Лекция 2.
Фреймворк CRISP-DM. Алгоритмы машинного обучения

—

Юрий Яровиков

План лекции

- Пара слов об искусственном интеллекте
- Стандарт CRISP-DM
- Метрические алгоритмы
- Логические алгоритмы

источник: vas3k.ru

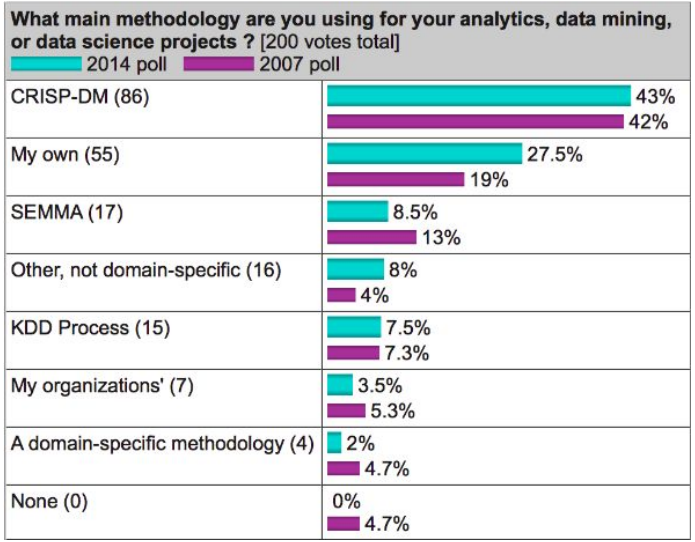
Что могут и что не могут машины

Машина может	Машина не может
Предсказывать	Создавать новое
Запоминать	Резко поумнеть
Воспроизводить	Выйти за рамки задачи
Выбирать лучшее	Убить всех людей

источник: vas3k.ru

Как автоматизировать Data Mining?

- “Хаотический” анализ данных хорош для стартапов и курсов машинного обучения
- Для достижения максимальной эффективности необходима методология работы с данными
- CRISP-DM — наиболее распространённая из них



Источник: knuggets.com

Фреймворк CRISP-DM

- CRoss Industry Standard Process for Data Mining
- Стандарт, описывающий общие процессы работы с данными
- Методология должна быть адаптирована к конкретному проекту

Фреймворк CRISP-DM: бизнес-анализ

- Определение бизнес-целей
  например: уменьшение оттока клиентов
- Оценка текущей ситуации
  ресурсы, риски, окупаемость
- Определение целей аналитики
  целевые метрики, критерии успеха
- Подготовка плана проекта
  оценка фаз проекта

Фреймворк CRISP-DM: анализ данных

- Сбор данных
  анализ источников, сбор недостающих данных
- Описание данных
  формирование БД, вычисление статистик
- Изучение данных
  исследование, выделение полезных атрибутов
- Проверка качества данных
  пропущенные значения, ошибки

Фреймворк CRISP-DM: подготовка данных

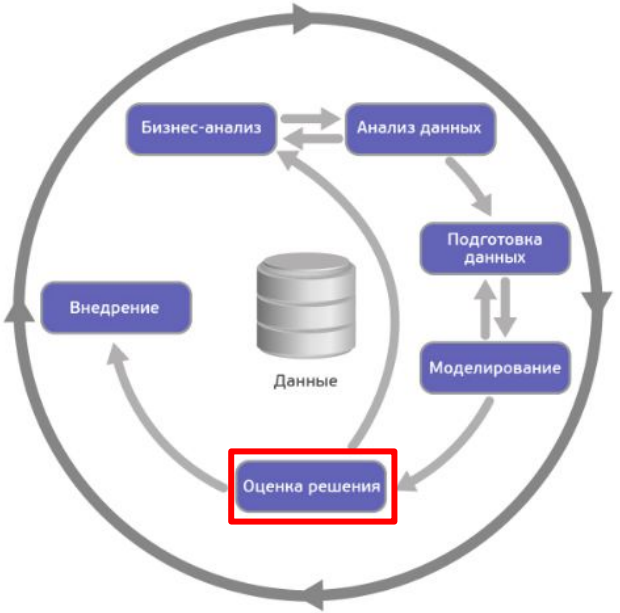
- Выборка данных
  выбор релевантных атрибутов
- Очистка данных
  восстановление значений, кодировка и пр.
- Генерация данных и признаков
  конструирование новых атрибутов
- Интеграция данных
  слияние данных нескольких источников
- Форматирование данных
  приведение к формату, пригодному для анализа

