квадратичное отклонение.
- При таком подходе каждая ось в преобразованном пространстве сжимается или растягивается таким образом, что соответствующее дискриминантное значение для данного объекта представляет собой число стандартных отклонений точки от главного центроида.

Стандартизованные коэффициенты

можно получить двумя способами:

- 1) по формуле (11), если исходные данные были приведены к стандартной форме;
- 2) преобразованием нестандартизованных коэффициентов к стандартизованной форме:

$$\mathbf{c}_i = \beta_i \sqrt{\frac{w_{ii}}{n - g}} \quad (12)$$

где w_{ii} – сумма внутригрупповых квадратов i -й переменной, определяемой по формуле (5).

- Стандартизованные коэффициенты полезно применять для уменьшения размерности исходного признакового пространства переменных.
- Если абсолютная величина коэффициента для данной переменной для всех дискриминантных функций мала, то эту переменную можно исключить, тем самым сократив число переменных.

Структурные коэффициенты

определяются коэффициентами взаимной корреляции между отдельными переменными и дискриминантной функцией. Если относительно некоторой переменной абсолютная величина коэффициента велика, то вся информация о дискриминантной функции заключена в этой переменной.

Структурные коэффициенты полезны при классификации групп.

Структурный коэффициент можно вычислить и для переменной в пределах отдельно взятой группы.

Тогда получаем *внутригрупповой структурный коэффициент*, который вычисляется по формуле:

$$s_{ij} = \sum_{k=1}^p r_{ik} c_{kj} = \sum_{k=1}^p \frac{w_{ik} c_{kj}}{\sqrt{w_{ii} w_{jj}}}$$

где s_{ij} – внутригрупповой структурный коэффициент для i -ой переменной и j -ой функции; r_{ik} – внутригрупповые структурные коэффициенты корреляции между переменными i и k ; c_{kj} – стандартизованные коэффициенты канонической функции для переменной k и функции j .

Структурные и стандартизованные коэффициенты

- *Структурные коэффициенты* по своей информативности несколько отличаются от стандартизованных коэффициентов.
- *Стандартизованные коэффициенты* показывают вклад переменных в значение дискриминантной функции. Если две переменные сильно коррелированы, то их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется только одна из этих переменных.
- Такое распределение величины стандартизованного коэффициента объясняется тем, что при их вычислении учитывается влияние всех переменных.
- Структурные же коэффициенты являются парными корреляциями и на них не влияют взаимные зависимости прочих переменных.

Коэффициент канонической корреляции

Другой характеристикой, позволяющей оценить полезность дискриминантной функции является *коэффициент канонической корреляции* r_i .

Каноническая корреляция является мерой связи между двумя множествами переменных. Максимальная величина этого коэффициента равна 1.

Будем считать, что группы составляют одно множество, а другое множество образуют дискриминантные переменные.

Коэффициент канонической корреляции для i -ой дискриминантной функции определяется формулой:

$$r_i = \sqrt{\frac{\lambda_i}{1 + \lambda_i}} \quad (14)$$

Остаточная дискриминация

- Так как дискриминантные функции находятся по выборочным данным, они нуждаются в *проверке статистической значимости*.
- Дискриминантные функции представляются аналогично главным компонентам. Поэтому для проверки этой значимости можно воспользоваться *критерием*, аналогичным дисперсионному критерию в методе главных компонент.
- Этот критерий оценивает *остаточную дискриминантную способность*, под которой понимается способность различать группы, если при этом исключить информацию, полученную с помощью ранее вычисленных функций.
- Если остаточная дискриминация мала, то не имеет смысла дальнейшее вычисление очередной дискриминантной функции.

Остаточная дискриминация

Полученная статистика носит название «*Λ-статистики Уилкса*» и вычисляется по формуле:

$$\Lambda = \prod_{i=k+1}^g (1/(1 + \lambda_i)) \quad (15)$$

где k – число вычисленных функций.

Чем меньше эта статистика, тем значимее соответствующая дискриминантная функция.

Остаточная дискриминация

Величина

$$\chi^2 = -[n - ((p + g)/2) - 1] \ln \Lambda_k, \quad k = 0, 1, \dots, g - 1$$

имеет хи–квадрат распределение с $(p - k)(g - k - 1)$ степенями свободы.

Вычисления проводятся в следующем порядке:

1. Находим значение критерия χ^2 при $k=0$. Значимость критерия подтверждает существование различий между группами. Кроме того, это доказывает, что первая дискриминантная функция значима и имеет смысл ее вычислять.
2. Определяем первую дискриминантную функцию и проверяем значимость критерия при $k=1$. Если критерий значим, то вычисляем вторую дискриминантную функцию и продолжаем процесс до тех пор, пока не будет исчерпана вся значимая информация.

Классифицирующие функции

- Ранее было рассмотрено получение канонических дискриминантных функций при известной принадлежности объектов к тому или иному классу.
 - Основное внимание уделялось *определению числа и значимости этих функций, и использованию их для объяснения различий между классами*. Все сказанное относилось к интерпретации результатов ДА.
-
- Однако наибольший интерес представляет *задача предсказания класса, которому принадлежит некоторый случайно выбранный объект*.
 - Эту задачу можно решить, используя информацию, содержащуюся в дискриминантных переменных. Существуют различные способы классификации.

Классифицирующие функции

- В процедурах классификации могут использоваться как сами дискриминантные переменные, так и канонические дискриминантные функции.
- В первом случае применяется *метод максимизации различий между классами* для получения функции классификации, различие же классов на значимость не проверяется и, следовательно, дискриминантный анализ не проводится.
- Во втором случае для классификации используются непосредственно дискриминантные функции и проводится более глубокий анализ.

Контрольные вопросы

Вариант 1

1. Разведочный анализ данных: предпосылки, суть и цели.
2. Количественный анализ – суть и предназначение. Фазы количественной обработки данных.

Вариант 2

1. Принцип «мягких вычислений» - суть и значение для ИАД.
2. Качественный анализ - суть и предназначение

Вариант 3

1. Data Mining: определение понятия и требования к знаниям.
2. Понятие корреляционно-регрессионного анализа данных. Этапы КРА.

Вариант 4

1. Стадии ИАД.
2. Дисперсионный анализ. Постановка задачи.

Вариант 5

1. Основные задачи Data Minig.
2. Однофакторный дисперсионный анализ. Основные этапы и соотношения.

Лекция 7-8

ЗАДАЧА КЛАССИФИКАЦИИ

Задача классификации

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 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.

Задача классификации

- *Классификация* - это закономерность, позволяющая делать вывод относительно определения характеристик конкретной группы.
- Для проведения *классификации* должны присутствовать признаки, характеризующие группу, к которой принадлежит 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*), которое также именуют контролируемым или управляемым обучением.
- Задачей *классификации* часто называют предсказание категориальной зависимой переменной (т.е. зависимой переменной, являющейся категорией) на основе выборки непрерывных и/или категориальных переменных.

Постановка задачи

Предполагается, что уже имеется какое-то количество n объектов, для каждого из которых известен некоторый набор из m признаков (факторов) и номер класса, к которому этот объект принадлежит, т.е. сырье данные, используемые для решения задачи классификации, имеют вид:

Номер наблюдения, i	Значения факторов			Значения переменной отклика (номер класса)
1	$x_{1,1}$...	$x_{1,m}$	y_1
...
i	$x_{i,1}$...	$x_{i,m}$	y_i
...
n	$x_{n,1}$...	$x_{n,m}$	y_n

Здесь значения переменной отклика – номер класса, которому принадлежит объект, т.е. $y_i \in \{1, \dots, K\}$, для всех $i = 1, \dots, n$, K – (известное) количество классов

Постановка задачи

Как и в задаче регрессионного анализа, предположим, что имеется n объектов, каждый из которых описывается m признаками.

Будем нумеровать объекты индексом i ($i = 1, \dots, n$), а признаки (значения которых могут быть получены непосредственным измерением) – индексом j ($j = 1, \dots, m$).

Для объекта с номером i обозначим через $x_{i,j}$ – значения признака j ;

y_i – значение зависимого признака объекта i .

Постановка задачи

Пример. Пусть объекты – это клиенты банка, наблюдаемый признак x – уровень их заработной платы, прогнозируемый признак y – состояние кредитной карты. Цель исследования – спрогнозировать, «уйдёт ли в минус» тот или иной клиент банка (владелец банковской карты).

