Интеллектуальные системы

Лекция 1

Интеллектуальная система (ИС, англ. intelligent system). Определения

- Из толкового словаря по искусственному интеллекту

Аверкин А. Н., Гаазе-Рапопорт М. Г., Поспелов Д. А. 1992
https://ru.wikipedia.org/wiki/Интеллектуальная_система

- Это техническая или программная система, способная решать задачи, традиционно считающиеся творческими, принадлежащие конкретной предметной области, знания о которой хранятся в памяти такой системы. Структура интеллектуальной системы включает три основных блока — базу знаний, механизм вывода решений и интеллектуальный интерфейс. (Близко к определению Экспертной системы по структуре и по области применимости, но есть наличие понятия творчества, а оно довольно широкое)
- В технологиях принятия решений
- Это информационно-вычислительная система с интеллектуальной поддержкой, решающая задачи без участия человека — лица, принимающего решение (ЛПР), в отличие от интеллектуализированной системы, в которой оператор присутствует. (отсюда опосредованно следует присутствие объекта управления)

- Энциклопедия эпистемологии и философии науки. http://epistemology_of_science.academic.ru/253/интеллектуальные_системы
- ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ (от лат. intellectus — ум, рассудок) — компьютерные системы, которые реализуют некоторые черты человеческого интеллекта, дающие возможность осиливать трудные задачи, решение которых человеком в реальное время невозможно. И. С. являются основным продуктом исследований искусственного интеллекта. Характерной особенностью компьютерных систем, являющихся интеллектуальными, считается использование как подсистемы представления знаний и фактов, так и подсистемы Решателя задач, который реализует когнитивные рассуждения.
- Довольно расплывчатое определение. Имеются ли в виду NP сложные задачи (в чем трудность), какие черты человеческого интеллекта, и каким человеком за реальное время не решаются, хотя здесь дана ссылка на задачи реального времени и соответственно опосредованная отсылка к адаптивному управлению.

Определение ИС данное на Интуите с ссылкой на Советский энциклопедический словарь М.: Советская энциклопедия, 1983. 1600 с., под ред. Прохорова А.М.

- Адаптивная система - это система, которая сохраняет работоспособность при непредвиденных изменениях свойств управляемого объекта, целей управления или окружающей среды путем смены алгоритма функционирования, программы поведения или поиска оптимальных, в некоторых случаях просто эффективных, решений и состояний. Традиционно, по способу адаптации различают самонастраивающиеся, самообучающиеся и самоорганизующиеся системы.

- Под алгоритмом понимается последовательность заданных действий, которые однозначно определены и выполнимы на современных ЭВМ за приемлемое время для решаемой задачи.

- ИС - адаптивная система, позволяющая строить программы целесообразной деятельности по решению поставленных перед ними задач на основании конкретной ситуации, складывающейся на данный момент в окружающей их среде.

- Интеллектуальная система - синоним термина Умная система - это сложная система, способная воспринимать, сравнивать, преобразовывать, создавать и хранить внутри себя модели определенных объектов

- Интеллектуальная система — совокупность взаимодействующих между собой относительно элементарных структур и процессов, объединенных в целое выполнением функции интеллекта (целенаправленного, опосредованного и обобщенного познания, активного отражения объективной реальности, логического и творческого мышления), несводимой к функции ее компонентов. Признаки ИС: 1) взаимодействует со средой и другими системами как единое целое, 2) состоит из иерархии подсистем более низкого уровня (Еремин А.Л., 2005 Ноогенез и теория интеллекта, для гуманитарных направлений)

Московский Государственный Университет им. М.В. Ломоносова.
Журнал «Интеллектуальные системы»

Примеры статей:

- Кончук Д.О., Окуловский Ю.С. Универсальная модель алгоритмов коллективного разума и ее реализация
- Рассмотрена модель алгоритмов коллективного разума, обобщающая известные алгоритмы искусственных иммунных систем, муравьиные алгоритмы, метод роя частиц и т.д. Данная модель является общей для многих видов алгоритмов коллективного разума, и позволяет создать единую формализацию этих методов. На базе данной модели создана программная реализация, позволяющая быстро разрабатывать алгоритмы коллективного разума, сравнивать различные методы, распараллеливать их выполнение, и т.д.

- Козлов В.Н. Геометрический подход к распознаванию зрительных образов (краткий обзор)

Статья, как это следует из названия, представляет собой обзор по дискретно-геометрическому подходу к распознаванию изображений.

- Дергунов А.В. База знаний повышения производительности MPI-приложений
- Для анализа MPI программ наиболее часто используют программные системы, которые осуществляют сбор и визуализацию трассы выполнения программы. Но при использовании таких инструментов пользователь сталкивается с проблемой анализа больших объемов информации. Другой проблемой является то, что часто встречающиеся ситуации, приводящие к потерям производительности MPI программ, явно не визуализируются, т.е. отсутствует база знаний повышения производительности. Поэтому возникает потребность в средствах, которые бы автоматизировали анализ трассы и подсказали пользователю, как повысить производительность его программы. В данной работе описывается программная система, выполняющая эту задачу. Ключевым компонетом этой системы является база знаний причин недостаточных производительности MPI приложений, которая состоит из правил, описанных на специальном языке, разработанном в рамках этой системы.

Основные направления Журнала «Интеллектуальные системы»

- распознавание образов;
- интеллектуальное программное обеспечение;
- базы данных и знаний;
- принципы принятия решений;
- экспертные системы и решатели задач;
- биоинформатика;
- автоматы и роботы с элементами искусственного интеллекта;
- компьютеры и нейрокомпьютеры;
- нечеткая математика и ее приложения;

К сфере решаемых ИС задач относятся задачи, обладающие, как правило, следующими особенностями:

- в них неизвестен алгоритм решения задач (такие задачи называют интеллектуальными задачами);
- в них используется помимо традиционных данных в числовом формате информация в виде изображений, рисунков, знаков, букв, слов, звуков;
- в них предполагается наличие выбора (не существует алгоритма - это значит, что нужно сделать выбор между многими вариантами в условиях неопределенности). Свобода действий является существенной составляющей интеллектуальных задач.

