Введение в Data Science
Занятие 13. Заключительное

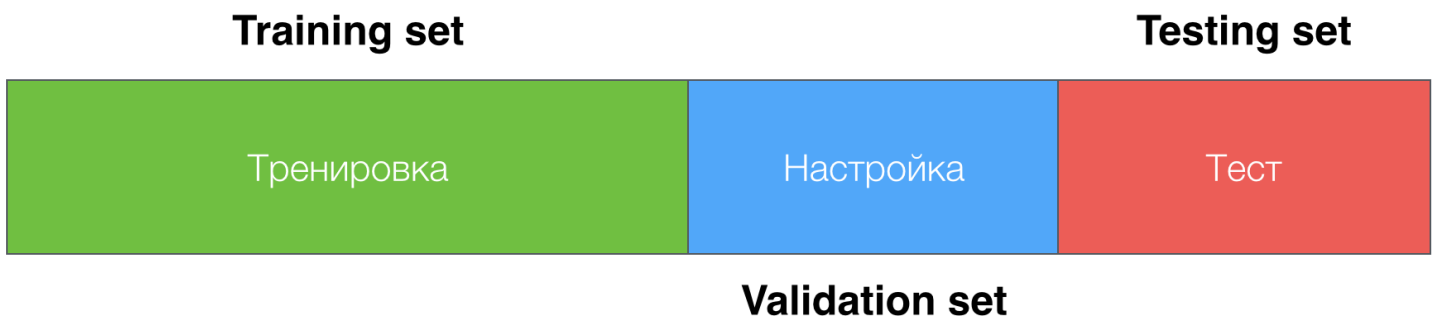
Николай Анохин Михаил Фирулик

31 мая 2014 г.

Предобработка данных

Заключение

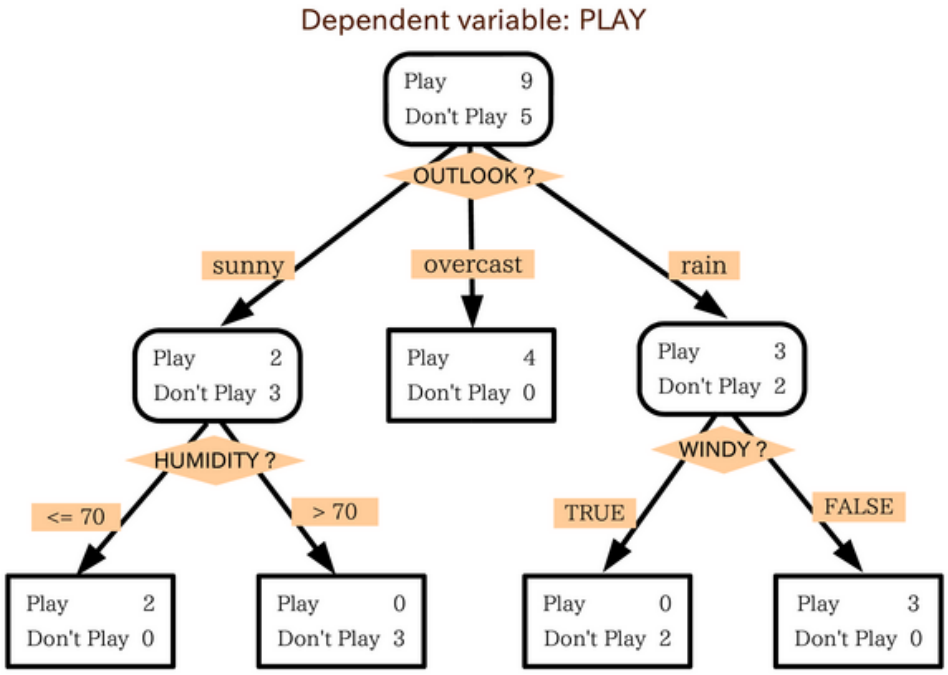
Выбор параметров модели



Предобработка данных

- **выбор признаков** / feature selection
- дискретизация признаков / feature discretization
- очистка данных / data cleansing
- уменьшение размерности / dimensionality reduction

Зачем выбирать признаки?



1. Качество
   подвержены влиянию случайных признаков: DT, KNN, ...
2. Скорость
   хотя отбор признаков на практике медленный
3. Интерпретируемость

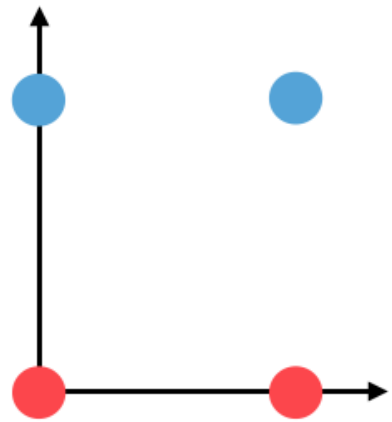
Подходы к выбору признаков

- Ручной

  лучше, если вы знаете, что делаете
- Автоматизированный
  - Схемо-независимый / Scheme-independent
  - Схемо-зависимый / Scheme-specific

Схемо-независимый подход

- Выбрать столько, чтобы идентифицировать каждый объект
- Техника near-hit, near-miss
- С помощью выбранного критерия качества
- С помощью алгоритма машинного обучения *Decision Tree, Linear Model*



Критерии качества признаков

Сколько?

