Большие данные 14 марта 2020 г

См тетрадочку!

Правило 3-трёх VVV

1. volume( обем)
2. veloсity (скорость обработки данных)
3. validation структурируемость

Ai ML

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ai | ML | DL | NN | Data Mining |
| Искусственный интеллект | Машинное обучение | Глубокое обучение | Нейронная сеть |  |

Вероятность p=(благоприятных ко всем возможным) (m/n)

Событие а и b

P(a) and P(b)

P(a)\*p(b)=1/2\*1/2=1/4

Выбрасывают два кубика(6 цифр) и надо чтобы выпало s меньше равно 3

N=6^2;

M=2

P=1/18

Искусственный интеллект решает задачи:

Направлен на решение задач или имитация работы человека для решение **неформализованных** или **труднодноформализованных** задач;

Локальный ИИ узкоспециализированный

Глобальный ИИ этот, как человек

**Анализ данных и машинное обучение.**

Анализ данных и машинное обучение - это наука, изучающая способы извлечения закономерностей из ограниченного кол-во примеров. Однако стоит отметить, что термины анализ данных и машинное обучение не тождественны. МО занимается строгим изучением методов извлечения закономерностей, а АД это направление решающее прикладные задач.

Подходы в АД:

1. собиранием data set и решаем прикладные задачу
2. не собираем dataset и берём готовую и исследуем на устойчивость алгоритма машинного обучения
3. создание синтетических dataset( выявляем закономерности и сделать новые записи на основе закономерности, чтобы увеличить dataSet)

Задачи в МО

прогноз-это набор методов, алгоритмов, статических методов- можем прогнозировать число (курс доллара)

1. обучение без учителя(кластеризация). задача объединения объекта на группы обладающая некоторыми свой-ми. Для первичной обработки, чтобы посмотреть какие у них есть связи.
2. Оценивание плотности. Задачи приближения и распределения объектов, задача выявления аномалий. В таких задачах сначала оценивается сначала распределения нормальных объектов, а затем аномальными является всё объекты, в рамках которых это распределения получает слишком низкую вероятность
3. визуализация –задача изображения многомерных объектов над двух или трёх мерной плоскостью, таким образом, чтобы сохранялась как можно больше зависимостей и отношений между ними
4. понижение размерности- генерация таких новых признаков, что их меньше чем исходных, но при этом задача решается не хуже исходной.
5. Задачи регрессии(линейной, логистической, полиномиальной)
6. Ранжирование
7. Обучение с учителем (классификация-это разбиение на классы)

7.1) бинарная классификация y={0,1}

7.2) многоклассовая классификая без пересечения y={k1,k1,…kL}

7.3) многоклассовая классификая с пересечением y={0,1}^L

7.4) полуконтролируемая или частично обучение. В задачах, в которых для одной части объектов известны и всё признаки, и ответы, а для другой части объектов только признаки.

Мы берём объект, потом выделяем признаки. После этого мы набрали множество Х объектов. Любой объект описывает признаки(атрибуты)-причём признаки это вектор, соответственно для каждого объекта есть какие-то свои признаки.

Рисунок 1.

Мы подаём dataset и попробуй распределить их как-то по табличке.

Рисунок 2

Сколько An атрибутов, столько и признаков.

Используемые языки для МО python и R. В питоне есть scikit-learn для МО.

Есть 3 метода для кластеризации:

1) DBSCAN

2) K-means

3) Agglomerative

В каждом из этих методов уже заложено евклидово пространство. И **минус** что для атрибута у нас просчитывается.

**Для вытянутых подходит манхэттенское.**

**Минус** ещё один, заключается в том, что 2 и 3 метод требуют на вход кол-во кластеров на которые надо разбить. Для этого есть чуть получше 1, так как на прямую мы не задаём кол-во кластеров, мы задаём 2 параметра. Eps, min-semple;

Визуализация, мы выбираем 2 атрибута и мы должны выбрать так, чтобы у нас

Понижение размеренности - это, например, из 100 мерного сделать 50 мерное.

Задачи МО-Нахождение целевой.

Рисунок 3

Мы должны аппроксимировать значения, y=ax+b

Наша задача простроить целевую ф-и которая отображает результат.

Мы должны предъобработать данные(выкинуть шумы, которые мешают из-за погрешностей).

Дальше переходим к анализу и выбору алгоритму.

Рисунок4

Задача классификации у нас есть множество объектов х и у(ответы на х).

Мы делим на обучающая и на тестовую выборку.

Наша задача будет заключатся в восстановлении х по y. Надо разбить dataset на 2 части.

70 на 30 или 80 на 20. Всё зависит от особенностей. В первом приближении разобьём 30 на 70, а потом смотрят на сколько хорошо, но это подгоняется мб 65 на 35;

Как нам разбить? это табличка можно разбить 80 отсчитать, остальное в тестовую.

Можно попробовать перемешать и заново разбить.

Можно взять 3-тью каждую третью строчку, тоже плохой вариант.

Можно взять пересекающейся рис 5, но там переобучение.

Может возникнуть недообучение. Это лучше чем переобучение это можем обнаружить с помощью метриков.

**Дальше нужно выбрать какой алгоритм выбрать**. Как определить какой алгоритм решает нашу задачу, что нам нужно ввести- функционал качества работы алгоритма(или ф-ю потерь)

TP true positive

TN true negative

FP false positive

FN false negative

Можно ввести ф-ю ошибки, и она называется logLoss

L(a,y) чем меньше разница в каждом кластере, тем лучше, и вот эту ф-и будем минимизировать.

Основные понятие статистики

Мат. Ожидание –(это (КЕШ), т.е. те значения, которой с высокой вероятностью ожидаем.

Дисперсия –это отклонения, которое мы предсказали.

Дисперсия связана со среднем квадратичным отклонением

Тогда в общем виде запишется L(a,y)=

W-это вес или коэ-т нашего объекта, тогда w0 это сдвиг.

Нам помогает алгоритм DTW алгоритм динамической трансформации временной шкалы.

Упростим:

И тогда получим

Чем плохо MSE?

Возьмём корень от этого L(a(x),y)=

И тогда получим L(a(x),y)=(RMSE)

Что нам позволяет определить MSE и RMSE? Позволяет нам сравнить 2 модели, она позволяет контролировать наше обучение. Мы начинаем обучать модель и mse и в тот момент, когда она останавливает на интерации или становится меньше, то мы берём предыдущею итерацию и используем её.

Чем плоха эта штука, у нас могут быть ошибки 2-х родов.

Введя коэ-т детерминации у нас получится отношение: см рисунок 7

MAE

У MSE были + он не дифференцируем, от шумов он менее чувствительнее.

Mse

+ прост, дифференцируемый

- чувствителен на шумы, даёт другую размерность(квадратичную)

RMSE

+

-

MAE

Прост, однако менее чувствителен к шумам

-однако не дифференцируемый

Автор себоятья рашка.