ИС - признаки

- Техническая или программная система позволяющая решать поставленные перед ней задачи, алгоритм решения которых заранее неизвестен, в изменяющихся условиях возникающих в данной предметной области за актуальное время. А также способная отображать состояние этой ПО на основе анализа и обработки поступающей неформализованной информации (данных). (последнее предложение близко к части задач решаемых Методами Машинного обучения)

- Виды интеллектуальных систем
- Интеллектуальная информационная система
- Экспертная система
- Расчётно-логические системы
- Гибридная интеллектуальная система
- Рефлекторная интеллектуальная система
- К расчётно-логическим системам относят системы, способные решать управленческие и проектные задачи по декларативным описаниями условий. При этом пользователь имеет возможность контролировать в режиме диалога все стадии вычислительного процесса. Данные системы способны автоматически строить математическую модель задачи и автоматически синтезировать вычислительные алгоритмы по формулировке задачи. Эти свойства реализуются благодаря наличию базы знаний в виде функциональной семантической сети и компонентов дедуктивного вывода и планирования.

- Рефлекторная система — это система, которая формирует вырабатываемые специальными алгоритмами ответные реакции на различные комбинации входных воздействий. Алгоритм обеспечивает выбор наиболее вероятной реакции интеллектуальной системы на множество входных воздействий, при известных вероятностях выбора реакции на каждое входное воздействие, а также на некоторые комбинации входных воздействий. Данная задача подобна той, которую реализуют нейронные сети.

Экспертные системы

- Заменяют человека эксперта в какой-то предметной области

Обычно основаны на продукционных моделях знаний, исчислении высказываний и предикатов

Содержат – базу фактов, машину вывода, систему объяснений выводов, интерфейс

Могут работать с нечеткими знаниями и использовать Байесовский подход, коэффициенты уверенности, теорию Демпстера-Шеффера, нечеткие множества

Например:

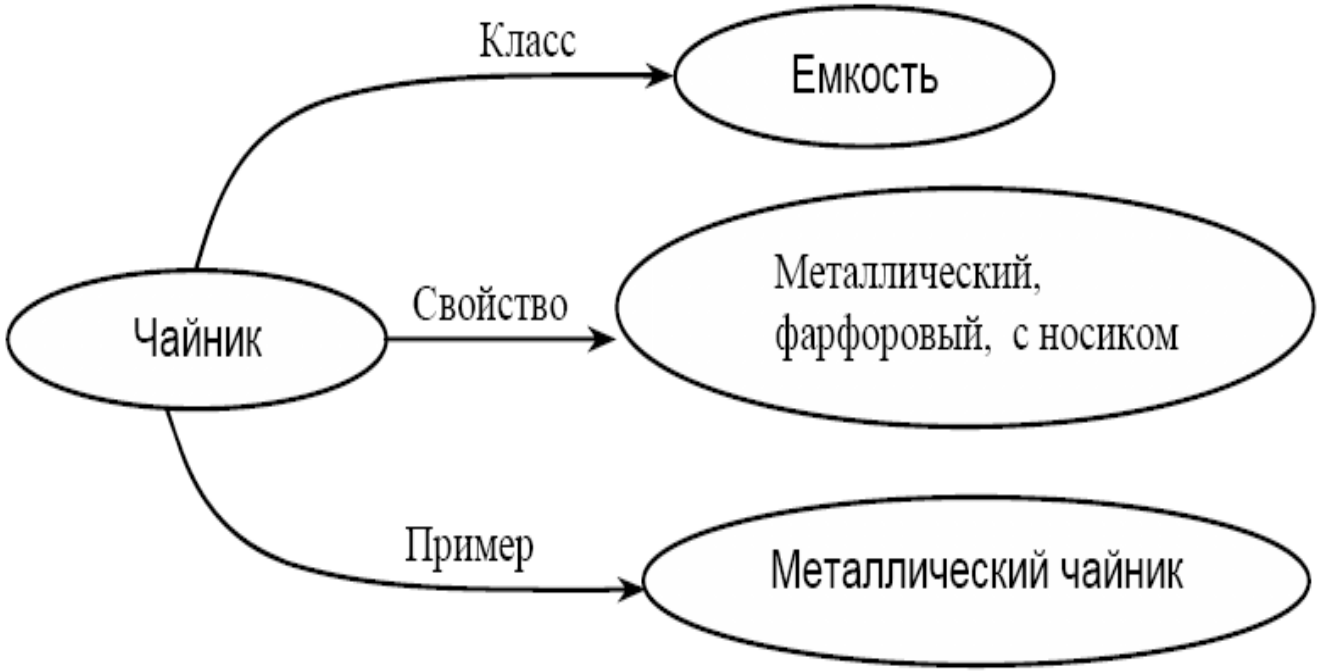
CLIPS - оболочка для построения ЭС, с оптимизированным алгоритмом выбора продукций (сопоставления с образцом, Rete)

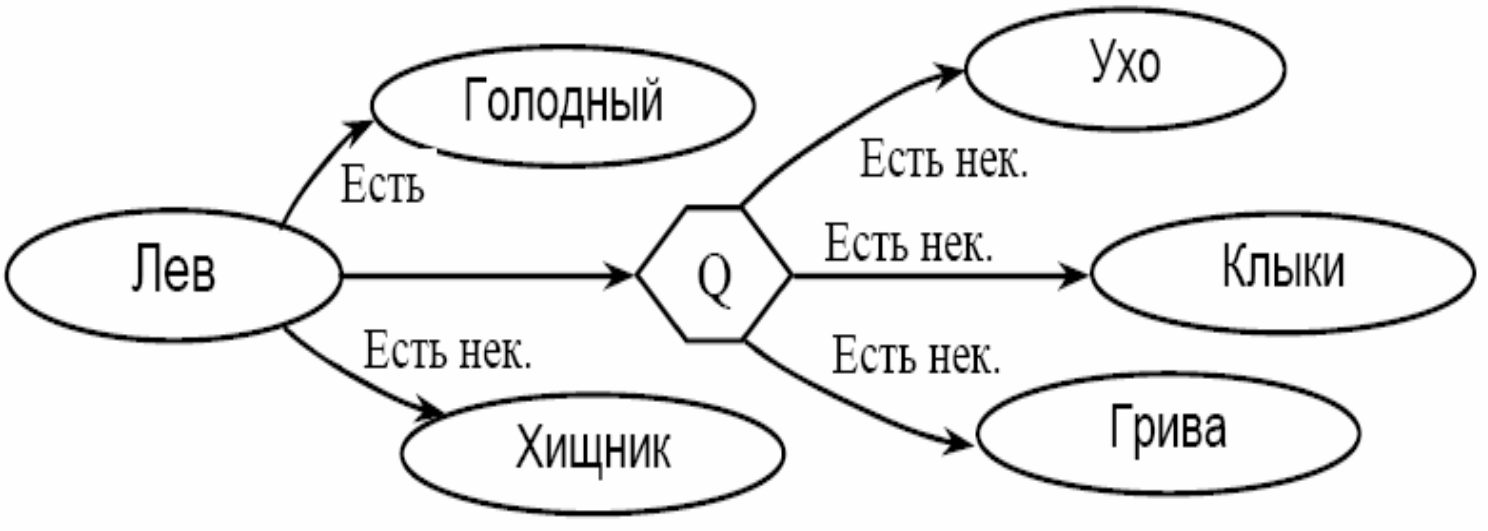
Cyc (Сайк) — проект по созданию объёмной онтологической базы знаний, позволяющей программам решать сложные задачи из области искусственного интеллекта на основе логического вывода и привлечения здравого смысла.

