Инженерия знаний. Методы структурирования знаний. Основные методы извлечения знаний. Методы формализаций знаний. Методы машинного обучения.

Лекция 2

Киселёв Глеб Андреевич
к.т.н., старший преподаватель
ФФМиЕН РУДН
тел.:+79067993329
email: kiselev@isa.ru

Познавательный процесс

- Получение нового знания на основе анализа данных (АД);
- Предсказанием и сохранением результатов исследований (пополнение БЗ прецедентами).

АД – поиск закономерностей

- Ранее не известных: *x4 и x5 возможно ли при диагнозе z1, если z1 линейно выводится из x1-x3?*
- Нетривиальных: *Если и гипотеза 1 линейно выводится из факта 1 и факта 2, то зачем АД?*
- Доступных интерпретации: *что значит <x1,…xN>?*
- Практически полезных;

Подготовка данных для поиска новых знаний (knowledge discovery)

1. Унификация представления данных;
2. Очистка данных;
3. Постановка гипотез об информативных признаках;
4. ML – обработка данных;
5. Валидация гипотез;
6. Пополнение базы прецедентов.

Унификация представления данных

- Преобразования данных из различных источников к единому представлению;
- Учёт неполноты данных;
- Учёт избыточности или недостаточности данных (должна быть гипотеза о методе анализа);

Очистка данных

- Выявление пропущенных данных, шумов, аномальных значений;
- Нормализация данных (преобразование непрерывных значений в дискретные, выбор общей шкалы данных);

Постановка гипотез об информативных признаках

- Выделение наборов компонент, на основе которых будет проведено исследование;

ML – обработка данных

- Построение модели ассоциации, последовательности, классификации или регрессии;

Валидация гипотез

- Предсказание на основе модели;
- Уточнение набора влияющих факторов;

Ассоциации. Задача поиска ассоциативных правил предполагает поиск частых наборов в большом числе наборов данных.

- По смежности в пространстве (Атрезия пищевода – неполное формирование пищевода, часто ассоциированное с трахео-пищеводным свищем);
- По смежности во времени (каузальные связи, формирование матриц условие-эффект);
- По сходству (признак при разной патологии);
- По контрасту (альтернативная связь или признаки отрицания).

Последовательности:

- Секвенциальный анализ. Отличие поиска ассоциативных правил от секвенциального анализа (анализа последовательностей) на примере магазина в том, что в первом случае ищется набор объектов в рамках одной транзакции, т.е. такие товары, которые чаще всего покупаются ВМЕСТЕ. В одно время, за одну транзакцию. Во втором же случае ищутся не часто встречающиеся наборы, а часто встречающиеся ПОСЛЕДОВАТЕЛЬНОСТИ. Т.е. в какой последовательности покупаются товары или через какой промежуток времени после покупки товара "А", человек наиболее склонен купить товар "Б". Т.е. данные по одному и тому же клиенту, но взятые из разных транзакций.

Классификации:

- Задача определения одного из параметров анализируемого объекта на основании значений других параметров (уровень гемоглобина и эритроцитов). Определяемый параметр часто называют зависимой (латентной) переменной, а участвующие в его определении – независимыми переменными. Алгоритмы – деревья решений, опорные вектора, knn и тд.

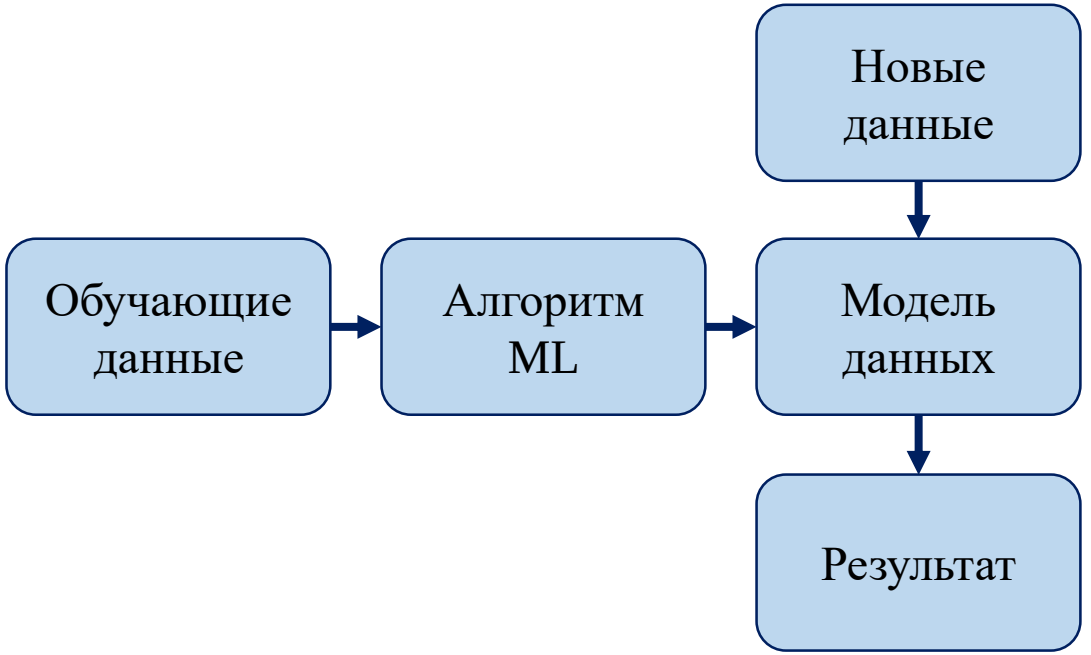
Регрессии:

- Задача построения модели взаимосвязи между несколькими входными переменными и выходной зависимой переменной. Модель может иметь как линейные, так и не линейные связи.

- 1-й – методы «черного ящика» (нейронные сети, методы математической статистики – вычисление весов или коэффициентов близости в некоторой метрике). Не предусматривает содержательного объяснения процесса обучения.

- 2-й – методы, основанные на знаниях, структуры символьных данных, интерпретируемые и отвечающие принципу транспарентности (прозрачности, открытости).

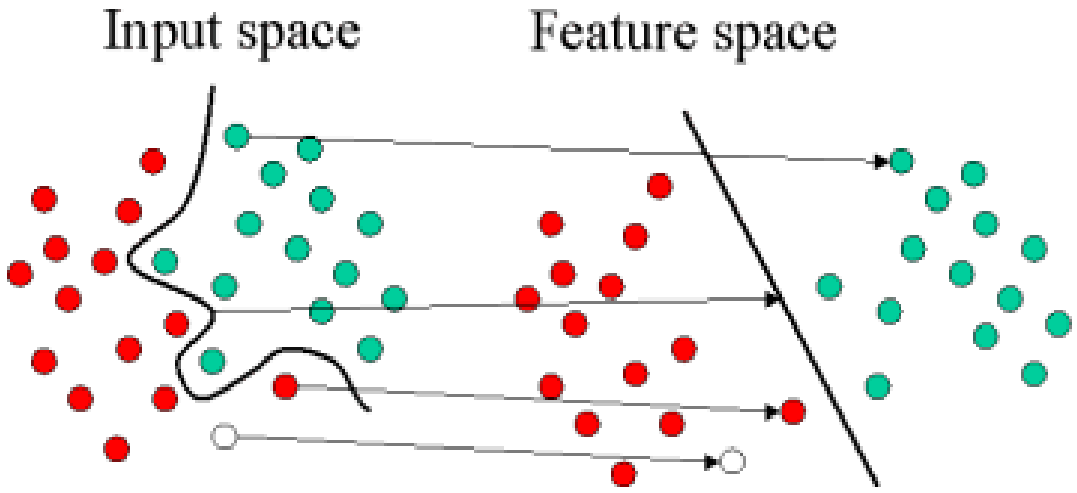
- **Обобщение (индуктивное)**– фундаментальная концепция машинного обучения. Это перенос знаний с ранее предъявленных данных на новые;
- Если слишком сильно «подгонять» модель под какой-то шаблон данных то модель **переобучится** (модель будет плохо работать с новыми данными);
- Алгоритмы ML могут быть как на основе статистически-выводимых уравнений – деревья решений, knn, опорные вектора, так и на основе нелинейных регуляторов – нейросети. *Первые методы используются, когда данных мало, вторые – когда много.*



	Признак 1	**…**	**Признак N-1**	**Признак N**
Диагноз Б	1		0	0
Диагноз А	0		0	1
Диагноз А	0		1	0
Диагноз А	0		1	0
Диагноз Б	1		0	1
Диагноз Б	1		0	1
Диагноз Б	1		0	0

Алгоритм опорных векторов

1. Подготовленный набор данных разделяется на 2 части – положительные и отрицательные примеры. *(Например, лечение подействовало, лечение не подействовало);*
2. Создается набор правил, покрывающий все положительные примеры, но не покрывающий отрицательные;
3. Алгоритм итеративно пытается обобщить описания положительных примеров (назовём их опорными примерами);
4. *Критерий предпочтения* позволяет выбрать лучшее правило в множестве опорных примеров (опорном множестве);
5. Алгоритм строит нелинейную разделяющую плоскость.
6. Алгоритм перегруппировывает исходные объекты с использованием комплекса математических функций, известных как ядра.
   Процесс перестановки объектов – это отображение или преобразование. В результате отображение объектов становится линейно разделимым и вместо построения сложной кривой нужно найти оптимальную линию для отделения объектов (ЗЕЛЕНОГО и КРАСНОГО цвета).