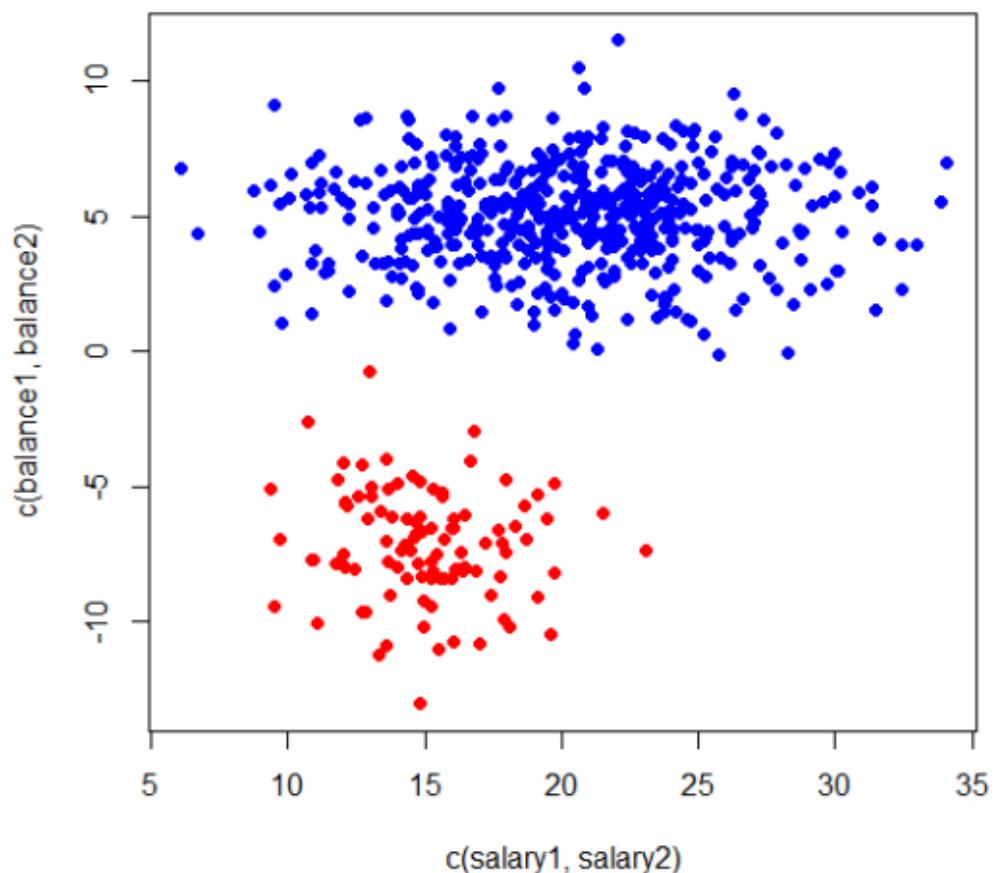
В этом примере $m = 1$. Данные n обследованных клиентов запишем как пары (x_i, y_i) , $i = 1, \dots, n$. Каждой такой паре можем поставить в соответствие точку на координатной плоскости.

«Разобьём» всех клиентов на два класса:

- «благонадёжные» (т.е. имеющие неотрицательный баланс на карте)
- «неблагонадёжные» (т.е. имеющие отрицательный баланс на карте).

Постановка задачи. Пример

Зависимость между зарплатой и балансом карты



Синие точки соответствуют клиентам банка, имеющим неотрицательный баланс кредитной карты, красным – отрицательный.

Одни и те же значения по оси абсцисс могут соответствовать как синим, так и красным точкам.

Однако можно заметить, что большим значениям признака x (заработной плате) соответствует большее число синих точек, чем красных.

Виды классификации

Вспомогательная

(искусственная) классификация:
производится по внешнему
признаку и служит для придания
множеству предметов
(процессов, явлений) нужного
порядка;

Естественная классификация:

производится по существенным
признакам, характеризующим
внутреннюю общность предметов и
явлений, предполагает и закрепляет
результаты изучения закономерностей
классифицируемых объектов.

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

Простая - деление родового
понятия только по признаку и
только один раз до раскрытия
всех видов.

Сложная - применяется для
деления одного понятия по
разным основаниям и синтеза
таких простых делений в единое
целое.

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

Процесс классификации

1 этап. Конструирование модели: описание множества предопределенных классов.

- Каждый пример набора данных относится к одному предопределенному классу.
- На этом этапе используется обучающее множество, на нем происходит конструирование модели.
- Полученная модель представлена классификационными правилами, деревом решений или математической формулой.

2 этап. Использование модели: классификация новых или неизвестных значений.

- Оценка правильности (точности) модели.
 1. Известные значения из тестового примера сравниваются с результатами использования полученной модели.
 2. Уровень точности - процент правильно классифицированных примеров в тестовом 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 для классификации новых примеров, класс которых неизвестен.

Основные методы классификации:

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.

21.05.2020

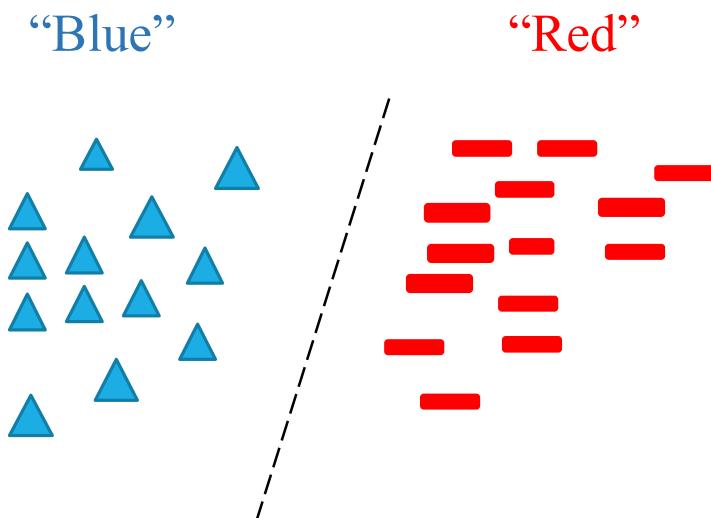
Лекция 8

ЗАДАЧА КЛАССИФИКАЦИИ. ПРОДОЛЖЕНИЕ

Метод опорных векторов

Support Vector Machine - SVM

- относится к группе **границых методов** (классы определяются при помощи границ областей);
- Назначение** - с помощью SVM решают задачи **бинарной классификации**;
- в основе метода** - понятие **плоскостей решений** (плоскость решения разделяет объекты с разной классовой принадлежностью).



Разделение классов прямой линией

Разделяющая линия задает границу, справа от которой - все объекты типа “blue”, слева - типа “red”.

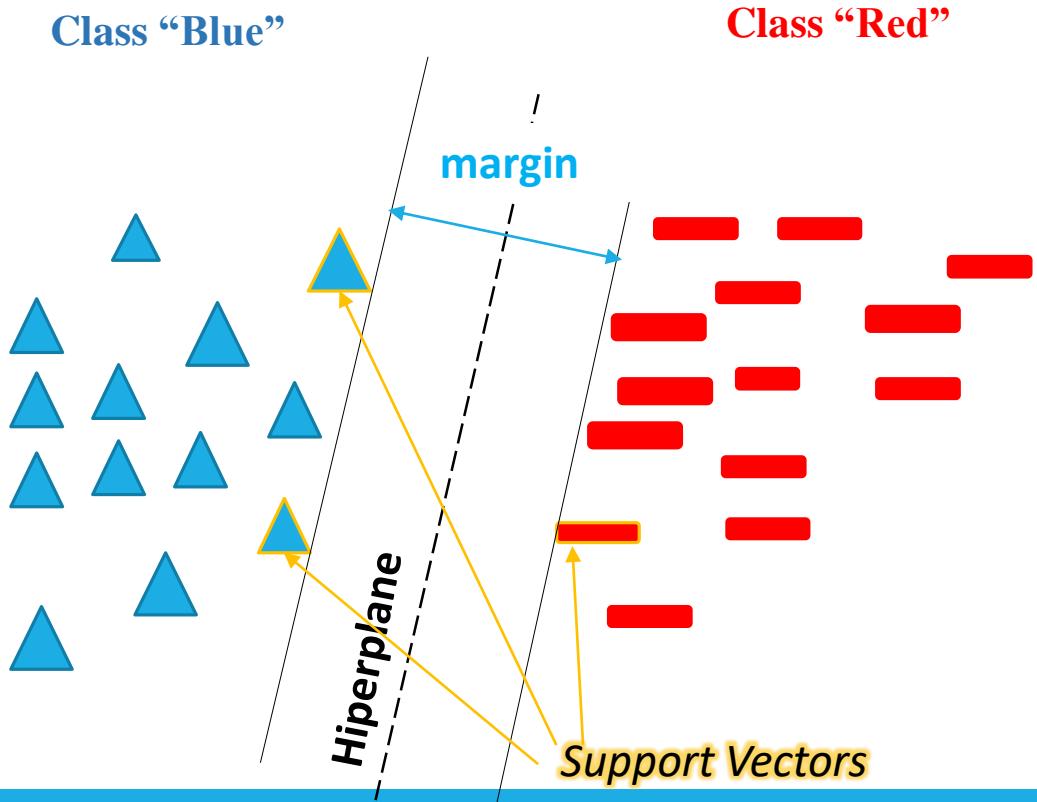
Новый объект, попадающий направо, классифицируется как объект класса **blue** или - как объект класса **red**, если он расположился по левую сторону от разделяющей прямой.

В этом случае
каждый объект характеризуется двумя
измерениями.

Метод опорных векторов

Цель метода опорных векторов - найти плоскость (*гиперплоскость*), разделяющую два множества объектов.

Метод отыскивает образцы, находящиеся на границах между двумя классами, т.е. **опорные векторы**.