16 05 2020

Есть 3 вида Машиного обучения Обучение с учителем, без учителя и обучение с потреблением

1 включает в себя: регрессию-имеет обычно вещественный аргумент

Классификация: разбиение и соотношение классов

Зависит от количества классов и соответственно разбивается на задачи

Самый простой вид классификации

С учителем:

1 бинарная классификация- самая простая (2 класса то что нам нужно и то что нам не относится)

2 много классовая классификация

3 много классовая без пересечений и много классовая с пересечениями

4 полу обучаемая кассификация – на всей части выборки есть и признаки и ответы, а на другой есть признаки, но нет ответов (из –за не достаточности данных, чаще всего встречается в медицинских систем)

Без учителя:

1. задачи кластеризация- выявить закономерности, а главная задача разбить на схожие группы.

1 задача визуализации и 2 уменьшении размерности

Работа над анализом данных начинается с постановки задачи, после этого начинаем собирать данные (выделить объект исследования, после этого нам нужен ещё признак объект мы должны выписать признаки, которые будут описывать объект)

У нас нет ограничения на количество признаков, но мы должны выписать всё возможные признаки, например, 100 признаков- хорошо это или плохо?!- чем больше признаков, тем сложнее работать с ними. Мы должны выбрать какой-то пул алгоритмов, чтобы работать с нашим dataset. Если у нас будет 100 признаков, то размерность у нас будет 100 мерное пространство. В этот момент начинается борьба чтобы уменьшить размерность

1 приближение, посмотрим сколько даёт каждый признак (уменьшение размерность -информационные выигрыши)

2 Придумать такие новые признаки, которые заменяют несколько старых признака, следовательно, в 3 раза мы бы могли бы уменьшить размерность. При этом они должны умешать качество на не значительное эпселент, которое мы могли бы контролировать.

После этого мы записали эти новые 66 признаков и тут вопрос на сколько линейно будет выводить результат

Что значит случайно отранжировать признак? Например, мы убрали признак, который мало влиял, но

4 задача –это задача визуализации отображение на 2-х или 3 мерной пространстве всех наших признаков, причём отобразить так, чтобы можно было увидеть задача визуализации.

Данные высокой размерности

Это данные big data и высокой размерности-это синонимы. Big data мб большое кол-во признаков

1 признак эффект существования многомерности

2 нам нужно объяснить природу того или иного действия, и природу многомерной анализированой структур

Многомерная структура-это множество статистических обследования объектов

[Q1,q1….qu] представляется ввиде таблицы, либо вектора

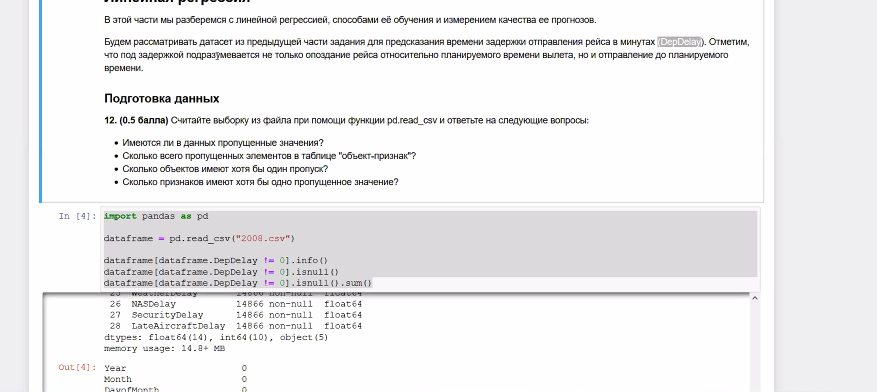
X=(x1,x2….xn)

Матрица A= (a11…..a1n

an1……ann)