MYCIN - наиболее известная диагностическая система, которая предназначена для диагностики и наблюдения за состоянием больного при менингите и бактериальных инфекциях.

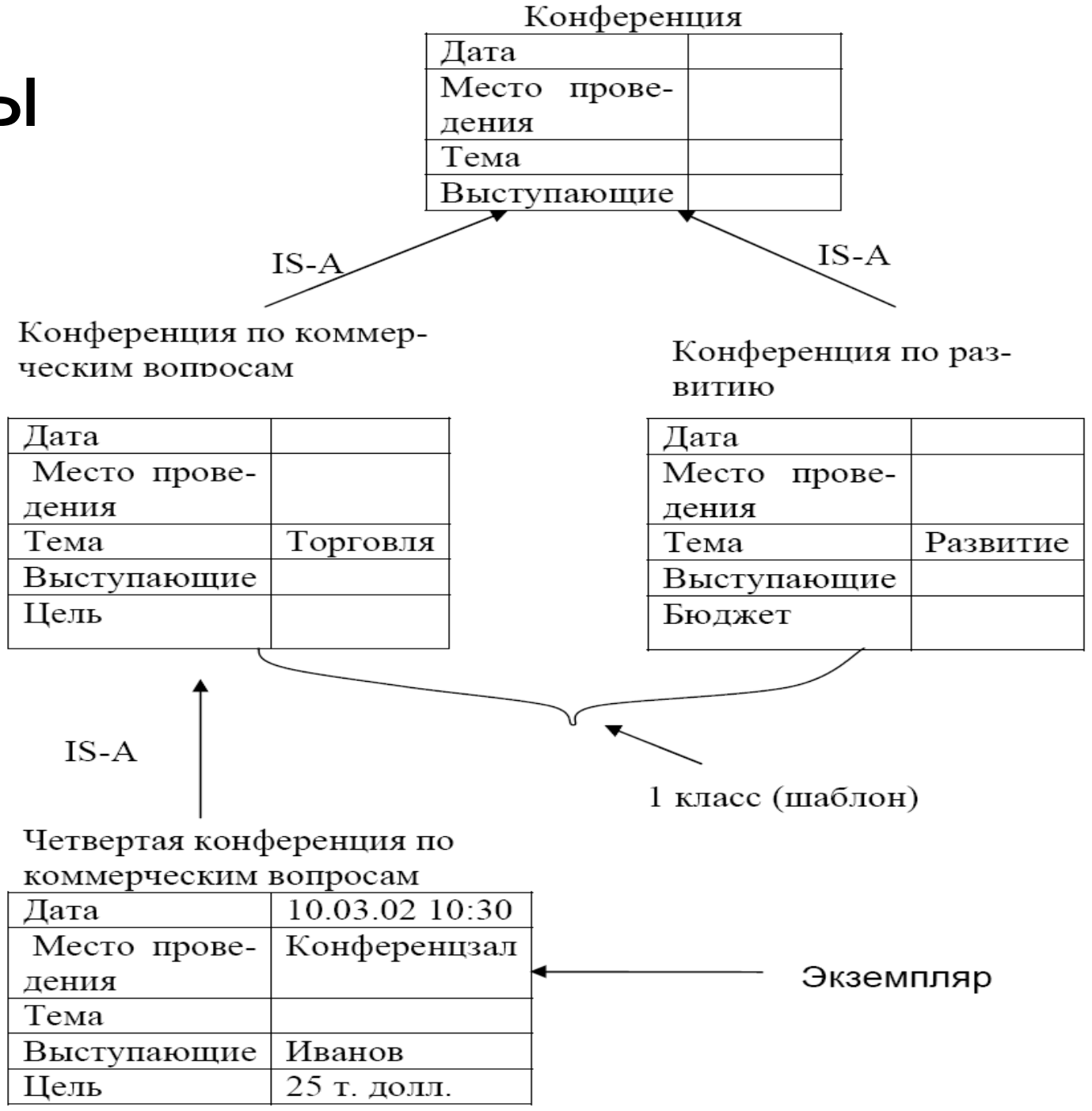

Компьютерные системы
Интеллектуальные системы
Системы, основанные на знаниях
Экспертные системы