Фреймворк CRISP-DM: моделирование

- Выбор алгоритмов
  учёт специфики моделей к конкретной задаче
- Подготовка плана тестирования
  разбиение данных на train и test, CV и пр.
- Обучение моделей
  fit!
- Оценка качества моделей
  predict!

Фреймворк CRISP-DM: оценка решения

- Оценка результатов
  формулировка результата в бизнес-терминах
- Оценка процесса
  что было сделано хорошо, что можно улучшить
- Определение следующих шагов
  внедрять модель или улучшать её



Фреймворк CRISP-DM: внедрение

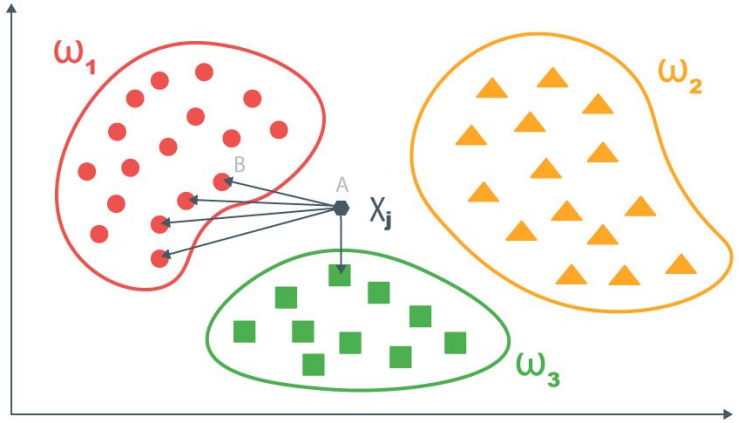
- План внедрения
  технический план, подготовка к эксплуатации
- Планирование мониторинга и поддержки
  отслеживание показателей, retrain
- Подготовка отчёта
  главный документ с информацией о проекте

Метрические алгоритмы

Метрический алгоритм — алгоритм, опирающийся на геометрическую структуру данных в пространстве объектов.

Алгоритм k ближайших соседей:

- Хотим предсказать класс объекта x
- Вычисляем $f_1(x), f_2(x), ..., f_n(x)$
- Находим k ближайших объектов из обучающей выборки
- Предсказание = самый популярный класс среди соседей



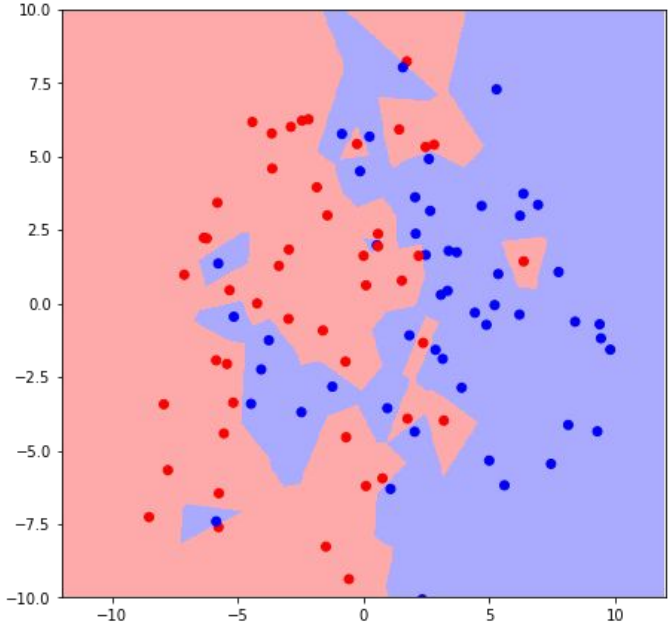
Алгоритм k ближайших соседей

Преимущества алгоритма:

- быстрый и лёгкий в понимании
- легко обобщается для задачи регресии (как?)