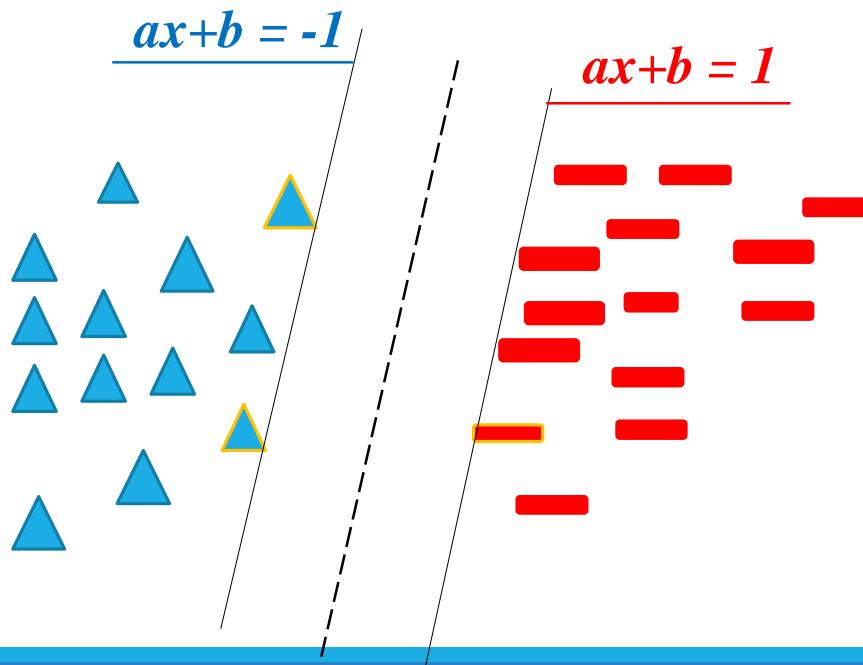
Опорными векторами называются объекты множества, лежащие на границах областей.

Классификация считается хорошей, если область между границами пуста.

Линейный метод опорных векторов (SVM)

Решение задачи **бинарной классификации** при помощи **метода опорных векторов** заключается в поиске некоторой линейной функции, которая правильно разделяет набор данных на два класса.

Рассмотрим задачу классификации, где число классов равно двум: поиск функции $f(x)$, принимающей значения **меньше нуля** для векторов **одного класса** и **больше нуля** - для векторов **другого класса**.



Дан тренировочный набор векторов пространства, для которых известна их принадлежность к одному из классов.

Семейство классифицирующих функций можно описать через функцию $f(x)$.

Гиперплоскость определена вектором a и значением b , т.е. $f(x)=ax+b$.

Линейный метод опорных векторов (SVM)

- В результате решения задачи (построения SVM-модели) будет найдена функция, принимающая значения меньше нуля для векторов одного класса и больше нуля - для векторов другого класса.
- Для каждого нового объекта отрицательное или положительное значение определяет принадлежность объекта к одному из классов.
- Наилучшей функцией классификации является функция, для которой ожидаемый риск минимален. Понятие **ожидаемого риска** в данном случае означает ожидаемый уровень ошибки классификации.
- Напрямую оценить ожидаемый уровень ошибки построенной модели невозможно, это можно сделать при помощи понятия эмпирического риска.
- Но следует учитывать, что минимизация эмпирического риска не всегда приводит к минимизации ожидаемого риска (особенно для небольших наборов тренировочных данных).

Эмпирический риск - уровень ошибки классификации на тренировочном наборе.

Математическая постановка задачи

Пусть имеется обучающая выборка:

$$(x_1, y_1), \dots, (x_n, y_n), \quad x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$$

Метод опорных векторов строит классифицирующую функцию F в виде

$$F(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b),$$

где $\langle \cdot, \cdot \rangle$ — скалярное произведение, \mathbf{w} — нормальный вектор к разделяющей гиперплоскости, b — вспомогательный параметр.

Те объекты, для которых $F(\mathbf{x}) = 1$ попадают в один класс, а объекты с $F(\mathbf{x}) = -1$ — в другой.

Выбор именно такой функции неслучаен: любая гиперплоскость может быть задана в виде $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ для некоторых \mathbf{w} и b .

Математическая постановка задачи

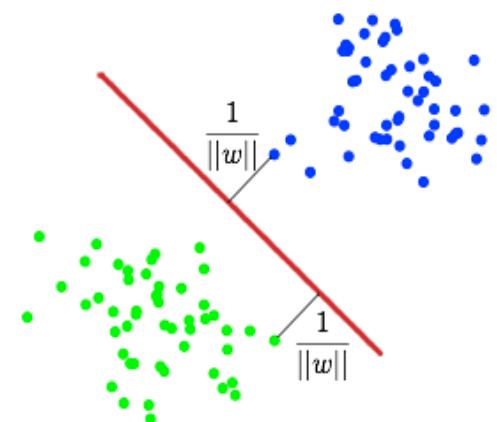
Далее, мы хотим выбрать такие w и b которые максимизируют расстояние до каждого класса.

Можно подсчитать, что данное расстояние равно $\frac{1}{\|w\|}$.

Проблема нахождения максимума эквивалентна проблеме нахождения минимума $\frac{1}{\|w\|}$.

Запишем все это в виде задачи оптимизации:

$$\begin{cases} \arg \min_{w,b} \|w\|^2, \\ y_i(\langle w, x \rangle + b) \geq 1, \quad i = 1, \dots, m. \end{cases}$$



Это является стандартной задачей квадратичного программирования и решается с помощью множителей Лагранжа.

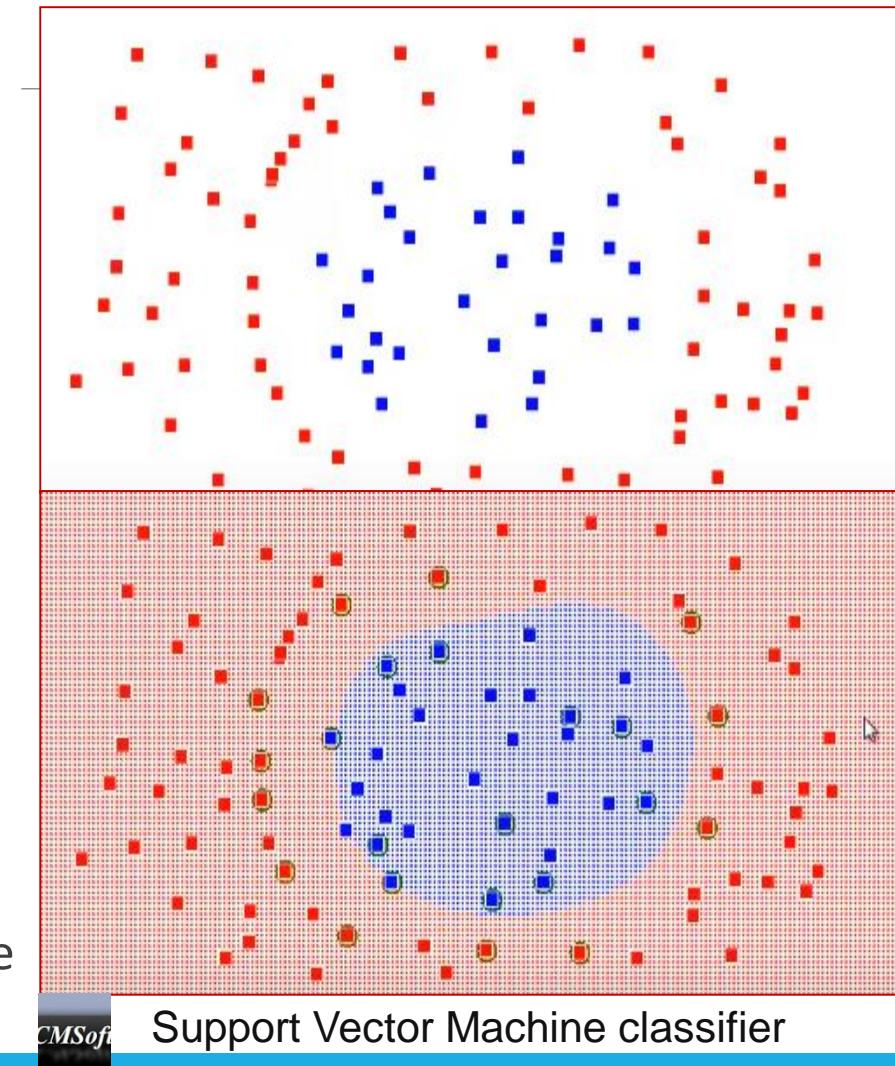
Проблемы классификации в SVM

Одна из проблем – не всегда можно легко найти линейную границу между двумя классами.

В таких случаях один из вариантов - увеличение размерности, т.е. перенос данных из плоскости в трехмерное пространство, где возможно построить такую плоскость, которая идеально разделит множество образцов на два класса.

Опорными векторами в этом случае будут служить объекты из обоих классов, являющиеся экстремальными.

Таким образом, при помощи добавления так называемого **оператора ядра** и дополнительных размерностей, находятся границы между классами в виде гиперплоскостей.



Линейная неразделимость

На практике случаи, когда данные можно разделить гиперплоскостью, или, как еще говорят, *линейно*, довольно редки.

В этом случае поступают так: все элементы обучающей выборки вкладываются в пространство \mathbf{X} более высокой размерности с помощью специального отображения $\varphi; \mathbb{R}^n \rightarrow X$.