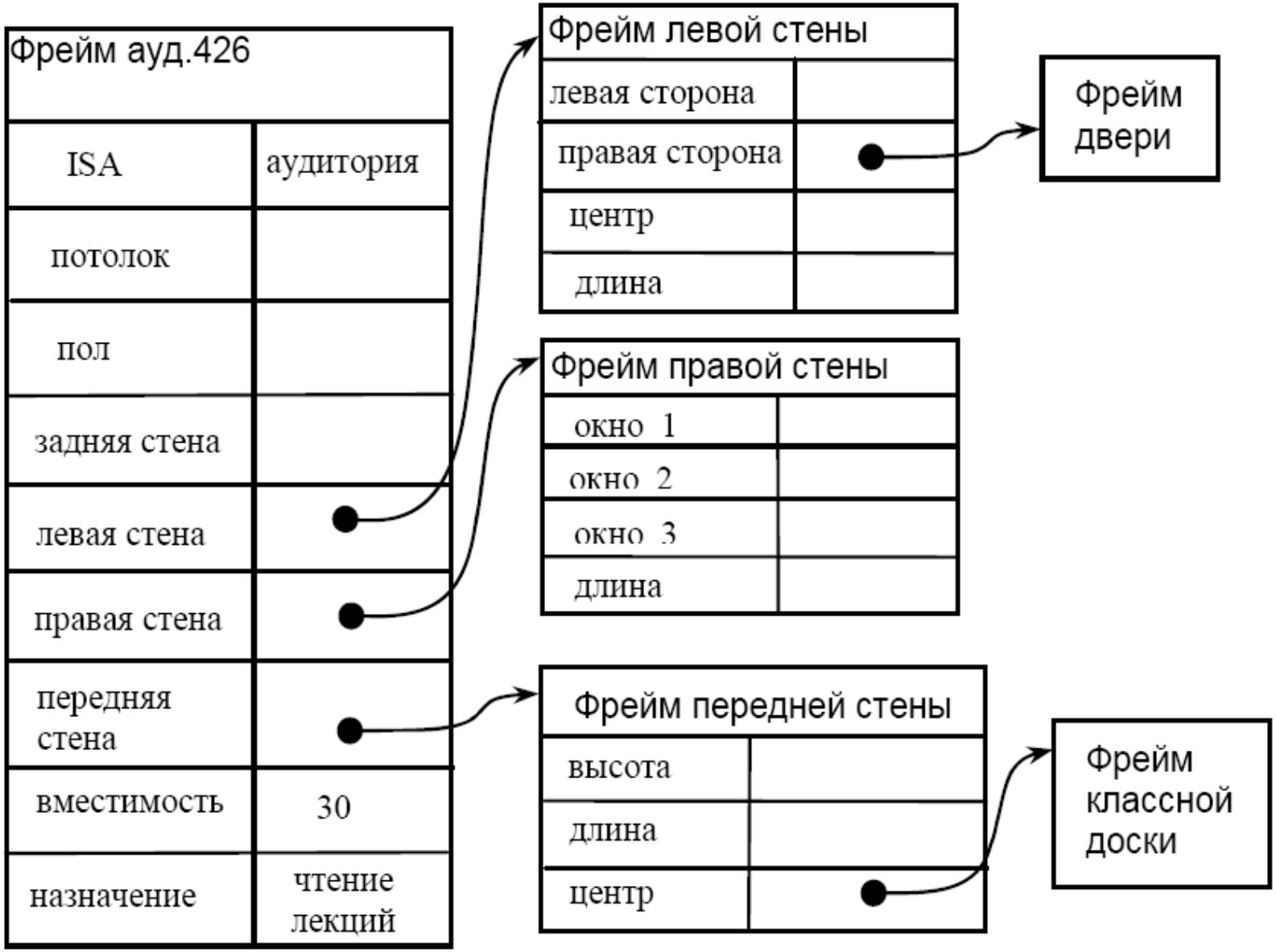
Семантическая сеть





Фреймы

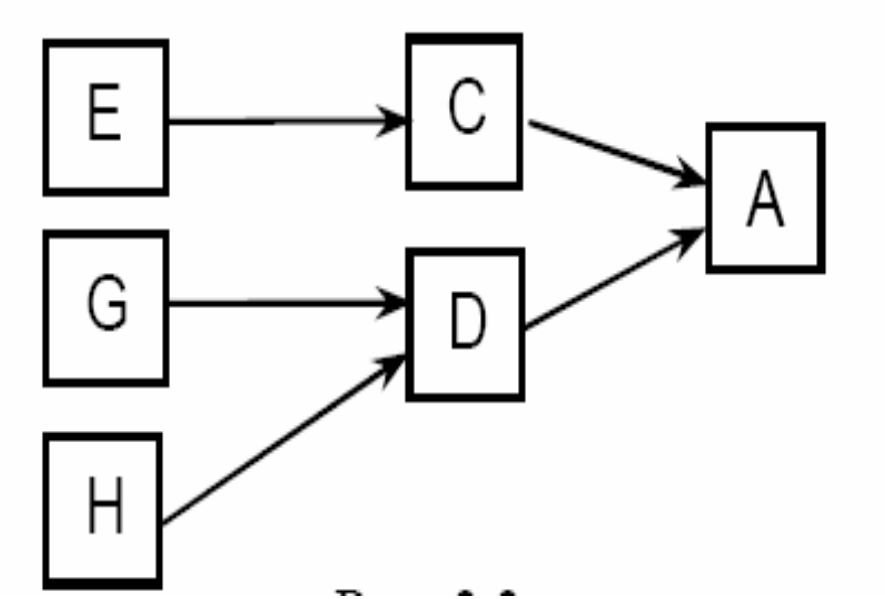

Конференция
Дата
Место проведения
Тема
Выступающие
IS-A
IS-A
Конференция по коммерческим вопросам
Конференция по развитию
Дата
Место проведения
Тема Торговля
Выступающие
Цель
Дата
Место проведения
Тема Развитие
Выступающие
Бюджет
IS-A
1 класс (шаблон)
Четвертая конференция по коммерческим вопросам
Дата 10.03.02 10:30
Место проведения Конференцзал
Тема
Выступающие Иванов
Цель 25 т. долл.
Экземпляр


Фрейм ауд.426
ISA
аудитория
потолок
пол
задняя стена
левая стена
правая стена
передняя стена
вместимость
30
назначение
чтение лекций
Фрейм левой стены
левая сторона
правая сторона
центр
длина
Фрейм двери
Фрейм правой стены
окно 1
окно 2
окно 3
длина
Фрейм передней стены
высота
длина
центр
Фрейм классной доски

Продукционная модель

Modus ponens

посылка p,
правило $p \rightarrow q$,
заключение q;



$C \wedge D \rightarrow A,\quad B \rightarrow A,\quad E \rightarrow C,\quad F \rightarrow C,\quad G \wedge H \rightarrow D,\quad J \wedge K \wedge L \rightarrow B.$

Modus tollens

Посылка ~q
Правило $p \rightarrow q$
Заключение ~p

ЛОГИЧЕСКИЕ МОДЕЛИ. Описания предметных областей, выполненные в логических языках, называются логическими моделями.
Языки логического типа опираются на исчисления, заимствованные из логики.
СЕТЕВЫЕ МОДЕЛИ. В основе моделей этого типа лежит конструкция, называемая семантической сетью.
ПРОДУКЦИОННЫЕ МОДЕЛИ. В моделях этого типа в прямой форме представляется информация о процедурах и условиях их применения.
<ситуации> → <заключение>.
<ситуации> → <действие>.
).
ФРЕЙМОВЫЕ МОДЕЛИ. Этим моделям присуща жесткая структура информационных единиц, которая называется протофреймом или фреймом-прототипом. В общем случае она выглядит следующим образом:
(имя_фрейма:
имя_слота_1 (значение_слота_1);
имя_слота_2 (значение_слота_2);
. . . . . .
имя_слота_N (значение_слота_N);
).
При конкретизации фрейма ему и слотам присваиваются конкретные имена и происходит заполнение слотов. Таким образом из фрейма-прототипа получаются фреймы- кземпляры.