4 задача или подход снижения размерности

Дискретная велечина

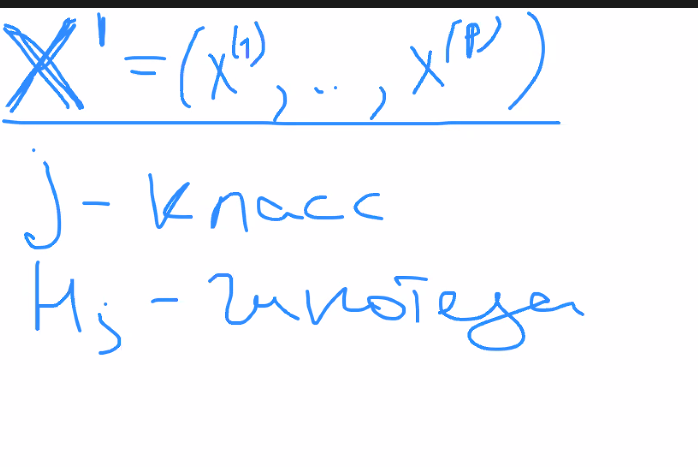


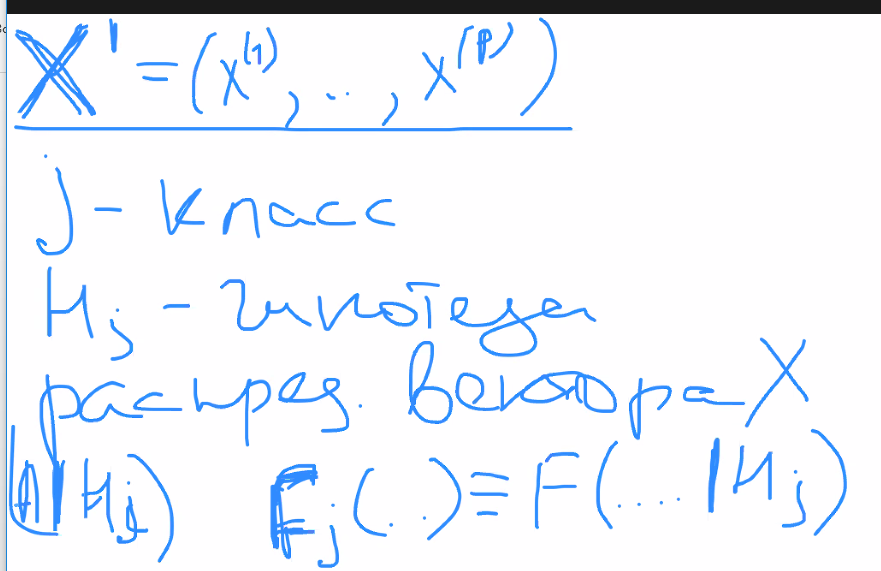


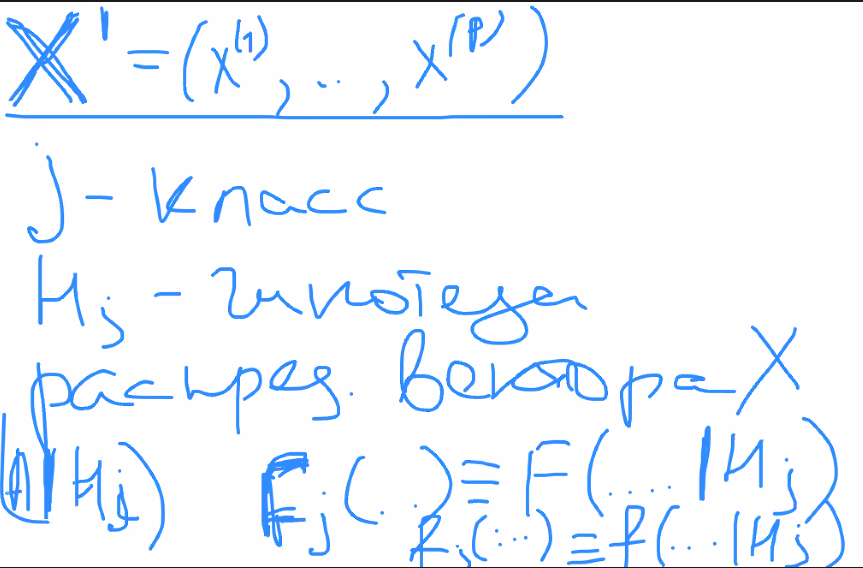
31 05 2020

Data set откуда он берётся.

Упорядочный набор –это упорядочный массив, мы распределяем по времени



Х- это наблюдение представляется ввиде вектора. Распределение вектора х, который принадлежности к этуму классу, то оно будет обозначатся ф-ей распределения F это само распределение, а f это плотность распределения 

Тогда полотность распреледеления 

Методы задания ф-и:

Графически

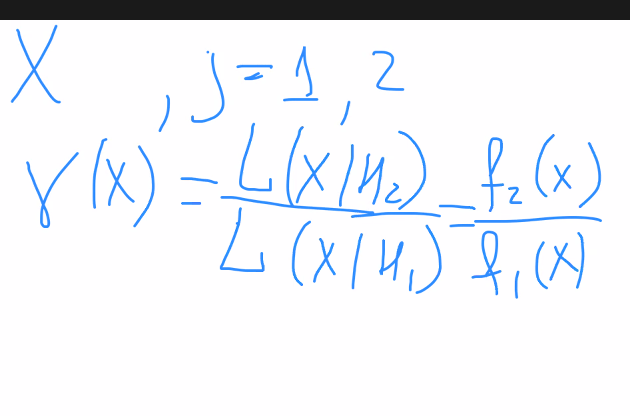
Аналитически

Функтации – отклонения(шумы) мы берём настолько большую выборку, что мы их не замечаем.

Задачи наблюдения Х:

Гипотеза h1 и есть гипотеза h2 которая принадлежит F2, и нам надо определить, кому она принадлежит

С паомощью критерия правдоподобия мы можем проверить j(x)=L(x/h2) / L(x/h1) = f2(x) /f1(x)



Ошибки 1-го рода

Ошибки 2-го рода

Пропускаются тогда, когда не должны(т.е. у нас алгоритм отвергает гипотезу. Отвергается правальный-это ошибка 1-го рода) А когда принимается не правильная гипотеза-это 2 рода.

Как мы понимаем, что есть ошибки 1 и 2 рода?-мы заранее подготовили данные и тестем их на алгоритме.

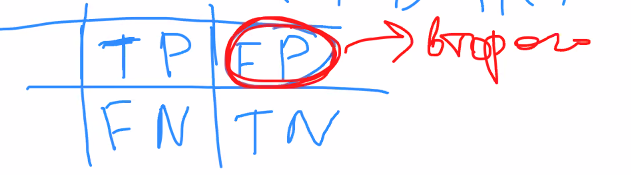
Таблица(матрица) ошибок:

tp true positeve

fn-false negative

Fp –false positeve

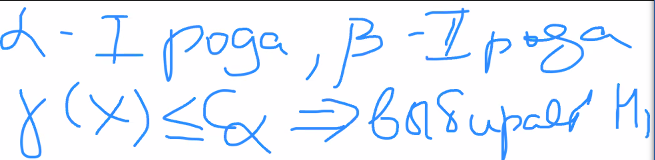
tn=true negative



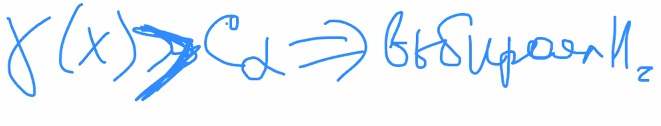
А что такое L ? в на один рис выше. L- это функция правдоподобия

А 1 рода ошибки b-второго ошибки

Выбирается h1

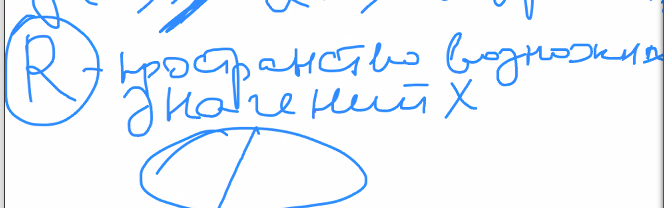


Выбирается h2

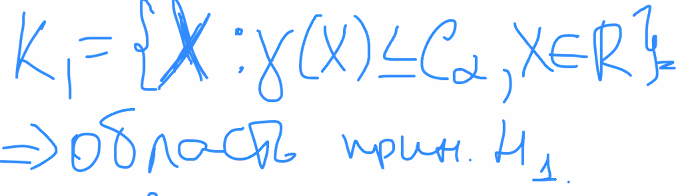


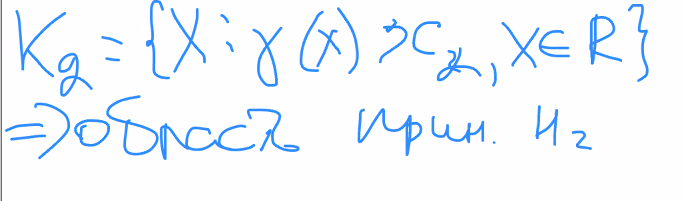
r- это пространство возможный значений Х

и тогда множество r разбивается на 2 не пересекающиеся множества



1 множество: когда гамма меньше с альфа (это критерий правдоподобности(ошибки первого рода)(это оценка ошибок 1-го рода, это установленное пороговое значение для первого случая)), тут R это пространство возможный значений



Примыкает h1

Тут важно, что к1 не пересекает к2

По другому эта штука называется.Это критическая область для принятия h1

Когда есть гипотеза, то у нас есть оплеорная верорятность. Она равняется



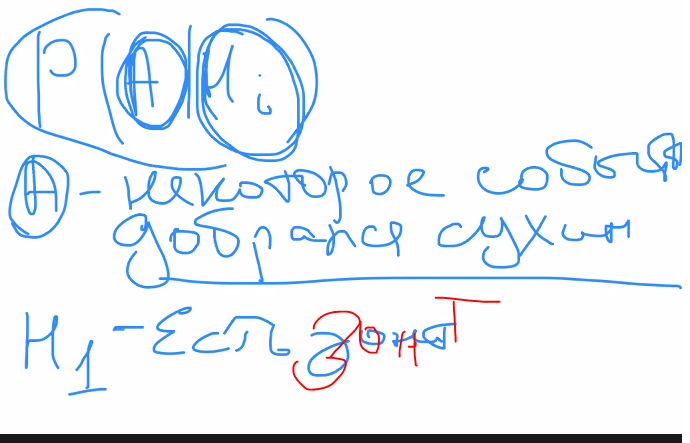
Благодаря ей, мы можем записать правило классификации:

------отступление---------------

Формула полной вероятности:



Что это за вероятность А-это некоторое событие. Это вероятность наступления А при условии H i.