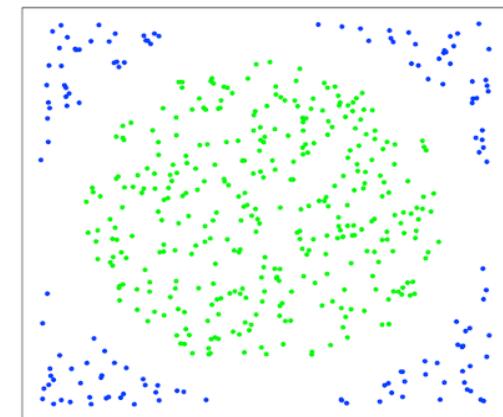
При этом отображение φ выбирается так, чтобы в новом пространстве \mathbf{X} выборка была *линейно* разделима

Классифицирующая функция F принимает вид

$$F(x) = \text{sign}(\langle w, \varphi(x) \rangle + b)$$

Выражение $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ называется ядром классификатора.

С математической точки зрения ядром может служить любая положительно определенная симметричная функция двух переменных. Положительная определенность необходимо для того, чтобы соответствующая функция Лагранжа в задаче оптимизации была ограничена снизу, т.е. задача оптимизации была бы корректно определена.



Чаще всего на практике встречаются следующие ядра:

Полиномиальное:

$$k(x, x') = (\langle x, x' \rangle + \text{const})^d$$

Радиальная базисная функция:

$$k(x, x') = e^{-\gamma \|x - x'\|^2}, \gamma > 0$$

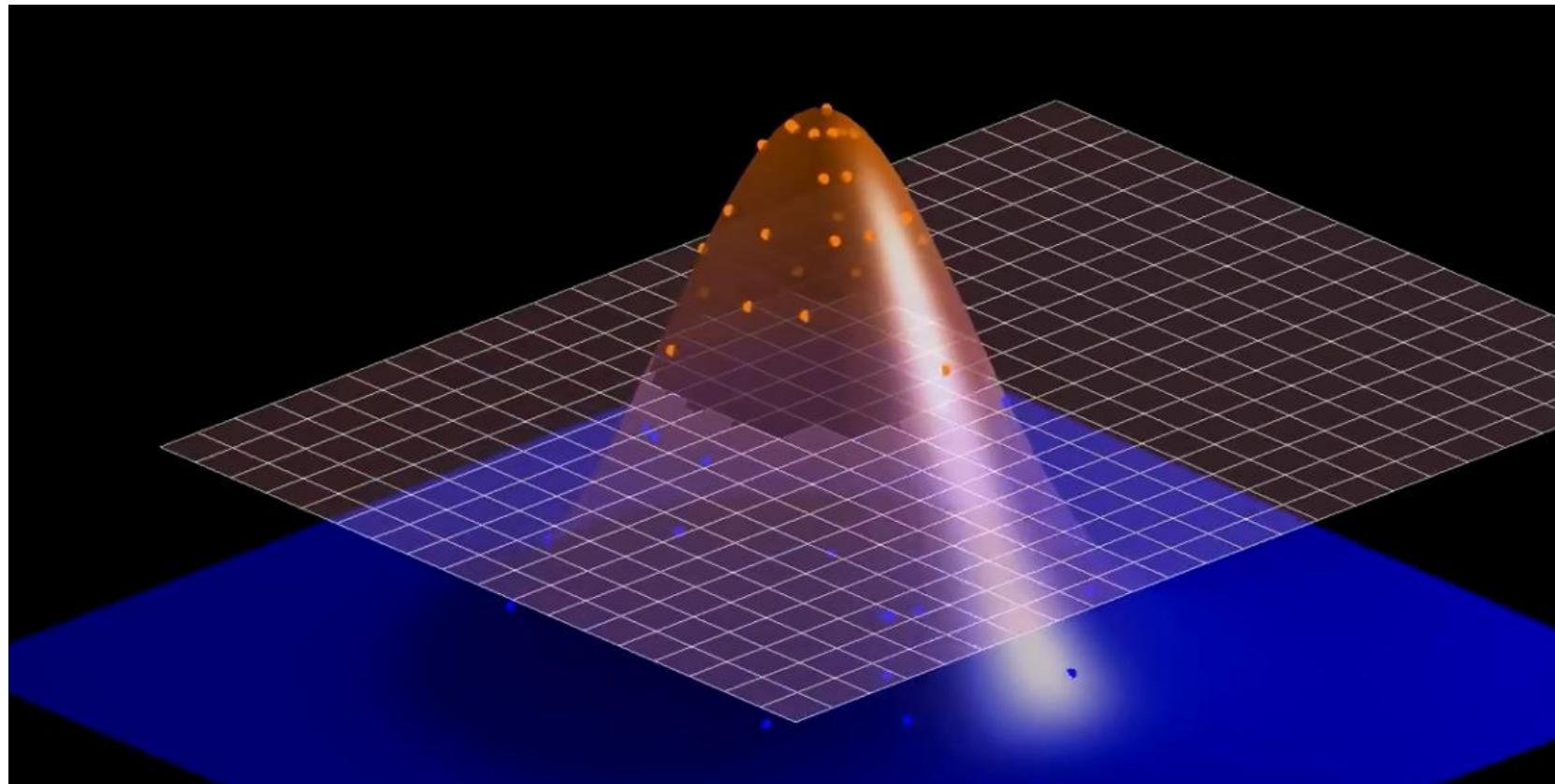
Гауссова радиальная базисная функция:

$$k(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$$

Сигмоид: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + c), \kappa > 0, c > 0$

Сложность построения SVM-модели заключается в том, что чем выше размерность пространства, тем сложнее с ним работать.

Один из вариантов работы с данными высокой размерности - это предварительное применение какого-либо метода понижения размерности данных для выявления наиболее существенных компонент, а затем использование метода опорных векторов.



Достоинства / недостатки SVM

Недостаток метода:

для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Достоинство метода: для классификации методом опорных векторов, в отличие от большинства других методов, достаточно небольшого набора данных.

При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных.

Метод опорных векторов позволяет:

- получить функцию классификации с минимальной верхней оценкой ожидаемого риска (уровня ошибки классификации);
- использовать линейный классификатор для работы с нелинейно разделяемыми данными, сочетая простоту с эффективностью.

Метод «ближайшего соседа»

(«nearest neighbour», «k-nearest neighbour»)

- Относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами.
- При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется.
- При данном подходе используется термин "*k-ближайший сосед*" ("*k-nearest neighbour*").
- Термин означает, что выбирается k "верхних" (ближайших) соседей для их рассмотрения в качестве множества "ближайших соседей".

Этапы подхода, основанного на прецедентах

Case Based Reasoning, CBR

сбор подробной информации о поставленной задаче;

сопоставление этой информации с деталями прецедентов, хранящихся в базе, для выявления аналогичных случаев;

выбор прецедента, наиболее близкого к текущей проблеме, из базы прецедентов ;

адаптация выбранного решения к текущей проблеме, если это необходимо;

проверка корректности каждого вновь полученного решения;

занесение детальной информации о новом прецеденте в базу прецедентов.

Достоинства / недостатки CBR

Недостатки метода:

- Данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт;
- Существует сложность выбора меры "близости" (метрики), также существует высокая зависимость результатов классификации от выбранной метрики;
- При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого - вычислительная трудоемкость;
- Типичные задачи данного метода - это задачи небольшой размерности по количеству классов и переменных.

Преимущества метода:

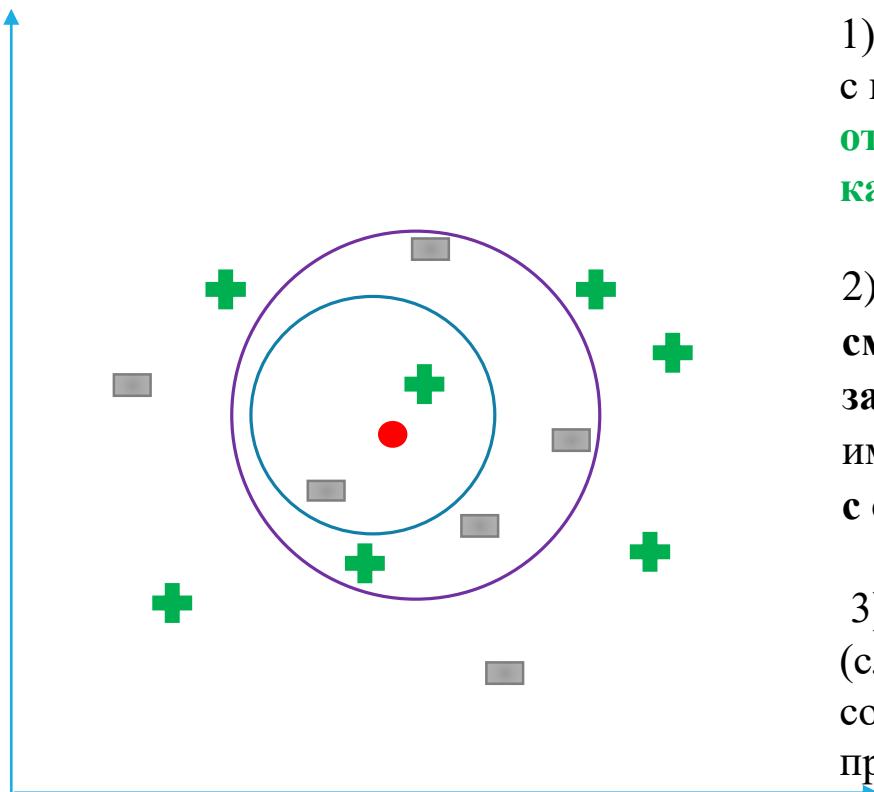
- Простота использования полученных результатов.
- Решения не уникальны для конкретной ситуации, возможно их использование для других случаев.
- Целью поиска является не гарантированно верное решение, а лучшее из возможных;
- С помощью данного метода решаются задачи классификации и регрессии.