- Онтология - это артефакт, структура, описывающая значения элементов некоторой системы.
- На формальном уровне онтология - это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно описывать классы, отношения, функции и индивиды.
- Онтология - это иерархически структурированное множество терминов, описывающих предметную область, которое может быть использовано как исходная структура для базы знаний.
- Основными компонентами онтологии могут являться:

- классы (или понятия),
- отношения (или свойства, атрибуты),
- функции,
- аксиомы,
- экземпляры (или индивиды).

Интеллектуальная информационная система (ИИС)

- комплекс программных, лингвистических и логико-математических средств для реализации основной задачи – осуществления поддержки деятельности человека и поиска информации в режиме продвинутого диалога на естественном языке. ИИС являются разновидностью интеллектуальной системы, а также одним из видов информационных систем.
- Вопросно-ответные системы:
- Например интеллектуальные поисковики – система старт: http://start.csail.mit.edu/index.php
- Виртуальные собеседники и помощники

Методы Машинного Обучения

- Обширный подраздел искусственного интеллекта, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа, и выделяющая знания из данных.

- Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении закономерностей в эмпирических данных. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

- Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации (Information Extraction), интеллектуальным анализом данных (Data mining).

Индуктивное обучение:

– Обучение с учителем:

* классификация
* восстановление регрессии
* структурное обучение (structured learning, нахождение структуры или зависимости имеющейся в исходных данных, применяется в компьютерном зрении)

– Обучение без учителя:

* кластеризация
* визуализация данных
* понижение размерности (необходимо представить данные в пространстве меньшей размерности, по возможности, минимизировав потери информации (метод главных компонент))

– Обучение с подкреплением (reinforcement learning)

– Активное обучение

– ...

Активное обучение — отличается тем, что обучаемый алгоритм имеет возможность самостоятельно назначать следующую исследуемую ситуацию, на которой станет известен верный ответ:

Обучение с частичным привлечением учителя (semi-supervised learning) — для части прецедентов задается пара «ситуация, требуемое решение», а для части — только «ситуация»

Трансдуктивное обучение (transduction) — обучение с частичным привлечением учителя, когда прогноз предполагается делать только для прецедентов из тестовой выборки

Многозадачное обучение (multi-task learning) — одновременное обучение группе взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение»

Многовариантное обучение (multiple-instance learning) — обучение, когда прецеденты могут быть объединены в группы, в каждой из которых для всех прецедентов имеется «ситуация», но только для одного из них (причем, неизвестно какого) имеется пара «ситуация, требуемое решение»

Общие постановки задач обучения по прецедентам

- Задано множество объектов X, множество допустимых ответов Y , и существует целевая функция (target function) $y^* : X \to Y$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \ldots, x_k\} \subset X$.

Пары «объект–ответ» (x_i, y_i) называются прецедентами.

- Совокупность пар

$X = (x_i, y_i)_{i=1..k}$ называется обучающей выборкой (training sample).

- Задача обучения по прецедентам заключается в том, чтобы по выборке X^k восстановить зависимость y^* , то есть построить решающую функцию (decision function)

$a : X \to Y$, которая приближала бы целевую функцию $y^*(x)$, причём не только на объектах обучающей выборки, но и на всём множестве X. Решающая функция a должна допускать эффективную компьютерную реализацию; по этой причине её называют алгоритмом??.

Объекты и признаки

- Признак (feature) f объекта x — то результат измерения некоторой характеристики объекта. Формально признаком называется отображение $f : X \to D_f$, где D_f — множество допустимых значений признака. В частности, любой алгоритм $a : X \to Y$ также можно рассматривать как признак.
- В зависимости от природы множества D_f признаки делятся на несколько типов.
- Если $D_f = \{0, 1\}$, то f — бинарный признак;
- Если $D_f = \{1, 2 .. s_j\}$ — конечное множество, то f — номинальный признак (категориальный или фактор); если j=2, то бинарный.
- Если D_f — конечное упорядоченное множество, то f — порядковый признак; Например, D_f= {Beginner, Elementary, Intermediate, Advanced, Proficiency} уровень владения английским языком)
- Если $D_f = R$ или $D_f = R^m$, то f — количественный признак.
- Если все признаки имеют одинаковый тип, $D_{f1} = \cdot \cdot \cdot = D_{fn}$, то исходные данные называются однородными, в противном случае — разнородными.

Пусть имеется набор признаков $f_1 .. f_n$. Вектор $\left(f_1(x),\ldots,f_n(x)\right)$ называют признаковым описанием объекта $x \in X$. Совокупность признаковых описаний всех объектов выборки X^k, записанную в виде матрицы $k \times n$, называют матрицей объектов признаков.

$$F = \left\| f_j(x_i) \right\|_{k \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_k) & \ldots & f_n(x_k) \end{pmatrix}$$

Матрица объектов–признаков является стандартным и наиболее распространённым способом представления исходных данных в прикладных задачах.

Ответы и типы задач

В зависимости от природы множества допустимых ответов Y задачи обучения по прецедентам делятся на следующие типы:

Если Y = {1, . . . , M}, то это задача классификации (classification) на M непересекающихся классов. В этом случае всё множество объектов *X* разбивается на классы $K_y = \{x \in X : y^*(x) = y\}$, и алгоритм a(x) должен давать ответ на вопрос «какому классу принадлежит x?». В некоторых приложениях классы называют образами и говорят о задаче распознавания образов (pattern recognition).

Если $Y = \{0, 1\}^M$, то это задача классификации на M пересекающихся классов. В простейшем случае эта задача сводится к решению M независимых задач классификации с двумя непересекающимися классами. (1 – принадлежит классу i, 0 - нет)

Если Y = R, то это задача восстановления регрессии (regression estimation).

Задачи прогнозирования (forecasting) являются частными случаями классификации или восстановления регрессии, когда $x \in X$ — описание прошлого поведения объекта x, $y \in Y$ — описание некоторых характеристик его будущего поведения.

Двухклассовая классификация. Наиболее простой в техническом отношении случай, который служит основой для решения более сложных задач.

Многоклассовая классификация. Когда число классов достигает многих тысяч (например, при распознавании иероглифов или слитной речи), задача классификации становится существенно более трудной.