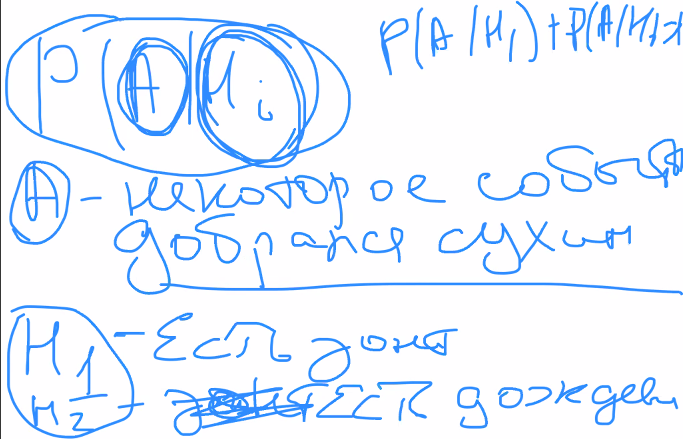


Добрался сухим на работу, есть зонт

Есть дождевик.

Мы отдельно считаем, отдельно вероятности h1 и h2 –это полная группа событий

P(a/h1)+p(a/h2)=1 тогда добрался сухим



P(a/hi) –это вероятность того что я добирался до работы A и h это не числа. Она может быть число(например при бросании костей) у нас произошло событие а при том, что наступило событие h, между ними не деление! Эта формула оприорной вероятности

Чему равняется формула p(A)?

Умножить на вероятность того или иного вроятность p(hj)

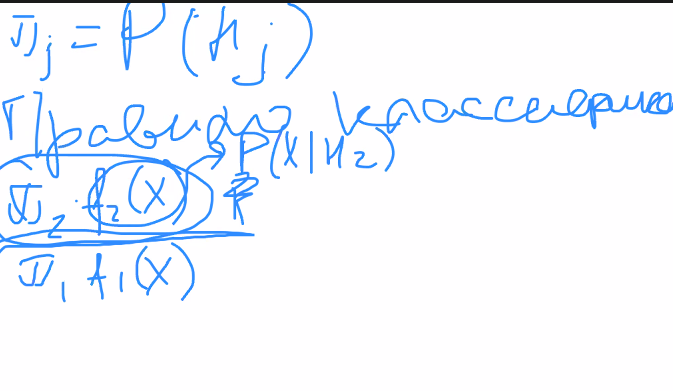


--------------------------------------------------------------------------------------------------------------------------------

Оприорная вероятность:

Пj=P(hj)

Правило классификатора

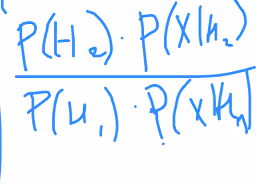


Отношения мб больше 1, следовательно можно сравнивать 1. Если это равно 1, то это плохо, так как мы можем отнести либо к 1, либо ко второму

Если больше 1 то относим к h2, а меньше к h1

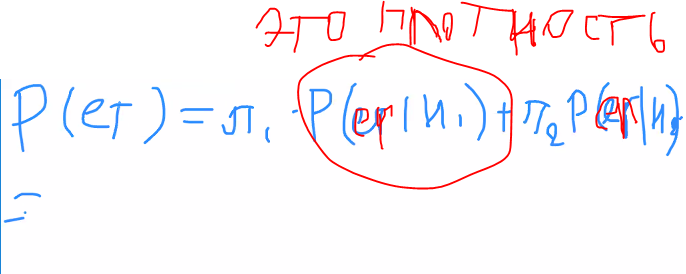


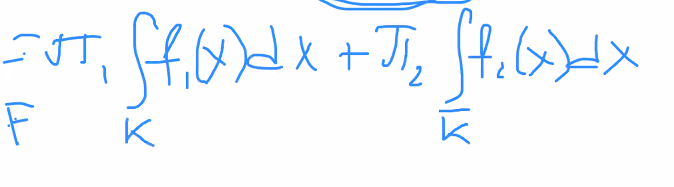
Байес уевский классификатор. Перепишем формулу. Он является частым случаем критерия правдоподобия(он принимим в жизни)

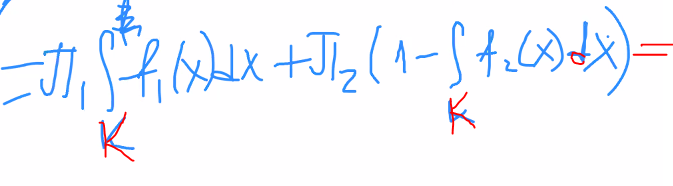


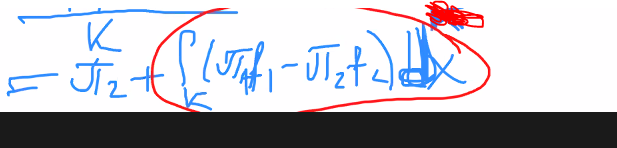
Где применятся.На почту приходит письма, как попадают в эту папку(как попадают в спам?) как раз прородителем был Байесвский прородитель. Он мало производительный

Мы хотим найти вероятность p(er) ошибки

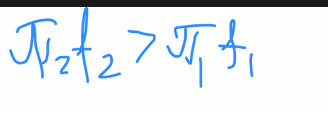




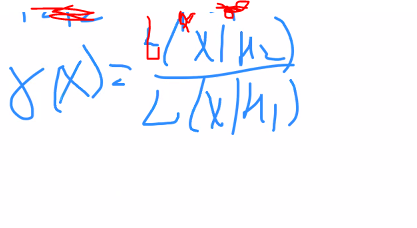


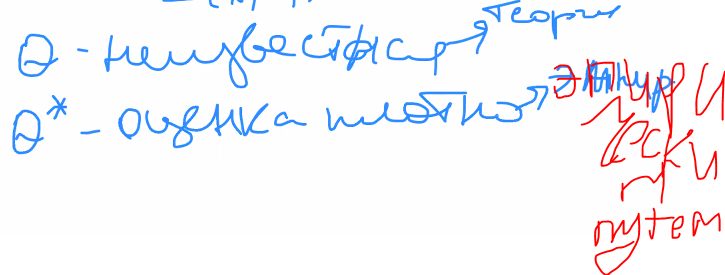


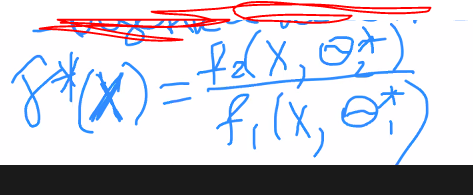
Тогда записываем следующий критерий



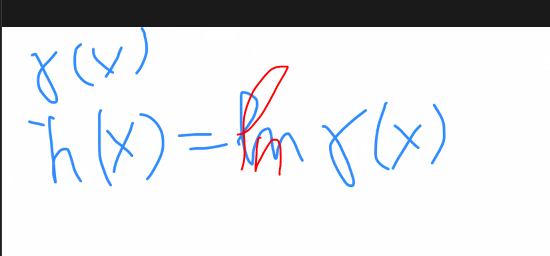
Это нам напоминает- Байесвский классификатор