Подход, основанный на прецедентах

Case Based Reasoning, CBR

- Вывод, основанный на прецедентах, представляет собой такой метод анализа данных, который делает заключения относительно данной ситуации по результатам поиска аналогий, хранящихся в базе прецедентов.
- Данный метод по сути относится к категории "**обучение без учителя**", (является "самообучающейся" технологией), благодаря чему рабочие характеристики каждой базы прецедентов с течением времени и накоплением примеров улучшаются.
- Разработка баз прецедентов по конкретной предметной области происходит на естественном для человека языке (может быть выполнена наиболее опытными сотрудниками компании - экспертами или аналитиками, работающими в данной предметной области).

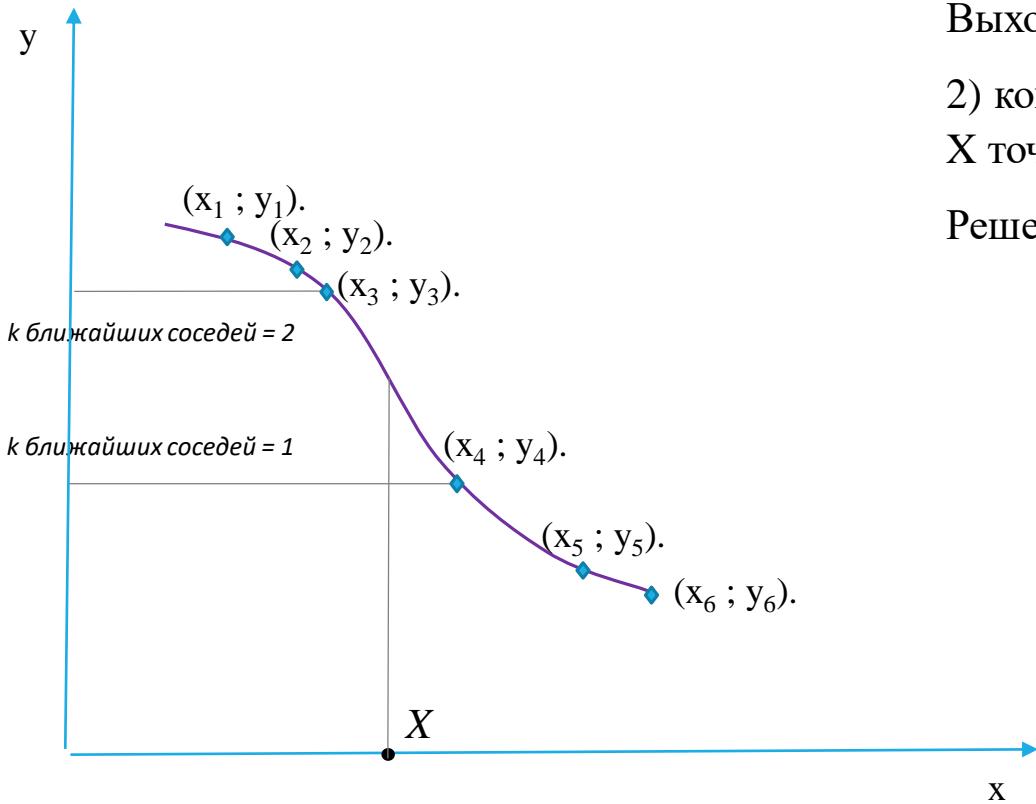
Решение задачи классификации новых объектов



- 1) Результат работы метода k -ближайших соседей с использованием одного ближайшего соседа:
отклик точки запроса будет классифицирован как знак плюс
- 2) Метод k -ближайших соседей (в случае 2) не сможет классифицировать отклик точки запроса, поскольку вторая ближайшая точка имеет знак минус и оба знака равнозначны (**победа с одинаковым количеством голосов**);
- 3) Результат работы метода k -ближайших соседей (случай 5 соседей): 2 точки со знаком "+" и 3 точки со знаком "-", алгоритм k -ближайших соседей присвоит знак "-" отклику точки запроса

Классификация объектов множества при разном значении параметра k

Решение задачи прогнозирования



1) метод k-ближайших соседей при $k = 1$, ищем набор примеров и выделяем из их числа ближайший к точке запроса X.

Выход Y равен y_4 ($Y = y_4$).

2) когда $k = 2$, выделяем уже две ближайшие к X точки (точки y_3 и y_4 соответственно).

Решение для Y в виде $Y = (y_3 + y_4)/2$.

Решение задачи прогнозирования осуществляется путем переноса описанных выше действий на использование произвольного числа ближайших соседей таким образом, что **выход Y точки запроса X вычисляется как среднеарифметическое значение выходов k-ближайших соседей точки запроса.**

Метод «ближайшего соседа» для задачи прогнозирования

- Независимые и зависимые переменные набора данных могут быть как непрерывными, так и категориальными.
- Для непрерывных зависимых переменных задача рассматривается как задача прогнозирования, для дискретных переменных - как задача классификации.
- Предсказание в задаче прогнозирования получается усреднением выходов k-ближайших соседей, а решение задачи классификации основано на принципе "*по большинству голосов*".

Метод «ближайшего соседа» для задачи прогнозирования

- **Критическим моментом** в использовании метода k-ближайших соседей является выбор параметра k . Он один из наиболее важных факторов, определяющих качество прогнозной либо классификационной модели.
- **Должно быть выбрано оптимальное значение параметра k** (это значение должно быть настолько большим, чтобы свести к минимуму вероятность неверной классификации, и одновременно, достаточно малым, чтобы k соседей были расположены достаточно близко к точке запроса).

Рассматриваем k как сглаживающий параметр, для которого должен быть найден компромисс между силой размаха (разброса) модели и ее смещенностью.

Метод «ближайшего соседа» для задачи прогнозирования

Оценка параметра k методом кросс-проверки

Один из вариантов оценки параметра k - проведение кросс-проверки (Bishop, 1995).

- **Кросс-проверка** - известный метод получения оценок неизвестных параметров модели. Основная идея метода - разделение выборки данных на v "складки". V "складки" здесь случайным образом выделенные изолированные подвыборки.
- По фиксированному значению k строится модель k -ближайших соседей для получения предсказаний на v -м сегменте (остальные сегменты при этом используются как примеры) и оценивается ошибка классификации.
- Для регрессионных задач наиболее часто в качестве оценки ошибки выступает сумма квадратов, а для классификационных задач удобней рассматривать точность (процент корректно классифицированных наблюдений).

Метод «ближайшего соседа» для задачи прогнозирования

- **Второй вариант выбора значения параметра k** - самостоятельно задать его значение. Однако этот способ следует использовать, если имеются обоснованные предположения относительно возможного значения параметра, например, предыдущие исследования сходных наборов данных.
- Метод k -ближайших соседей показывает достаточно неплохие результаты в самых разнообразных задачах.

Инструменты Data Mining, реализующих метод k-ближайших соседей и CBR-метод:

CBR Express и Case Point (Inference Corp.),

Apriori (Answer Systems), DP Umbrella (VYCOR Corp.),
KATE tools (Acknosoft, Франция),

Pattern Recognition Workbench (Unica, США),

а также некоторые статистические пакеты, например,
Statistica и др.

Литература

- B. Scholkopf, G. Ratsch, K. Muller, K. Tsuda, S. Mika An Introduction to Kernel-Based Learning Algorithms / IEEE Neural Networks, 12(2):181-201, May 2001
- Chickering D, Geiger D., Heckerman D. Learning Bayesian networks: The combination of knowledge and statistical data / Machine Learning. 1995. 20. P. 197-243
- Heckerman D Bayesian Networks for Data Mining / Data Mining and Knowledge Discovery. 1997. № 1. P. 79-119
- etc, Friedman N., Geiger D., Goldszmidt M. Bayesian Network Classifiers / Machine Learning. 1997. 29. P. 131-165
- Brand E., Gerritsen R / Naive-Bayes and Nearest Neighbor DBMS. 1998. № 7

Лекция 9.1

ПОИСК АССОЦИАТИВНЫХ ПРАВИЛ



Ассоциация - одна из задач Data Mining.

Целью поиска ассоциативных правил (association rule) является нахождение закономерностей между связанными событиями в базах данных.

Очень часто покупатели приобретают не один товар, а несколько. В большинстве случаев между этими товарами существует взаимосвязь. Так, например, покупатель, приобретающий макаронные изделия, скорее всего, захочет приобрести также кетчуп. Эта информация может быть использована для размещения товара на прилавках.