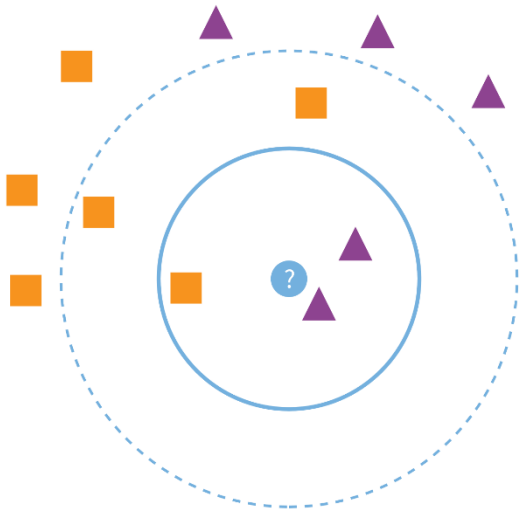
Хороший пример тут:
https://habr.com/ru/articles/105220/

Метод К-ближайших соседей

Дано: набор данных из n наблюдений $X_i(i=1,...,n)$ для каждого задан класс $C_j(j=1,...,m)$, формируются пары X_i, C_j.

Алгоритм KNN состоит из 2 фаз: обучения и классификации. В процессе обучения алгоритм запоминает пары X_i, C_j и использует параметр k – число соседей, по которым происходит классификация.

В фазе классификации, на вход алгоритма подается объект, который располагается на плоскости, описывающей параметры некоторой метрики. Для объекта определяется k ближайших по метрике соседей с имеющимися метками. На основании меток большинства соседей выбирается необходимая метка.



Пример взят с
https://loginom.ru/blog/knn

Кружком представлен объект, который требуется классифицировать, отнеся к одному из двух классов «треугольники» и «квадраты». Если выбрать k=3, то из трёх ближайших объектов два окажутся «треугольниками» и один «квадратом». Следовательно новому объекту будет присвоен класс «треугольник». Если задать k=5, то из пяти «соседей» два будут «треугольники» и три «квадраты», в результате классифицируемый объект будет распознан как «квадрат».

Т-Критерий Стьюдента

Дано:

- Гипотеза H1 – 2 выборки зависимы друг от друга (я считаю, что пациенты, которых лечили препаратом А чувствуют себя лучше, чем те, которых лечили препаратом Б);
- Гипотеза H0 – обратная (консервативная) H1 (препараты идентичны по своему эффекту);
- Данные выборок распределены по нормальному распределению;
- Данные количественны (чсс, уровень метаболита в крови, количество койко-дней);
- Выборки независимы (нет людей, которых лечили препаратом А И препаратом Б).

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{m_1^2 + m_2^2}}$$

в числителе разница средних арифметических по количеству наблюдений

в знаменателе корень квадратный суммы ошибок репрезентативности по этим группам

- Вычисляем количество степеней свободы ((длина выборки 1) + (длина выборки 2) – 2);
- После вычисления t – проверяем в таблице http://dmo.econ.msu.ru/Teaching/ru/stat/Student.htm
- Получаем вероятность правдивости гипотезы H1.

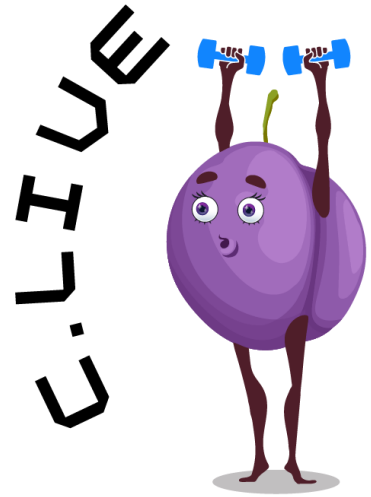
Подробнее:
http://www.machinelearning.ru/wiki/index.php?title=Критерий_Стьюдента

https://www.codecamp.ru/blog/pandas-t-test/

А теперь в Python:

```
from scipy. stats import ttest_ind

group1 = df[df['method']=='A'] group2 = df[df['method']=='B']

#perform independent two sample t-test

ttest_ind(group1['score'], group2['score'])

Ttest_indResult(statistic=-2.6034304605397938, pvalue=0.017969284594810425)
```

Спасибо за внимание!



Руководитель проекта Когнитивный ассистент
старший преподаватель, к.т.н. Киселёв Г.А.
+79067993329
kiselev@isa.ru