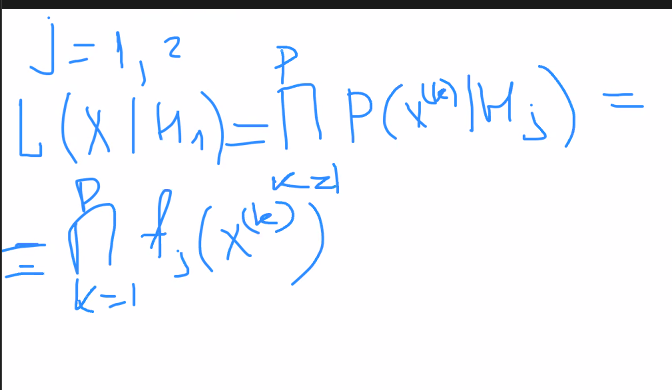




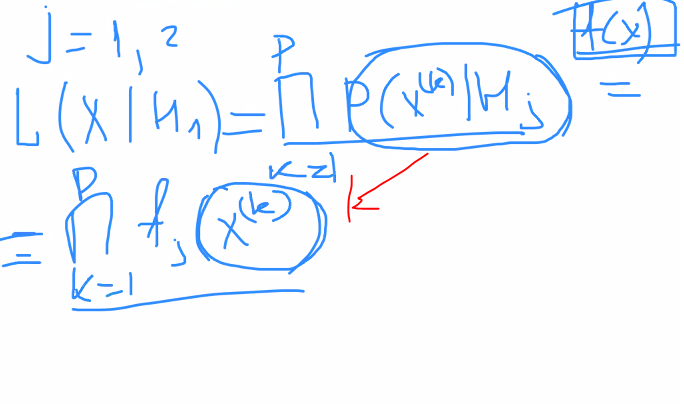
Ln



Для 2 случаев

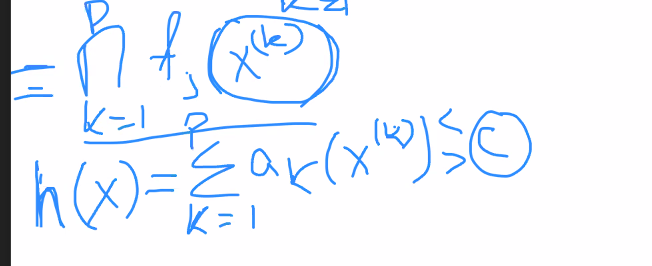


F(x) если она не прерывна, то она будет принимать значение

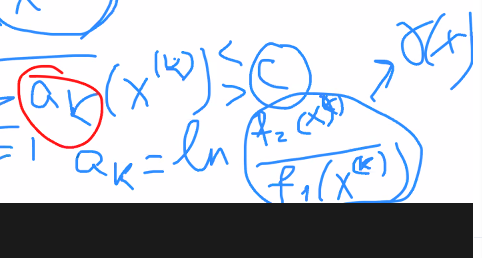


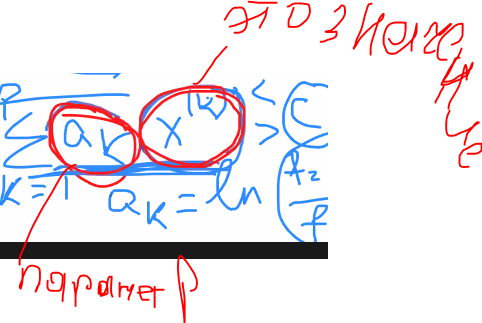
По мимо критерия правдоподобия нам не обходимо:

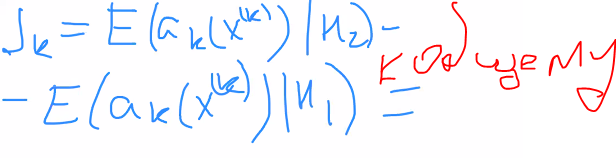
У нас же есть какой-то набор значений, который мы отнесём этот набор значений к классу h1. Нам надо ввести область попадания кипотез: если попало туда, то отнесём туда и сюда.

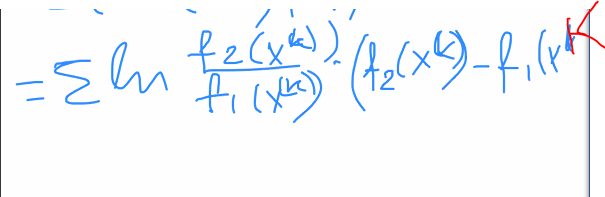


Ак описывает логарифмический смысл ак и оно равняется:

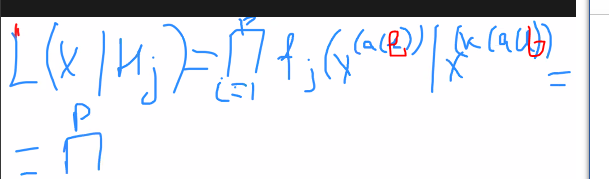




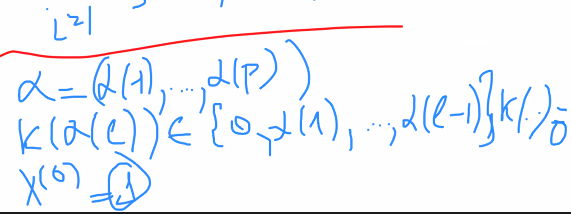


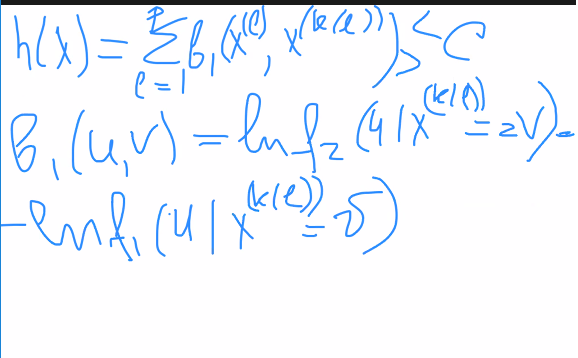


Следующая модель для 2-х дискретных распределений (деревообразная стрруктура):