Объект 1

Объект 2

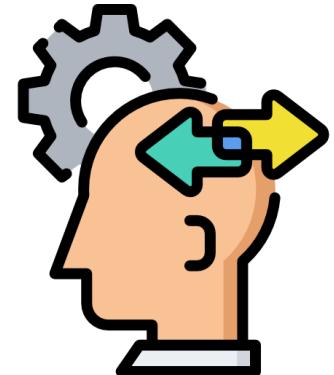
Найденная взаимосвязь

Сфера применения

- **розничная торговля:** определение товаров, которые стоит продвигать совместно; выбор местоположения товара в магазине; анализ потребительской корзины; прогнозирование спроса;
- **перекрестные продажи:** если есть информация о том, что клиенты приобрели продукты А, Б и В, то какие из них вероятнее всего купят продукт Г?
- **маркетинг:** поиск рыночных сегментов, тенденций покупательского поведения;
- **сегментация клиентов:** выявление общих характеристик клиентов компании, выявление групп покупателей;



Поиск шаблонов
поведения покупателей



Сегментация



Введение в ассоциативные правила

Рассмотрим основы на примере торговой компании.

- **Транзакция** - это множество событий, которые произошли одновременно.
- Регистрируя все бизнес-операции в течение всего времени своей деятельности, торговые компании накапливают огромные собрания транзакций.
- Каждая такая транзакция представляет собой набор товаров, купленных покупателем за один визит.
- **Транзакционная или операционная база данных** (Transaction database) представляет собой двумерную таблицу, которая состоит из номера транзакции (TID) и перечня покупок, приобретенных во время этой транзакции.
- **TID** - уникальный идентификатор, определяющий каждую сделку или транзакцию

Введение в ассоциативные правила

- На основе имеющейся базы данных нам нужно найти закономерности между событиями, то есть покупками.
- Ассоциативное правило состоит из двух наборов предметов, называемых условием и следствием, записываемых в виде $X \rightarrow Y$.**
- Допустим, имеется транзакционная база данных D. Присвоим значениям товаров переменные

TID	Приобретенные покупки	→	TID	Приобретенные покупки
100	Хлеб, молоко, печенье		100	a, b, c
200	Молоко, сметана		200	b, d
300	Молоко, хлеб, сметана, печенье		300	b, a, d, c
400	Колбаса, сметана		400	e, d
500	Хлеб, молоко, печенье, сметана		500	a, b, c, d
600	Конфеты		600	f

Введение в ассоциативные правила

- Рассмотрим набор товаров (Itemset), включающий, например, {Хлеб, молоко, печенье}.
- Выразим этот набор с помощью переменных:

$$abc=\{a,b,c\}$$

- Этот набор товаров встречается в нашей базе данных три раза, т.е. **поддержка этого набора товаров равна 3:**

$$SUP(abc)=3.$$

- При минимальном уровне поддержки, равной трем, набор товаров abc является часто встречающимся шаблоном.

Введение в ассоциативные правила

- **Поддержкой** называют количество или процент транзакций, содержащих определенный набор данных.
- Для данного набора товаров поддержка, выраженная в процентном отношении, равна 50%.

$$SUP(abc) = (3/6) * 100\% = 50\%$$

- Поддержку иногда также называют **обеспечением набора**.
- Таким образом, набор представляет интерес, если его поддержка выше определенного пользователем **минимального значения (min support)**.
- Эти наборы называют часто **встречающимися (frequent)**.

Поддержка ассоциативного правила

Рассмотрим правило «из покупки молока следует покупка печенья» для базы данных, которая была приведена ранее.

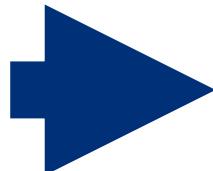
- **Поддержка ассоциативного правила – это число транзакций, которые содержат как условие, так и следствие.**

Например, для ассоциации $A \rightarrow B$ можно записать

$$support(A \rightarrow B) = S(A \rightarrow B) = P(A \cap B) =$$

= (количество транзакций, содержащих A и B)/(общее число транзакций)

- Правило имеет поддержку s , если $s\%$ транзакций из всего набора содержат одновременно наборы элементов A и B или, другими словами, содержат оба товара.
- *Молоко - это товар A, печенье - это товар B. Поддержка правила "из покупки молока следует покупка печенья" равна 3, или 50%.*



Достоверность правила

Достоверность правила показывает, какова вероятность того, что из события A следует событие B.

$$\text{confidence}(A \rightarrow B) = C(A \rightarrow B) = P(A|B) = P(A \cap B)/P(A) = \\ (\text{количество транзакций, содержащих } A \text{ и } B) / \\ (\text{количество транзакций, содержащих только } A)$$



Достоверность правила

- Правило "Из А следует В" справедливо с достоверностью с, если с% транзакций из всего множества, содержащих набор элементов А, также содержат набор элементов В.
- Число транзакций, содержащих молоко, равно четырем, число транзакций, содержащих печенье, равно трем, достоверность правила равна $(3/4)*100\%$, т.е. 75%.
- Достоверность правила "из покупки молока следует покупка печенья" равна 75%, т.е. 75% транзакций, содержащих товар А, также содержат товар В.



Значимость

- Методики поиска ассоциативных правил обнаруживают все ассоциации, которые удовлетворяют ограничениям на поддержку и достоверность, наложенным пользователем.
- Это приводит к необходимости **рассматривать десятки и сотни тысяч ассоциаций**, что делает **невозможным обработку** такого количества данных вручную.
- **Число правил желательно уменьшить** таким образом, чтобы **проанализировать только наиболее значимые** из них.
- Значимость часто вычисляется как разность между поддержкой правила и в целом и произведением поддержки только условия и поддержки только следствия.

Субъективные меры значимости

- **Лифт** – это отношение частоты появления условия в транзакциях, которые также содержат и следствие, к частоте появления следствия в целом.

$$\text{lift}(A \rightarrow B) = L(A \rightarrow B) = C(A \rightarrow B) / S(B).$$

- **Левередж** – это разность между наблюдаемой частотой, с которой условие и следствие появляются совместно (т.е. с поддержкой ассоциации), и произведением частот появления (поддержек) условия и следствия по отдельности.

$$T(A \rightarrow B) = S(A \rightarrow B) - S(A)S(B).$$

- **Улучшение** показывает, полезнее ли правило случайного угадывания. Если $I(A \rightarrow B) > 1$, это значит, что вероятнее предсказать наличие набора B с помощью правила, чем угадать случайно.

$$I(A \rightarrow B) = S(A \rightarrow B) / (S(A)S(B)).$$

Методы и алгоритмы поиска ассоциативных правил

- **Алгоритм AIS.** В алгоритме AIS кандидаты множества наборов генерируются и подсчитываются "на лету", во время сканирования базы данных.
- **Алгоритм SETM.** Создание этого алгоритма было мотивировано желанием использовать язык SQL для вычисления часто встречающихся наборов товаров. Как и алгоритм AIS, SETM также формирует кандидатов "на лету", основываясь на преобразованиях базы данных.

Неудобство алгоритмов AIS и SETM - излишнее генерирование и подсчета слишком многих кандидатов, которые в результате не оказываются часто встречающимися. Для улучшения их работы был предложен алгоритм **Apriori**.

Алгоритм Apriori

В основе алгоритма **Apriori** лежит понятие частого набора. Под частотой понимается простое количество транзакций в которых содержится данный предметный набор.

Частый предметный набор – предметный набор с поддержкой больше заданного порога либо равной ему. Этот порог называется минимальной поддержкой.

Работа данного алгоритма состоит из нескольких этапов, каждый из этапов состоит из следующих шагов:

1. формирование кандидатов;
2. подсчет кандидатов.

Алгоритм Apriori

Формирование кандидатов

(candidate generation) - этап, на котором алгоритм, сканируя базу данных, создает множество i -элементных кандидатов (i - номер этапа).

На этом этапе поддержка кандидатов не рассчитывается.



Подсчет кандидатов (candidate counting) - этап, на котором вычисляется поддержка каждого i -элементного кандидата. Здесь же осуществляется отсечение кандидатов, поддержка которых меньше минимума, установленного пользователем (min_sup).

Оставшиеся i -элементные наборы называем часто встречающимися.

Алгоритм Apriori

- Чтобы сократить пространство поиска ассоциативных правил, алгоритм Apriori использует свойство **антимонотонности**.
- Свойство утверждает, что если предметный набор Z не является частым, то добавление некоторого нового предмета A к набору Z не делает его более частым. Данное полезное свойство позволяет значительно уменьшить пространство поиска ассоциативных правил.
- На первом этапе алгоритма Apriori формируются частые однопредметные наборы – множество F_1 .

Алгоритм Apriori

- Для поиска F_k , то есть k -предметных наборов, алгоритм Apriori сначала создает множество F_k кандидатов в k -предметные наборы путем связывания множества F_{k-1} с самим собой.
- Затем F_k сокращается с использованием свойства антимонотонности.
- Предметные наборы множества F_k , которые остались после сокращения, формируют F_k .

Алгоритм Apriori

После того, как все частые предметные наборы найдены, можно переходить к *генерации на их основе ассоциативных правил*.

Для этого к каждому частому предметному набору s можно применить процедуру, состоящую из 2 шагов.

1). Генерируются все возможные поднаборы s .