Непересекающиеся классы.

Пересекающиеся классы. Объект может относиться одновременно к нескольким классам.

Нечёткие классы. Требуется определять степень принадлежности объекта каждому из классов, обычно это действительное число от 0 до 1.

Распознавание рукописных символов (цифр)

Научиться распознавать рукописный символ по его изображению.

Предполагается, что на картинке нарисована только одна цифра. При распознавании рукописного текста (сложная задача) вначале нужно выделить такое изображение. В некоторых случаях (распознавание индекса) — это уже сделано.

Изображения цифр от 0 до 9 закодированы известным образом. Требуется определить, какая цифра нарисована.

Например, код (признаковое описание) — битовая матрица размера 32 × 32. 1 — пиксел черный, 0 — пиксел белый.

Изображение перед кодированием масштабируется, чтобы все изображения имели примерно одинаковый размер.

Элементы матрицы запишем по строкам получим вектор x длины $32^2 = 1024$ — признаковое описание объекта.

$X = \{0, 1\}^{1024}$.

Множество всех возможных кодов разбивается на 10 классов:

$Y = \{0, 1, 2, \ldots, 9\}$

Получили задачу классификации: по $x \in X$ требуется определить класс $y \in Y$.

Входы $x_1, x_2, \ldots, x_d$ — бинарные признаки.

Обучение проходит на обучающей выборке (наборе «прецедентов») $(x^{(i)}, y^{(i)})$ $(i = 1, 2, \ldots, N)$.

Обучающая выборка в примере optdigit

http://www.ics.uci.edu/~mlearn/MLRepository.html содержит 1934 прецедента.

Пример. В задачах медицинской диагностики в роли объектов выступают пациенты. Признаки характеризуют результаты обследований, симптомы заболевания. и применявшиеся методы лечения. Примеры бинарных признаков — пол, наличие головной боли, слабости, тошноты, и т. д. Порядковый признак — тяжесть состояния (удовлетворительное, средней тяжести, тяжёлое, крайне тяжёлое). Количественные признаки — возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т. д. Признаковое описание пациента является, по сути дела, формализованной историей болезни. Накопив достаточное количество прецедентов, можно решать различные задачи: классифицировать вид заболевания (дифференциальная диагностика); определять наиболее целесообразный способ лечения; предсказывать длительность и исход заболевания; оценивать риск осложнений; находить синдромы — наиболее характерные для данного заболевания совокупности симптомов. Ценность такого рода систем в том, что они способны мгновенно анализировать и обобщать огромное количество прецедентов — возможность, недоступная человеку.

Узнавание образцов. Имеется некоторое количество картинок, на каждой из которых нарисована кошка (или треугольник, или жираф, или самолет, или конкретный человек) и, возможно, некоторое количество картинок, на каждой из которых она (он) отсутствует. Построить алгоритм, определяющий наличие кошки (треугольника и т.д.) на картинке. Распознавание голосовых команд. Имеется некоторое количество звуковых файлов, содержащих записи произнесения (utterance) голосовых команд из конечного ассортимента (например, слов “да” и “нет” или цифр для голосового набора номера) и знание, в каком файле какая команда. Построить алгоритм, понимающий голосовые команды. Распознавание речи. Имеется некоторое количество звуковых файлов, содержащих записи естественной речи на каком-то языке, и текстовых файлов — их расшифровок. Построить алгоритм, распознающий речь и записывающую ее в виде текста. Геологическая диагностика. Про некоторое количество разработанных нефтяных месторождений известны данные их предварительной геологической разведки (например, сейсмограммы) и результаты их эксплуатации. Построить алгоритм, предсказывающий эксплуатационные характеристики, в первую очередь мощность, разведанных, но еще не вскрытых месторождений.

Экономическое прогнозирование. Имеются данные о еженедельных объемах продаж нескольких тысяч видов товаров в нескольких сотнях магазинов за несколько лет. Построить алгоритм, предсказывающий спрос на ближайший месяц.

Выдача кредитов. Предсказание ухода клиентов.

Модель алгоритмов и метод обучения.

Опр. Моделью алгоритмов называется параметрическое семейство отображений

$A = \{g(x,\theta) | \theta \in \Theta\}$, где $g : X \times \Theta \to Y$ — некоторая фиксированная функция, Θ — множество допустимых значений параметра θ, называемое пространством параметров или пространством поиска (search space).

Пример. В задачах с n числовыми признаками $f_j : X \to R \quad j = 1,...,n$ широко используются линейные модели с вектором параметров $\theta = (\theta_1,...,\theta_n) \in \Theta = R^n$:

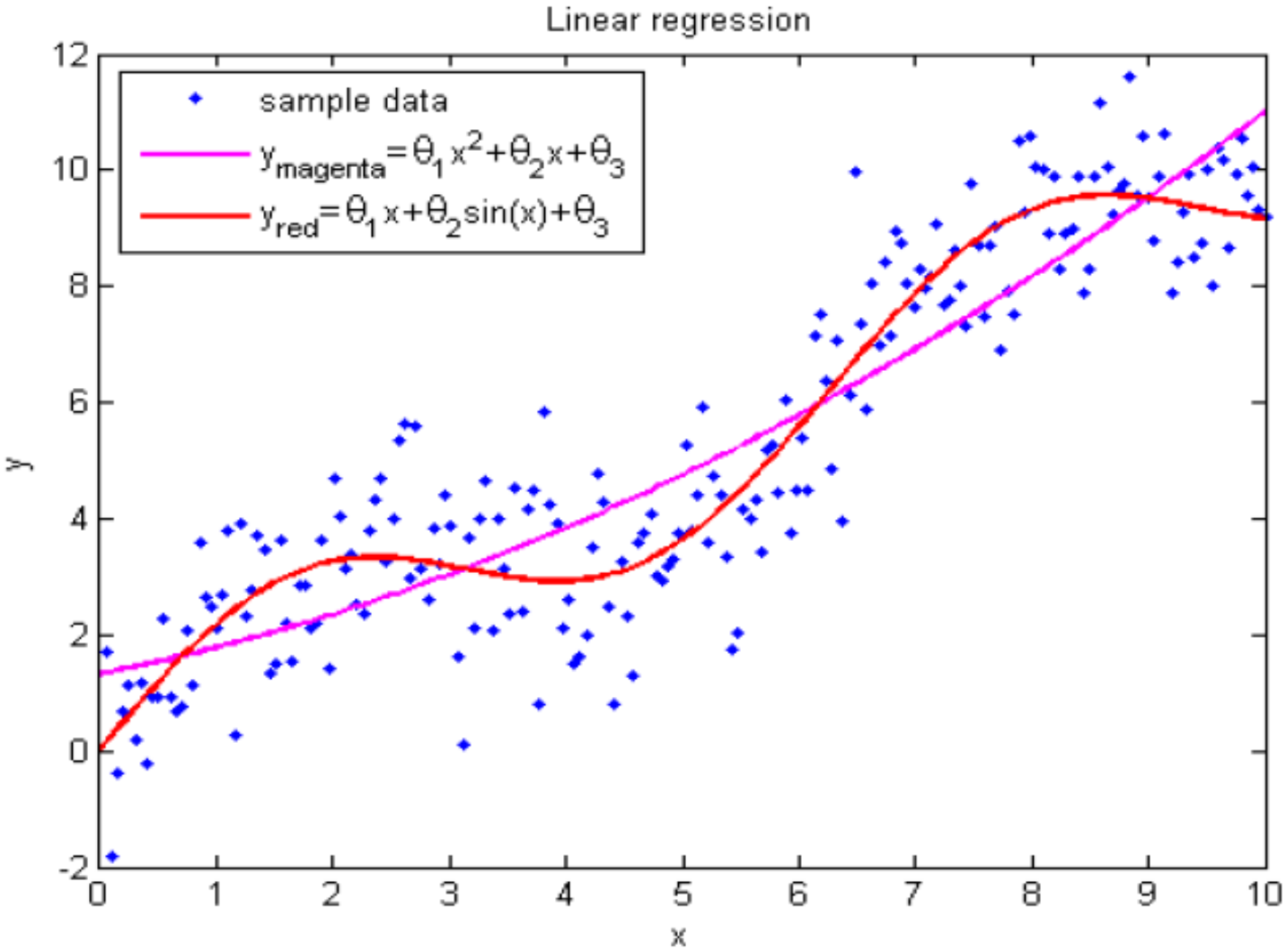
$$g(x,\theta) = \sum_{j=1}^{n} \theta_j f_j(x)$$ - для задач восстановления регрессии $Y \to R$

$$g(x,\theta) = sign \sum_{j=1}^{n} \theta_j f_j(x)$$ - для задач классификации $Y \to \{-1,+1\}$

Пример. Один из классических подходов к аппроксимации функций одной переменной по заданным точкам $(x_i, y_i) \in R^2, i = 1..k$, заключается в построении полиномиальной модели. Если ввести n признаков $f_j(x) = x^{j-1}$, то функция $g(x, \theta)$ из предыдущего примера будет определять полином степени n − 1 над исходным признаком x.

Процесс подбора оптимального параметра модели θ по обучающей выборке X^k называют настройкой (fitting) или обучением (training, learning) алгоритма $a \in A$.

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



Опр. Метод обучения (learning algorithm) — это отображение $\mu : (X \times Y)^k \to A$, которое произвольной конечной выборке $X^k = (x_i, y_i)_{i=1}^k$ ставит в соответствие некоторый алгоритм $a \in A$. Говорят также, что метод μ строит алгоритм a по выборке X . Метод обучения должен допускать эффективную программную реализацию.
Итак, в задачах обучения по прецедентам чётко различаются два этапа.

На этапе обучения метод *μ* по выборке X^k строит алгоритм $a = \mu(X^k)$.
На этапе применения алгоритм a для новых объектов x выдаёт ответы $y = a(x)$.

Этап обучения наиболее сложен. Как правило, он сводится к поиску параметров модели, доставляющих оптимальное значение заданному функционалу качества.

Функционал качества

Опр. Функция потерь (loss function) — это неотрицательная функция $Z(a,x)$, характеризующая величину ошибки алгоритма a на объекте x. Если $Z(a,x)=0$, то ответ a(x) называется корректным.

Опр. Функционал качества алгоритма a на выборке X^k:

$$Q(a,X^k)=\frac{1}{k}\sum_{i=1}^{k}Z(a,x_i)$$

Функционал Q также называют функционалом средних потерь или эмпирическим риском, так как он вычисляется по эмпирическим данным $(x_i,y_i)_{i=1}^{k}$

Функция потерь, принимающая только значения 0 и 1, называется бинарной. В этом случае $Z(a,x)=1$ означает, что алгоритм a допускает ошибку на объекте x, а функционал Q называется частотой ошибок алгоритма a на выборке X^k.

Наиболее часто используются следующие функции потерь, при Y $\subseteq$ R:

$Z(a,x) = \left[a(x) \neq y^*(x)\right]$ - индикатор ошибки, применяющийся в задачах классификации, квадратные скобочки 0 – если условие не выполняется, 1 – выполняется.

$Z(a,x) = \left|a(x) - y^*(x)\right|$ - отклонение от правильного ответа, функционал Q называется средней ошибкой алгоритма a на выборке X^k.

$Z(a,x) = \left(a(x) - y^*(x)\right)^2$ - отклонение от правильного ответа, функционал Q называется средней квадратичной ошибкой алгоритма a на выборке X^k, применяется обычно в задачах регрессии.

Классический метод обучения, называемый минимизацией эмпирического риска (empirical risk minimization, ERM), заключается в том, чтобы найти в заданной модели A алгоритм a, доставляющий минимальное значение функционалу качества Q на заданной обучающей выборке X^k.

$$\mu(X^k) = \arg\min_{a \in A} \left(Q\left(a, X^k\right) \right)$$

Пример. В задаче восстановления регрессии (Y = R) с n числовыми признаками $f_j : X \to R$, $j = 1,...,n$, и квадратичной функцией потерь метод минимизации эмпирического риска есть ничто иное, как метод наименьших квадратов:

$$\mu(X^k) = \arg\min_{\theta} \left(\sum_{i=1}^{k} \left(g\left(x_i, \theta\right) - y_i \right)^2 \right)$$

Вероятностная постановка задачи обучения

В задачах обучения по прецедентам элементы множества X — это не реальные объекты, а лишь доступные данные о них. Данные могут быть неточными, поскольку измерения значений признаков $f_j(x)$ и целевой зависимости $y^*(x)$ обычно выполняются с погрешностями. Данные могут быть неполными, поскольку измеряются не все мыслимые признаки, а лишь физически доступные для измерения.

В результате одному и тому же описанию x могут соответствовать различные объекты и различные ответы. В таком случае $y^*(x)$, строго говоря, не является функцией. Устранить эту некорректность позволяет вероятностная постановка задачи.