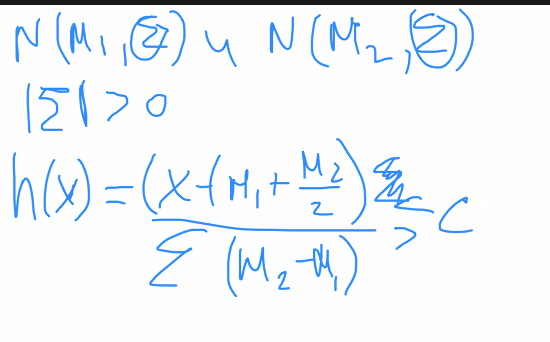
И можем переписать следующее,ф –ю правдопадобию

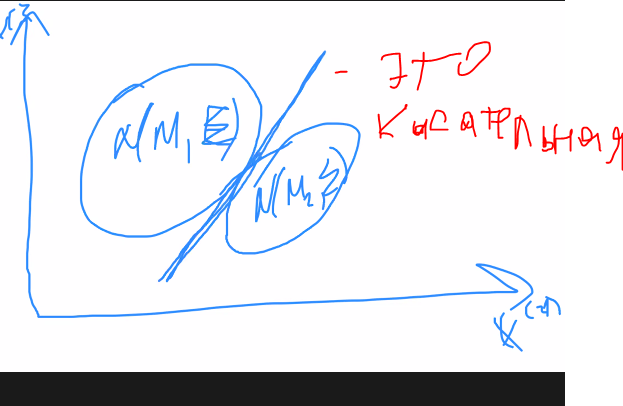
 



От дискретных можно перейти к прямым распределениям

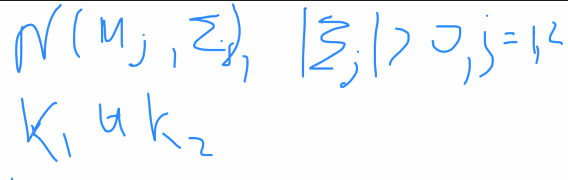
Особеность модели фишера является- простота



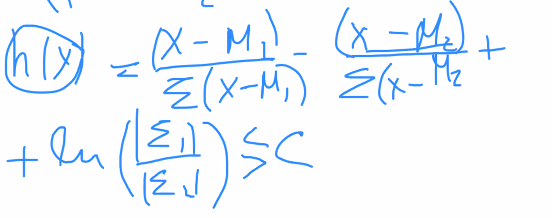


Следующая модель: с разными кореляционными матрицами и распределение у нас будет следующие:

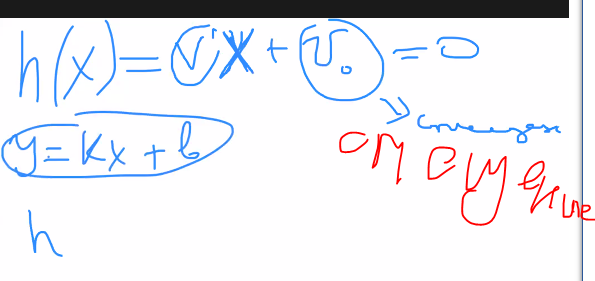
К1 и к2 области

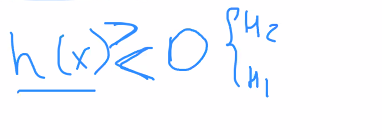


Получается некоторый полином 2-го порядка



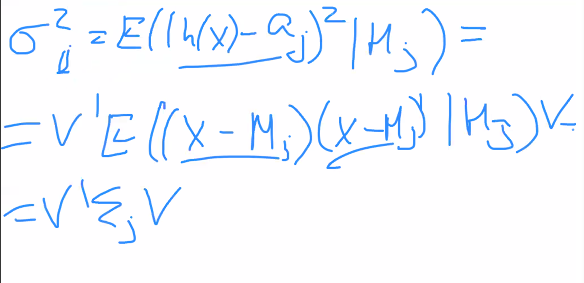
Как будет выглядеть гиперплоскость: должна быть похожа на y=kx +b

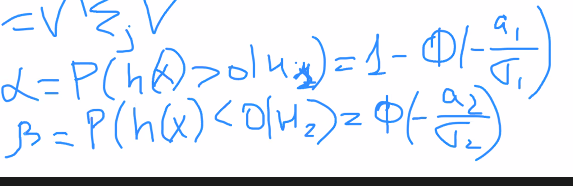
Vо смещение 

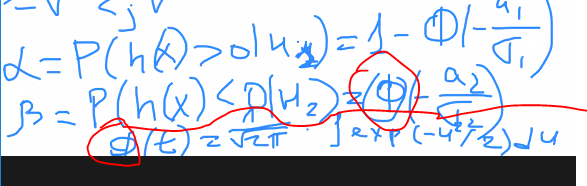


Какая ошибка будет минимизироваться?-

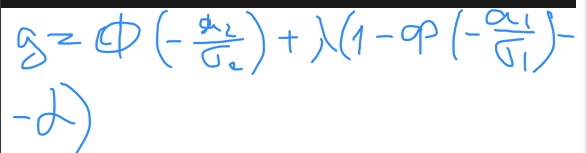
Среднее квадратическое отклонение

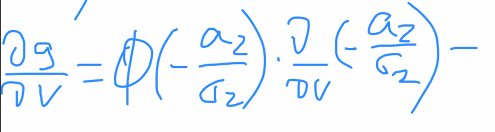


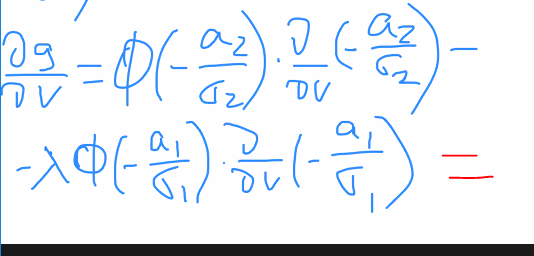




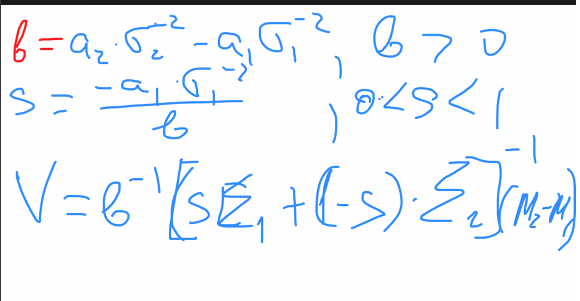
Для нахождения используется метод лангранжа