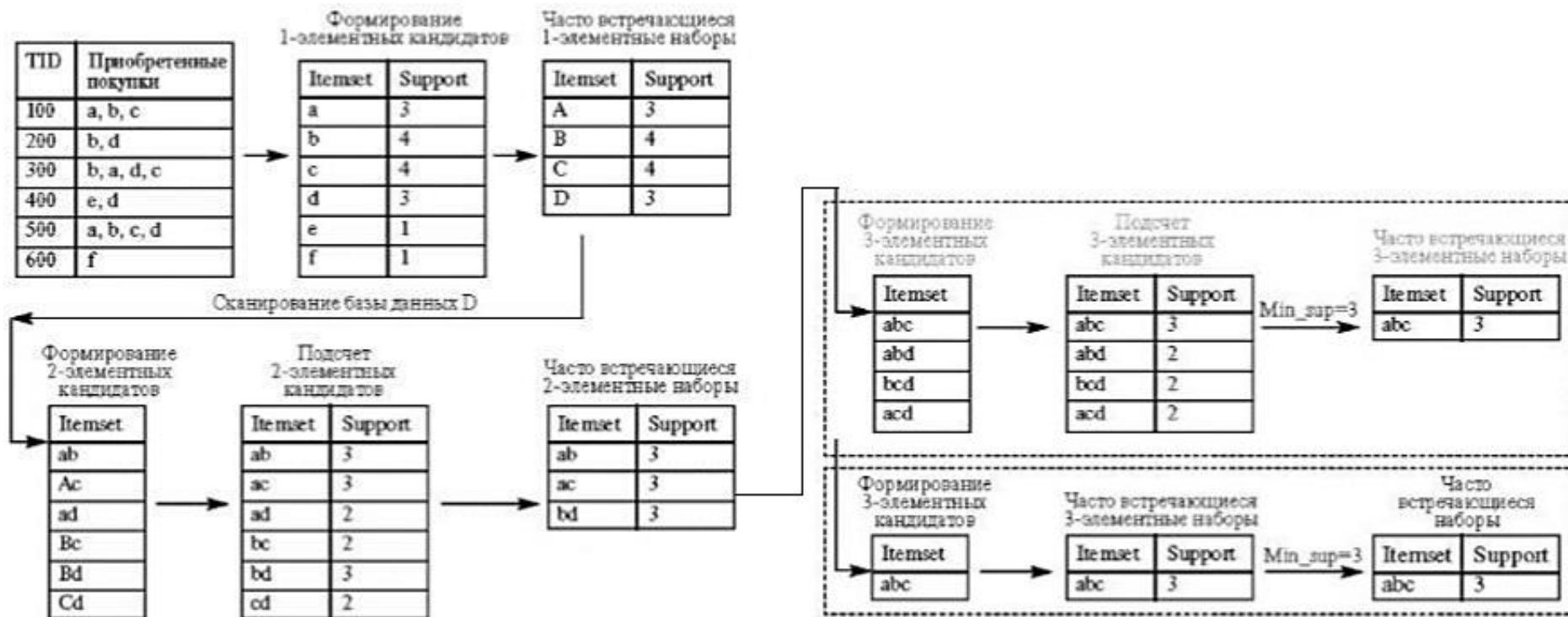
2). Если поднабор ss является непустым поднабором s , то рассматривается ассоциация $R: ss \rightarrow (s - ss)$, где $s - ss$ представляет собой набор s без поднабора ss .

R считается ассоциативным правилом, если удовлетворяет условию заданного минимума поддержки и достоверности.

Данная процедура повторяется для каждого подмножества ss из s .

Алгоритм Apriori

Рассмотрим работу алгоритма Apriori на примере базы данных D.
Минимальный уровень поддержки равен 3.



Разновидности Apriori

- **AprioriTid.** Интересная особенность этого алгоритма - то, что база данных D не используется для подсчета поддержки кандидатов набора товаров после первого прохода.
- С этой целью используется кодирование кандидатов, выполненное на предыдущих проходах.
- В последующих проходах размер закодированных наборов может быть намного меньше, чем база данных, и таким образом экономятся значительные ресурсы.

Разновидности Apriori

- **AprioriHybrid.** Анализ времени работы алгоритмов Apriori и AprioriTid показывает, что в более ранних проходах Apriori добивается большего успеха, чем AprioriTid.
- Однако AprioriTid работает лучше Apriori в более поздних проходах. Кроме того, они используют одну и ту же процедуру формирования наборов-кандидатов.
- Основанный на этом наблюдении, алгоритм AprioriHybrid предложен, чтобы объединить лучшие свойства алгоритмов Apriori и AprioriTid.
- AprioriHybrid использует алгоритм Apriori в начальных проходах и переходит к алгоритму AprioriTid, когда ожидается, что закодированный набор первоначального множества в конце прохода будет соответствовать возможностям памяти.
- Однако, переключение от Apriori до AprioriTid требует вовлечения дополнительных ресурсов.

Разновидности Apriori. AprioriHybrid.

- **Алгоритм DHP**, также называемый алгоритмом хеширования (J. Park, M. Chen and P. Yu, 1995 год). В основе его работы - вероятностный подсчет наборов-кандидатов, осуществляемый для сокращения числа подсчитываемых кандидатов на каждом этапе выполнения алгоритма Apriori.
- К другим усовершенствованным алгоритмам относятся: PARTITION, DIC, алгоритм "выборочного анализа".
- **PARTITION алгоритм** (A. Savasere, E. Omiecinski and S. Navathe, 1995 год). Этот алгоритм разбиения (разделения) заключается в сканировании транзакционной базы данных путем разделения ее на непересекающиеся разделы, каждый из которых может уместиться в оперативной памяти.
- **Алгоритм DIC**, Dynamic Itemset Counting (S. Brin R. Motwani, J. Ullman and S. Tsur, 1997 год). Алгоритм разбивает базу данных на несколько блоков, каждый из которых отмечается так называемыми "начальными точками" (start point), и затем циклически сканирует базу данных.

Ссылки на используемые источники:

Apriori — масштабируемый алгоритм поиска ассоциативных правил

Agrawal R, Imielinski T, Swami AN. «Mining Association Rules between Sets of Items in Large Databases.» SIGMOD. June 1993, 22(2):207-16

Market Basket Analysis

Выводы

- **Задачей поиска ассоциативных правил** является определение часто встречающихся наборов объектов в большом множестве наборов.
- **Секвенциальный анализ заключается** в поиске частых последовательностей. Основным отличием задачи секвенциального анализа от поиска ассоциативных правил является установление отношения порядка между объектами.
- **Наличие иерархии в объектах** и ее использование в задаче поиска ассоциативных правил позволяет выполнять более гибкий анализ и получать дополнительные знания.

Результаты решения задачи представляются в виде ассоциативных правил, условная и заключительная часть которых содержит наборы объектов.

Основными характеристиками ассоциативных правил являются поддержка, достоверность и улучшение.

Поддержка (support) показывает, какой процент транзакций поддерживает данное правило.

Достоверность (confidence) показывает, какова вероятность того, что из наличия в транзакции набора условной части правила следует наличие в ней набора заключительной части.

Улучшение (improvement) показывает, полезнее ли правило случайного угадывания.

Задача поиска ассоциативных правил решается в два этапа.
На первом выполняется поиск всех частых наборов объектов.
На втором из найденных частых наборов объектов
генерируются ассоциативные правила.

Алгоритм Apriori использует одно из свойств поддержки,
гласящее: поддержка любого набора объектов не может
превышать минимальной поддержки любого из его
подмножеств.

Лекция 9.2

БАЙЕСОВСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ

Байесовские методы. Байесовская классификация



Классификация технологических методов ИАД [1]

Байесовские методы. Байесовская классификация

Суть методов кросс-табуляции

Кросс-табуляция является простой формой анализа, широко используемой в генерации отчетов средствами систем оперативной аналитической обработки (OLAP).

Двумерная кросс-таблица представляет собой матрицу значений, каждая ячейка которой лежит на пересечении значений атрибутов.

Расширение идеи кросс-табличного представления на случай гиперкубической информационной модели является основой многомерного анализа данных, поэтому эта группа методов рассматривается как симбиоз многомерного оперативного анализа и интеллектуального анализа данных.

К методам ИАД группы кросс-табуляции относится также использование байесовских сетей (*Bayesian Networks*).

Байесовские методы. Байесовские сети

Байесовская сеть (или байесова сеть, байесовская сеть доверия) - **графическая вероятностная модель, представляющая собой множество переменных и их вероятностных зависимостей.**

Математический аппарат байесовых сетей создан американским ученым Джудой Перлом, лауреатом Премии Тьюринга (2011 г.).

Формально, **байесовская сеть** - это направленный ациклический граф, каждой вершине которого соответствует случайная переменная, а дуги графа кодируют отношения условной независимости между этими переменными. Вершины могут представлять переменные любых типов, быть взвешенными параметрами, скрытыми переменными или гипотезами.

Так как **байесовская сеть** - это **полная модель** для переменных и их отношений, она может быть использована для того, чтобы давать ответы на вероятностные вопросы.

Это процесс вычисления апостериорного распределения переменных по переменным-свидетельствам называют вероятностным выводом, что дает универсальную оценку для приложений, где нужно выбрать значения подмножества переменных, которое минимизирует функцию потерь, например, вероятность ошибочного решения.

Байесовские методы. Байесовские сети

Байесовская сеть позволяет получить ответы на следующие типы вероятностных запросов:

- 1) нахождение вероятности свидетельства,
- 2) определение априорных маргинальных вероятностей,
- 3) определение апостериорных маргинальных вероятностей, включая:
 - *прогнозирование*, или прямой вывод - определение вероятности события при наблюдаемых причинах),
 - *диагностирование*, или обратный вывод (абдукция), - определение вероятности причины при наблюдаемых следствиях,
 - *межпричинный (смешанный) вывод* или трансдукция, - определение вероятности одной из причин наступившего события при условии наступления одной или нескольких других причин этого события.
- 4) вычисление наиболее вероятного объяснения наблюдаемого события,
- 5) вычисление апостериорного максимума.