Недостатки алгоритма:

- необходимо хранить всю обучающую выборку
- неустойчив к масштабу признаков



алгоритм одного ближайшего соседа

Логические алгоритмы классификации

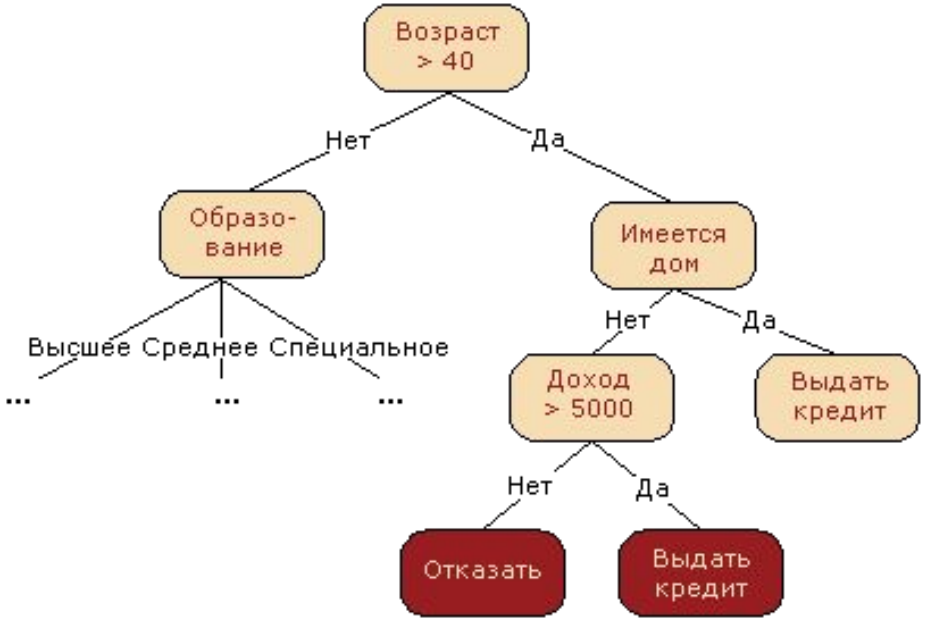
Логический алгоритм — алгоритм, использующий логические закономерности в данных.

Примеры простых решающих правил:

- Если в анкете указан домашний телефон и зарплата клиента $>$ \$2000 и размер кредита $<$ \$5000, то кредит выдать.
- Если возраст пациента $>$ 60 и пациент ранее перенёс инфаркт, то операцию не делать.

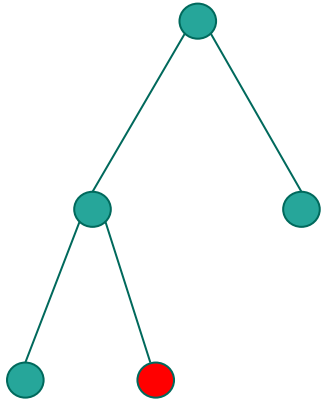
Решающие деревья

- В каждой вершине дерева находится вопрос
- В зависимости от ответа на вопрос, алгоритм направляется в нужную ветвь дерева
- Листы дерева соответствуют решению алгоритма