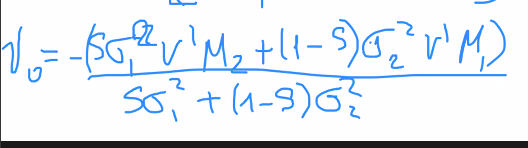


Предпологаем, что ошибки 1 и 2 рода меньше 50

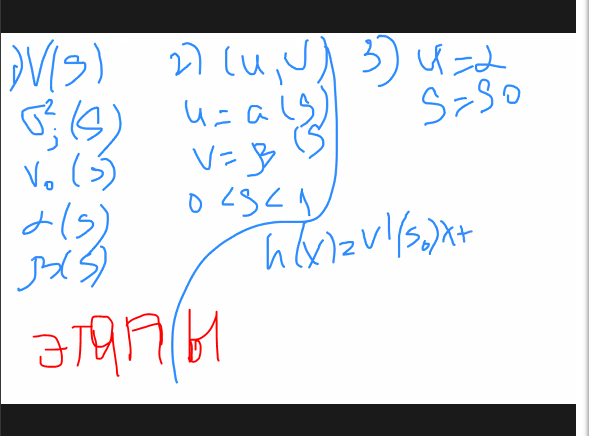


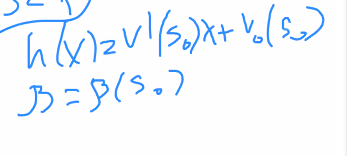
Чтобы найти v0



Вычисляем значение v(s)

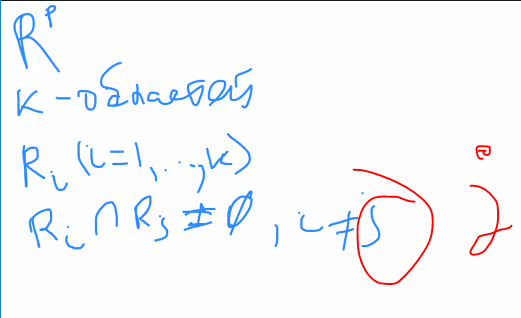
После этого мы сможем квадратичное отклонение найти

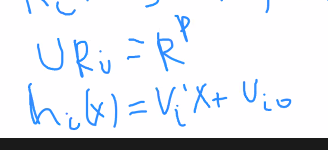




Для обучения используется только один параметр, этот параметр s

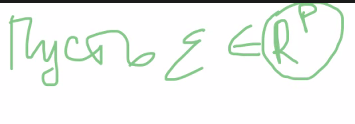
Бывают кусочно линейные классификаторы



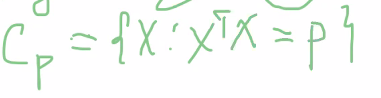


06 06 2020  
Многомерность распределение:

1. Эллипсоидальное распределение

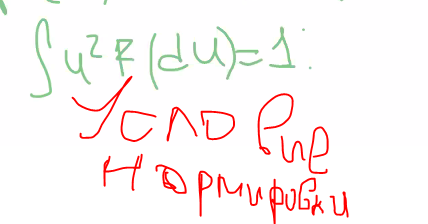


Оно равномерно распределено. R –это пространство

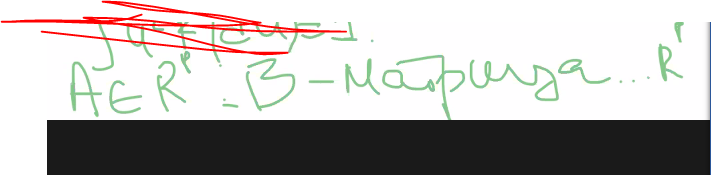


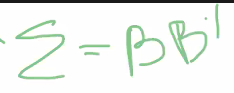
Не отрицательная случайная величина





Тут на ещё нужна матрица В не вырожденого линейного преобразования





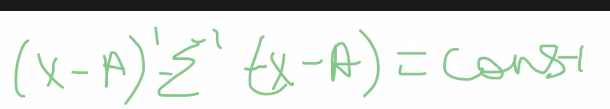
Если



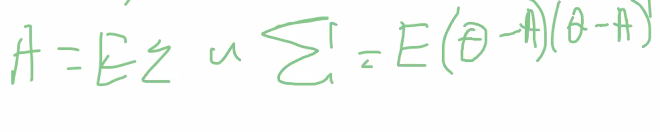
Тетета имеет эллипсоидальное распределение



Почему называется эллипсоидальное?



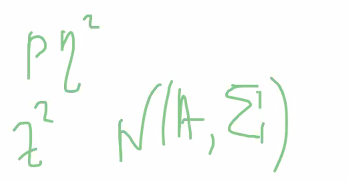
Где А

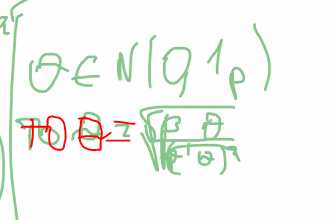


Где тетта постоянна

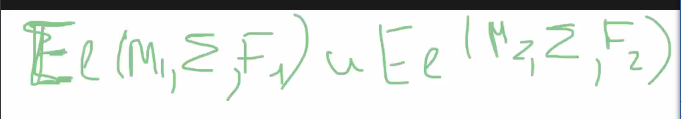


Но если рассматривать частный случай





Выборки рассматривать выборки



Можно сказать, что при дополнительном предположении существует доп



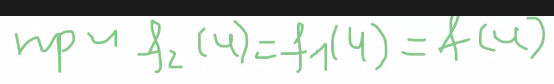
То тогда отношение правдоподобия примет следующий вид



Причём у нас t –это то что мы писали раньше



Следует след. Равенство.



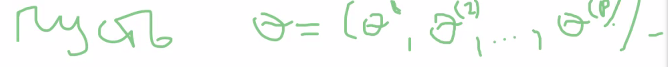
В этом случае при





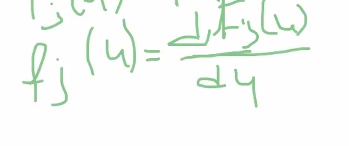
Монотонная ф-я разности

2 многомерное распределение, которые приводят и трансформируют к нормальному распределению





Плотность распределения :



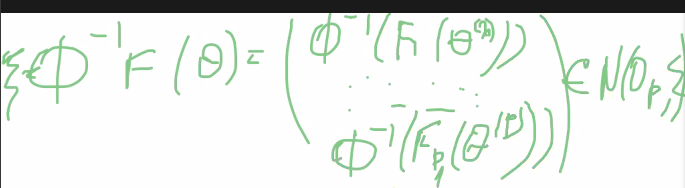
Т-нормальное распределение, причём это не отрицательно определённая матрица размером п на п