Байесовские методы. Байесовская классификация

- Ранее байесовская классификация использовалась для формализации знаний экспертов в экспертных системах.
- В настоящее время байесовская классификация применяется и в качестве одного из методов Data Mining.
- Байесовские методы получили достаточно широкое распространение и активно используются в самых различных областях знаний.

История вопроса: формула Байеса была опубликована в 1763 году спустя 2 года после смерти Томаса Байеса.

Методы, использующие ее, получили широкое распространение только к концу 20 века - расчеты требуют определенных вычислительных затрат, и они стали возможны лишь с развитием информационных технологий.

Байесовские методы. Байесовская классификация

Байесовский метод опирается на теорему о том, что если плотности распределения классов известны, то алгоритм классификации, имеющий минимальную вероятность ошибок, можно выписать в явном виде.

Для оценивания плотностей классов по выборке применяются различные подходы (в частности, параметрический, непараметрический и оценивание смесей распределений).

В *байесовских классификаторах* используется критерий, минимизирующий вероятность принятия ошибочного решения, поэтому байесовские алгоритмы являются статистически оптимальными.

Но при этом алгоритмы требуют в идеале полного знания многомерных функций распределения наблюдаемых признаков для каждого класса. Необходимость такого знания обусловлена использованием *формулы Байеса*, которая лежит в основе *байесовских методов* принятия решения.

Байесовский подход основан на предположении, что задача выбора решения сформулирована в терминах теории вероятностей и известны все представляющие интерес вероятностные величины. В основе байесовской классификации лежит *правило Байеса*.

Байесовская классификация. Постановка задачи

Рассмотрим обучающую выборку из n объектов, каждый из которых принадлежит одному из K классов и характеризуется набором m числовых признаков a_1, a_2, \dots, a_m .

Пусть имеется n_k объектов k -ого класса, так что

$$N = \sum n_k \quad K_k = 1.$$

Значение j -ого признака i -ого объекта из k -ого класса обозначим x_{ijk} .

Тогда этот объект можно охарактеризовать вектором-строкой $x_{ik} = (x_{i1k}, \dots, x_{ijk}, \dots, x_{imk})$. Эту строку будем рассматривать как i -ю реализацию векторной случайной величины ξ_k , подчиняющейся распределению вероятностей с плотностью $p(x_1, \dots, x_m | k)$, своей для каждого класса k .

Пусть теперь наблюдается объект, для которого необходимо определить, к какому классу он относится. Объект характеризуется только набором m числовых признаков x_1, \dots, x_m .

Байесовская классификация. Общая структура байесовского классификатора

В основе классификатора лежит следующее *правило*. Классификатор вычисляет *апостериорную вероятность* $P(k|x)$ каждого класса k , которому может принадлежать испытуемый объект, и относит этот объект к апостериорно наиболее вероятному классу \hat{k} :

$$\hat{k} = \arg \max_k \ln P(k|x_1, \dots, x_m)$$

Апостериорная вероятность вычисляется по *формуле Байеса*:

$$P(k|x_1, \dots, x_m) = P(k)p(x_1, \dots, x_m|k)/p(k),$$

где $P(k)$ – априорная вероятность того, что объект относится к k -ому классу, $p(k)$ и $p(x_1, \dots, x_m|k)$ - безусловная и условная многомерные плотности распределения вектора признаков, компоненты которого обычно статистически зависимы.

Таким образом, байесовский классификатор предполагает, что многомерная совместная плотность распределения признаков известна для всех классов.

Байесовская классификация. Общая структура байесовского классификатора

Аналитическое представление многомерной плотности вероятности известно только для нормального распределения.

При этом многомерная нормальная плотность распределения дает подходящую модель для одного важного случая, а именно когда значения векторов признаков x для данного класса k представляются непрерывнозначными, слегка искаженными версиями единственного типичного вектора, или вектора-прототипа, μ_k .

Именно этого ожидают, когда классификатор выбирается так, чтобы выделять те признаки, которые, будучи различными для образов, принадлежащих различным классам, были бы, возможно, более схожи для образов из одного и того же класса.

Байесовская классификация. Общая структура байесовского классификатора

Многомерная нормальная плотность распределения в общем виде представляется выражением

$$p(x) = \frac{1}{(2\pi)^{\frac{m}{2}} \det R^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T R^{-1}(x-\mu)}$$

где μ – m -компонентный вектор среднего значения,
 R – ковариационная матрица размера $m \times m$,
 T – знак транспонирования.

Если все недиагональные элементы равны нулю, то $p(x)$ сводится к произведению одномерных нормальных плотностей компонент вектора x .

Байесовская классификация. Общая структура байесовского классификатора

Для многомерного нормального распределения удаётся выразить в аналитически замкнутой форме (с точностью до несущественных слагаемых) алгоритм байесовской классификации:

$$\hat{k} = \arg \max_k \left(\ln P(k) - \frac{1}{2} \ln \det R_k - \frac{1}{2} (x_k - \mu_k) R_k^{-1} (x_k - \mu_k)^T \right)$$

где μ_k – m -вектор-строка математических ожиданий значений признаков объектов класса k ,

R_k – $m \times m$ -матрица ковариаций векторов признаков класса k .

Диагональные элементы матрицы образуют m -вектор D_k дисперсий признаков объектов класса k .

Алгоритм байесовской классификации с обучением состоит из двух этапов:

- 1) этап обучения;
- 2) этап классификации

Байесовские методы. Байесовская классификация

Достоинства байесовских сетей как метода Data Mining

- в модели определяются зависимости между всеми переменными, это позволяет легко обрабатывать ситуации, в которых значения некоторых переменных неизвестны;
- байесовские сети достаточно просто интерпретируются и позволяют на этапе прогностического моделирования легко проводить анализ по сценарию «что, если»;
- байесовский метод позволяет естественным образом совмещать закономерности, выведенные из данных, и, например, экспертные знания, полученные в явном виде;
- использование байесовских сетей позволяет избежать проблемы переучивания (overfitting), то есть избыточного усложнения модели, что является слабой стороной многих методов (например, деревьев решений и нейронных сетей).

Байесовские методы. Наивно-байесовский подход (Naive-Bayes Approach)

Суть метода наивно-байесовской классификации

В наивном байесовском классификаторе делается предположение о независимости признаков объекта. Если пренебречь статистическими связями между компонентами вектора признаков, тогда матрица R_k будет диагональной с вектором D_k на главной диагонали и классификатор станет **наивным байесовским классификатором**.

Также предполагается, что **маргинальная плотность распределения** $p(x_j | k)$ любого признака является нормальной для любого класса.

Но на практике так бывает далеко не всегда, то есть наблюдаемые данные не подчиняются нормальному закону распределения (в общем случае закон вообще неизвестен) и имеет место статистическая зависимость, поэтому область применения классификатора сужается.

Байесовские методы. Наивно-байесовский подход (Naive-Bayes Approach)

Наивно-байесовский подход имеет следующие недостатки:

- перемножать условные вероятности корректно только тогда, когда все входные переменные действительно статистически независимы;
- хотя часто данный метод показывает достаточно хорошие результаты при несоблюдении условия статистической независимости, но теоретически такая ситуация должна обрабатываться более сложными методами, основанными на обучении байесовских сетей;
- невозможна непосредственная обработка непрерывных переменных - требуется их преобразование к интервальной шкале, чтобы атрибуты были дискретными; однако такие преобразования иногда могут приводить к потере значимых закономерностей;
- на результат классификации в наивно-байесовском подходе влияют только индивидуальные значения входных переменных, комбинированное влияние пар или троек значений разных атрибутов здесь не учитывается.

Оптимальный байесовский классификатор

Так как **оптимальный байесовский классификатор** является модификацией наивного байесовского классификатора, то в качестве решающего правила также берётся рассмотренная ранее формула.

Одна из идей оптимизации наивно-байесовского классификатора состоит в том, чтобы, максимально используя обучающую выборку и гауссову копуля-функцию, обойти два «наивных» предположения.

Модификация позволяет:

- 1) учесть статистические связи между наблюдаемыми признаками;
- 2) адаптировать классификатор к неизвестному действительному распределению путем приведения сглаженных маргинальных функций распределения признаков к нормальному виду.

Другими словами, с помощью нелинейных гауссовых копула-функций негауссовы данные преобразуются в гауссовы, которые можно подавать на вход классификатору.

Литература

1. Л. В. Щавелёв Способы аналитической обработки данных для поддержки принятия решений. (СУБД. - 1998. - № 4-5)
2. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин) URL: <http://www.ccas.ru/voron>
3. Chickering D, Geiger D., Heckerman D. Learning Bayesian networks: The combination of knowledge and statistical data / Machine Learning. 1995. 20. P. 197-243
4. Heckerman D Bayesian Networks for Data Mining / Data Mining and Knowledge Discovery. 1997. № 1. P. 79-119
5. etc, Friedman N., Geiger D., Goldszmidt M. Bayesian Network Classifiers / Machine Learning. 1997. 29. P. 131-165
6. Brand E., Gerritsen R / Naive-Bayes and Nearest Neighbor DBMS. 1998. № 7