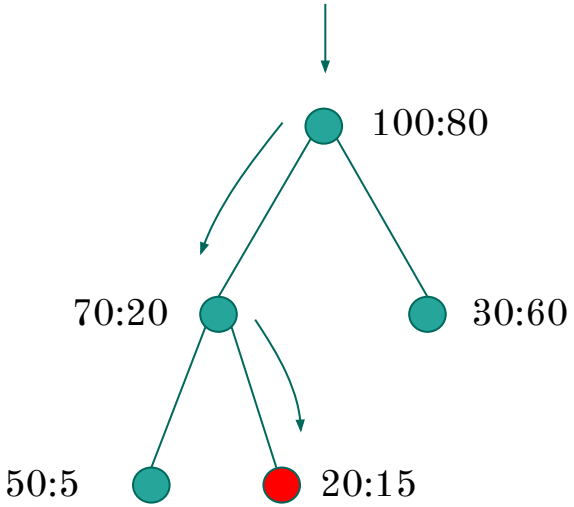
Обучение решающего дерева

- Находимся в красной вершине



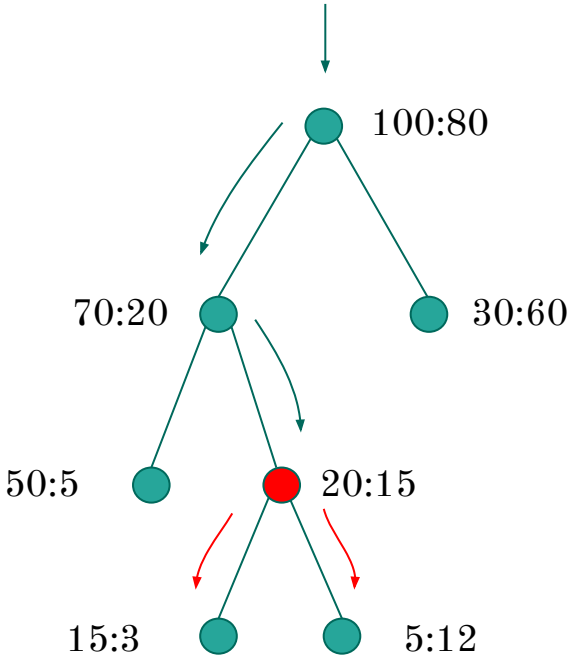
Обучение решающего дерева

- Находимся в красной вершине
- До красной вершины дошла часть объектов обучающей выборки



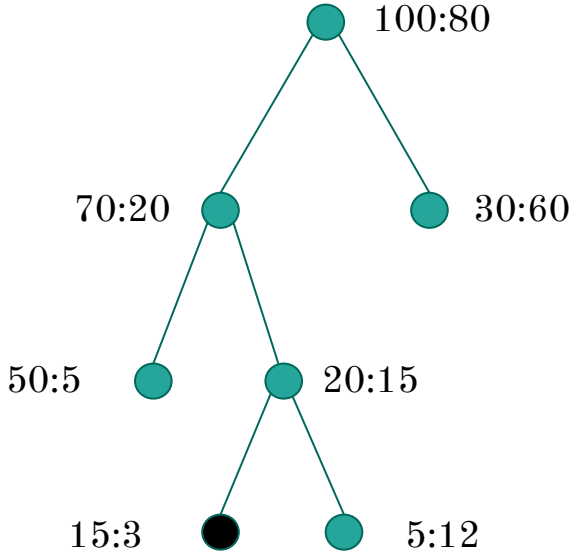
Обучение решающего дерева

- Находимся в красной вершине
- До красной вершины дошла часть объектов обучающей выборки
- Находим решающее правило так, чтобы объекты, дошедшие до красной вершины, хорошо разделялись по искомым классам



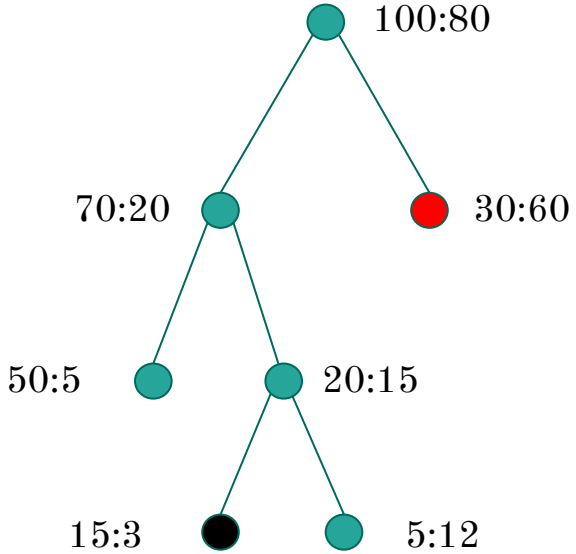
Обучение решающего дерева

- Находимся в красной вершине
- До красной вершины дошла часть объектов обучающей выборки
- Находим решающее правило так, чтобы объекты, дошедшие до красной вершины, хорошо разделялись по искомым классам
- Одна из нижних вершин стала терминальной