Вместо существования неизвестной целевой зависимости $y^*(x)$ предположим существование неизвестного вероятностного распределения на множестве $X \times Y$ с плотностью $p(x, y)$, из которого случайно и независимо выбираются k наблюдений $\left(x_i, y_i\right)_{i=1}^{k}$.

Такие выборки называются простыми или случайными одинаково распределёнными (independent identically distributed, i.i.d.).

Вероятностная постановка задачи считается более общей, так как функциональную зависимость $y^*(x)$ можно представить в виде вероятностного распределения $p(x, y) = p(x)p(y \mid x)$, положив $p(y \mid x) = \delta\left(y - y^*(x)\right)$, где δ - дельта функция.

Принцип максимума правдоподобия.

При вероятностной постановке задачи вместо модели алгоритмов $g(x,\theta)$, аппроксимирующей неизвестную зависимость $y^*(x)$, задаётся модель совместной плотности распределения объектов и ответов $\varphi(x,y,\theta)$, аппроксимирующая неизвестную плотность $p(x,y)$. Затем определяется значение параметра θ, при котором выборка данных X^k максимально правдоподобна, то есть наилучшим образом согласуется с моделью плотности.

Если наблюдения в выборке X^k независимы, то совместная плотность распределения всех наблюдений равна произведению $p(x,y)$ плотностей в каждом наблюдении.

$$p\left(X^k\right) = p\left((x_1, y_1), \ldots, (x_k, y_k)\right) = p(x_1, y_1) \cdot \ldots \cdot p(x_k, y_k)$$

Заменив, $p(x, y)$ на модель плотности $\varphi(x, y, \theta)$, получим функцию , правдоподобия: $L\left(\theta, X^k\right) = \prod_{i=1}^{k} \varphi(x_i, y_i, \theta)$

Чем выше значение правдоподобия, тем лучше выборка согласуется с моделью. Значит, нужно искать значение параметра θ, при котором значение $L\left(\theta, X^k\right)$ максимально. После того, как значение параметра θ найдено, искомый алгоритм $a_\theta(x)$ строится по плотности φ(x, y, θ).

Связь максимизации правдоподобия с минимизацией эмпирического риска.

Вместо максимизации $L\left(\theta, X^k\right)$ удобнее минимизировать функционал $-\ln L\left(\theta, X^k\right)$, поскольку он аддитивен по объектам выборки. Этот функционал совпадает с функционалом эмпирического риска Q, если определить вероятностную функцию потерь $Z\left(a_\theta, x\right) = -k \ln \varphi\left(x, y, \theta\right)$, ведь чем хуже пара $\left(x_i, y_i\right)$ согласуется с моделью φ, тем меньше значение плотности $\varphi(x_i,y_i,\theta)$ и выше величина потери $Z\left(a_\theta, x\right)$. Верно и обратное — для многих функций потерь возможно подобрать модель плотности $\varphi(x,y,\theta)$ таким образом, чтобы минимизация эмпирического риска была эквивалентна максимизации правдоподобия.

Проблема переобучения и понятие обобщающей способности

Минимизацию эмпирического риска следует применять с известной долей осторожности. Если минимум функционала $Q\left(a, X^k\right)$ достигается на алгоритме a, то это ещё не гарантирует, что a будет хорошо приближать целевую зависимость на произвольной контрольной выборке $X^l = \left(x'_i, y'_i\right)_{i=1}^l$.

Когда качество работы алгоритма на новых объектах, не вошедших в состав обучения, оказывается существенно хуже, чем на обучающей выборке, говорят об эффекте переобучения (overtraining) или переподгонки (overfitting). При решении практических задач с этим явлением приходится сталкиваться очень часто.

Легко представить себе метод, который минимизирует эмпирический риск до нуля, но при этом абсолютно не способен обучаться. Получив обучающую выборку X^k, он запоминает её и строит алгоритм, который сравнивает предъявляемый объект x с обучающими объектами x_i из X^k (поиск в БД). В случае совпадения $x = x_i$ алгоритм выдаёт правильный ответ y_i. Иначе выдаётся произвольный ответ. Эмпирический риск принимает наименьшее возможное значение, равное нулю. Однако этот алгоритм не способен восстановить зависимость вне материала обучения. Отсюда вывод: для успешного обучения необходимо не только запоминать, но и обобщать.

Обобщающая способность (generalization ability) метода μ характеризуется величиной $Q\left(\mu\left(X^k\right), X^l\right)$ при условии, что выборки X^k и X^l являются представительными. Для формализации понятия «представительная выборка» обычно принимается стандартное предположение, что выборки X^k и X^l — простые, полученные из одного и того же неизвестного вероятностного распределения на множестве X.

Опр. Метод обучения μ называется состоятельным, если при заданных достаточно малых значениях ε и η справедливо неравенство:

$P_{X^k, X^l}\left\{Q\left(\mu\left(X^k\right), X^l\right) > \varepsilon\right\} < \eta$, ε – *точность,* (1-η) – *надежность.*

Эмпирические оценки обобщающей способности применяются в тех случаях, когда не удаётся воспользоваться теоретическими.

Пусть дана выборка $X^K = (x_i, y_i)_{i=1}^K$. Разобьём её N различными способами на две непересекающиеся подвыборки — обучающую X^k длины k и контрольную X^l длины l = K− k. Для каждого разбиения n = 1, . . . , N построим алгоритм $a_n = \mu(X_n^k)$ и вычислим значение $Q_n = Q(a_n, X_n^l)$.

Среднее арифметическое значений Q_n по всем разбиениям называется оценкой скользящего контроля (cross-validation, CV):

$$CV(\mu, X^K) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^k), X_n^l),$$

в простейшем случае разбиения генерируются случайным образом, N берется от 20 до 100.

Стандартом «де факто» считается методика t×q-кратного скользящего контроля (t×q-fold cross-validation), когда выборка случайным образом разбивается на q блоков равной (или почти равной) длины, каждый блок по очереди становится контрольной выборкой, а объединение всех остальных блоков — обучающей. Выборка X^K по-разному t раз разбивается на q блоков. Итого получается $N = t \times q$ разбиений. Данная методика даёт более точные оценки за счёт того, что все объекты ровно по t раз встречаются в контроле. Недостатками скользящего контроля являются: вычислительная неэффективность, высокая дисперсия, неполное использование имеющихся данных для обучения из-за сокращения длины обучающей выборки с K до k.