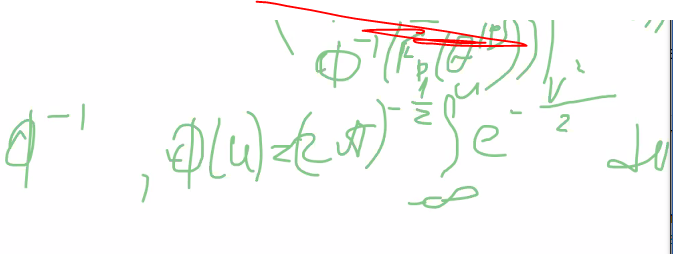
Вектор ф-я одномерного распределения



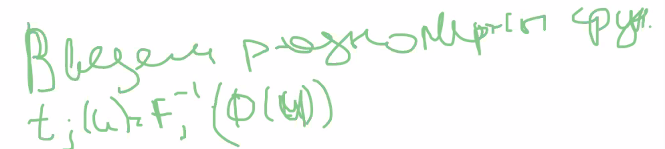
Тогда получается, что

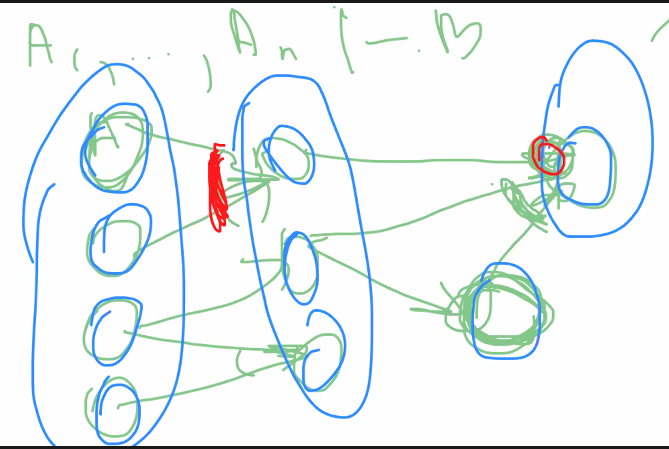


Ф-1 это ф-я обратная от ф, которая даёт нормальное распределение



Получается мы можем ввести p одномерных ф-ий





Найти X\*



Предполагаются дискретным и как правило конечным. Задаются связи и его элементы, мы задаём пространство поиска решений. У нас задача должна быть предствалена каким либо образом.

Самое распространённое-это теория графов

Когда мы рассматриваем интеллектуальные системы, мы можем выделить признаки интеллектуальной системы

1. Развитые коммуникативные способности (характеризует некоторый способ взаимодействия юзера с системой (сири, ок гугл)
2. Способность решать сложные формализованные задачи
3. Способность обучения- получения знаний из разных источников
4. Адаптивность- способность изменения и адаптация системы во внешнем мире

Цель исследования

1 решения задач трудных человек(из-за времени

Трудно формализованная задача:

1. Эта задачи не могут быть заданы в числовой форме
2. Не может быть выражено ввиде точной целевой ф-и
3. Не существует алгоритмичного решения задач
4. Есть алгоритм, но он решает долго

2 большая размерность поиска решения, что переводит к большому перебору

3 данные динамические, изменяются во времени

Экспертные системы решают задачи в конкретной определённой области.

Экспертная система, реализованная на пк, которая решает трудно формулизованные задачи, но это слабый ИИ. Чтобы результат был не хуже чем у эксперта

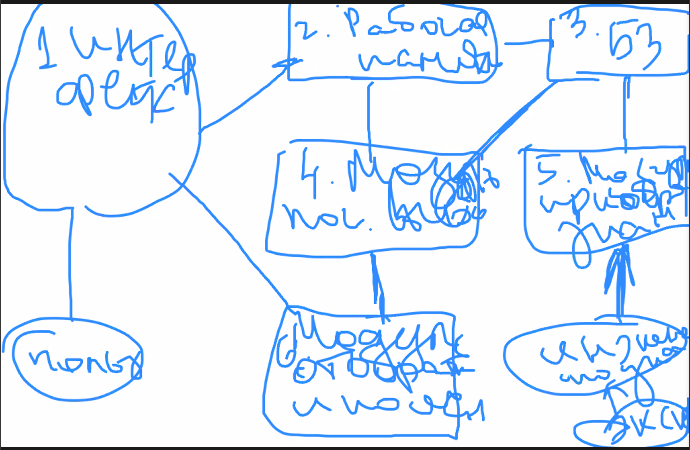


Рис 1 модуль (интерфейс) он взаимодействует с 2 рабочей памятью, а она работает с базой знаний -3. 4 модуль логического вывода. 3 взаимодействует с 4 и 5 модулем приобретений знаний

С 5 работает инженер по знаниям, а эти знания он берёт от эксперта.

4 передает на модуль пояснения. Пояснения связан с 1

Всё эти компоненты реализуются как единое целое

1 Модуль (лингвистический процессор) общение пользователя с системой. Язык желательно, чтобы был естественным. Блок должен принимать входные данные и изменять их на язык для программы. Отдавать должен данные понятные пользователю

2 Модуль Это как правило бд

3 Модуль Она содержит совокупность знаний, заложенных при создании и эти знания нужна для решения задач. Знания в идеале должны быть представлены в виде формальной структуре. данные представлены в виде Фактов и правил. Например Факты мб рентабельность, задолженность, рекламации, удалённость, финансовые состояния. Правила, если рентабьельность у нас высокое и задолженности нет, то финан. состояние хорошее. Предназначены для хранения долгосрочных данных.

Модуль логического вывода. Выполнения процедур над данными Реализация ввиде программного обеспечения. Вывод на основе базы данных.

Модуль отображения и решения. Он позволяет выводить решения и объяснить, юзеру как он получил тот или иной результат (при условии, что человек шарит в теме). Он даёт возможность даёт проанализировать это, а так же после анализа изменить запрос.

Модуль приобретения знаний- для извлечения знаний. Источники знаний могут быть разные. Эти знания считаются наиболее ценными, но они наиболее плохо формализованы. Передачи от эксперта к экспертной системы, наиболее узкое место. Этот блок делается на формировании бд. Её наполняет специалист(инженер) по знаниям.

У нас может быть статичная экспертная система, если говорится о фундаментальных вещах

Но есть динамические, но можно не учитывать измения во времени, во время решения задачи.

У в динамической систеиме есть предъугодатеватель и второй сопрояжения со внешним миром.

В режиме приобретение знаний, с ситемой работает инженер

Три обязательных механизма

1 выбор

2 заложить механизм поиска решения… логический вывод

3

Экспертные системы направлены много задач, но мона выделить

1. Задачи наблюдаемых данных
2. Задачи диагностики-выяления причин которые привели к тому или иному процессу
3. Задачи коррекции-эта тоже саммая диагностика, но допленена рекомендациями, чтобы не было отклонения
4. Задачи прогнозирования( с помощью знаний текущих и прошлых знаний)
5. Задачи планирования
6. Задача диспетчеризация
7. Проектирования
8. Мониторинга и управления

Теперь поговорим, а чем же должна обладать идеальная экспертная система

1. Способность решать сложные задачи в реальных предметных областях
2. Компетенстность, результаты высокого качества при затарате минимального времени и мы можем убедится как это получилось
3. –
4. Способность к интелектуальному рассуждению, логические рассуждения причём общего вида и частного вида

Логические рассуждения общего вида