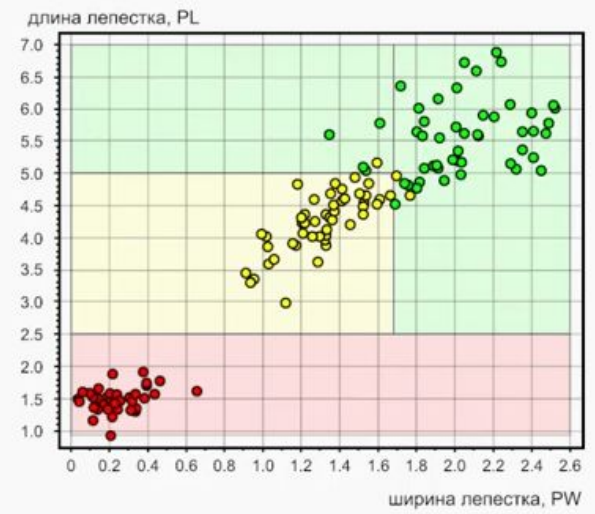
Обучение решающего дерева

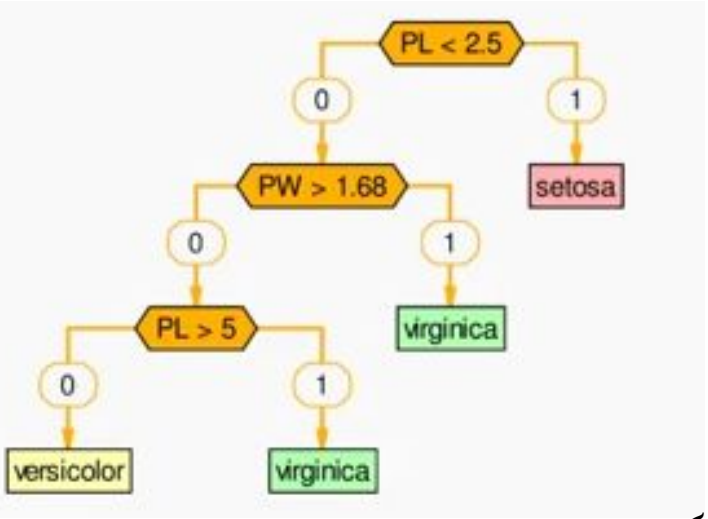
- Находимся в красной вершине
- До красной вершины дошла часть объектов обучающей выборки
- Находим решающее правило так, чтобы объекты, дошедшие до красной вершины, хорошо разделялись по искомым классам
- Одна из нижних вершин стала терминальной
- Повторяем с другой вершиной



Решающее дерево на примере Ирисов Фишера

Задача Фишера о классификации ирисов на три класса. В выборке по 50 объектов каждого класса, у каждого объекта 4 признака





В осях двух самых информативных признаков два класса разделились без ошибок, на третьем — три ошибки.

Решающий лес

- Построим совокупность решающих деревьев, каждое из которых будем обучать по случайной подвыборке и случайному подмножеству признаков
- Каждое дерево приняло решение о классификации
- Будем принимать окончательное решение обычным голосованием
- Выходит эффективно

Резюме

Преимущества решающих деревьев:

- интерпретируемость
- допускаются разнотипные данные
- возможность обхода пропусков

Недостатки решающих деревьев:

- склонны к переобучению
- фрагментация

Случайный лес — хороший способ устранить недостатки