- Фиксированное количество
  Пример: лучшие 100 признаков
- Percentile
  Пример: лучшие 20 процентов

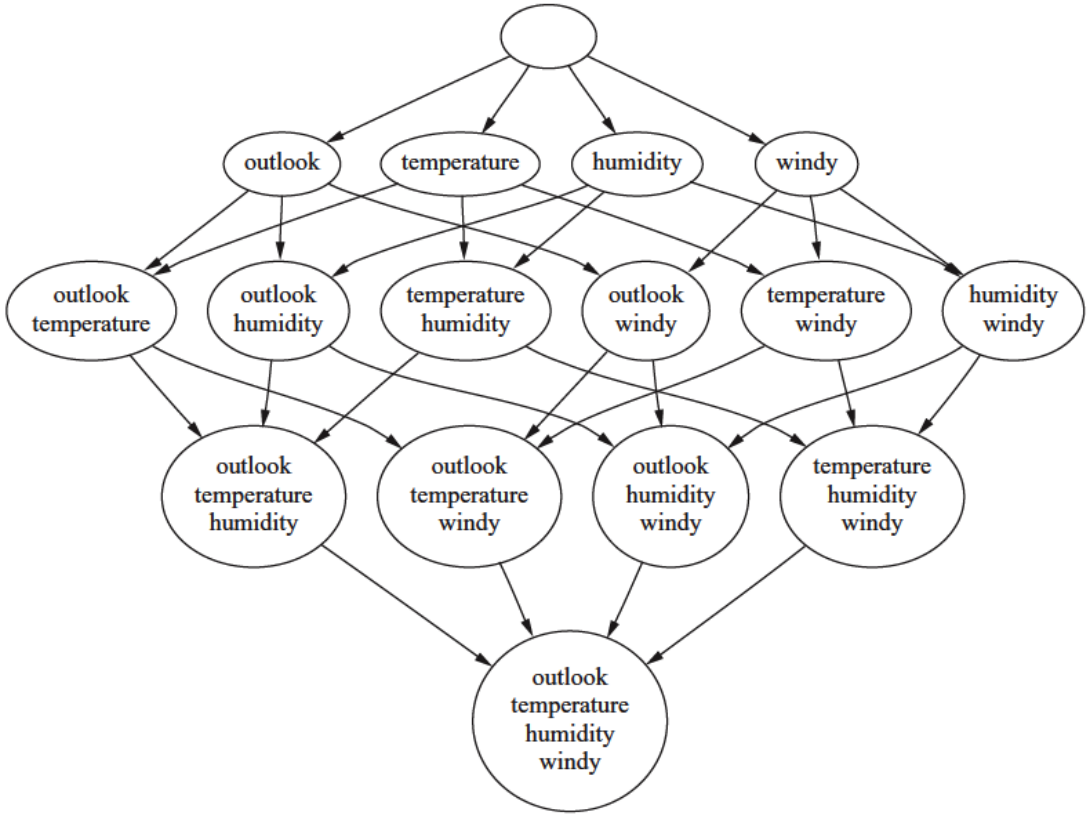
Как?

- Mutual Information

$$I(X, Y) = \sum_x \sum_y p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

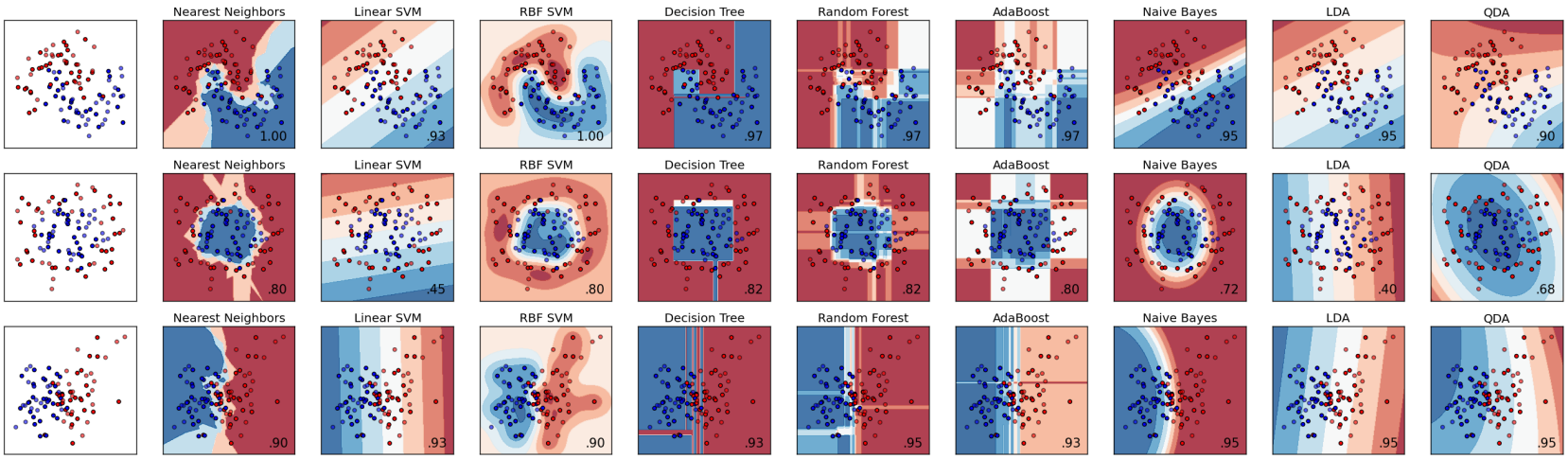
- Statistical Tests
  Chi^2, *binomial*, ...

Схемо-зависимый поиск в пространстве признаков

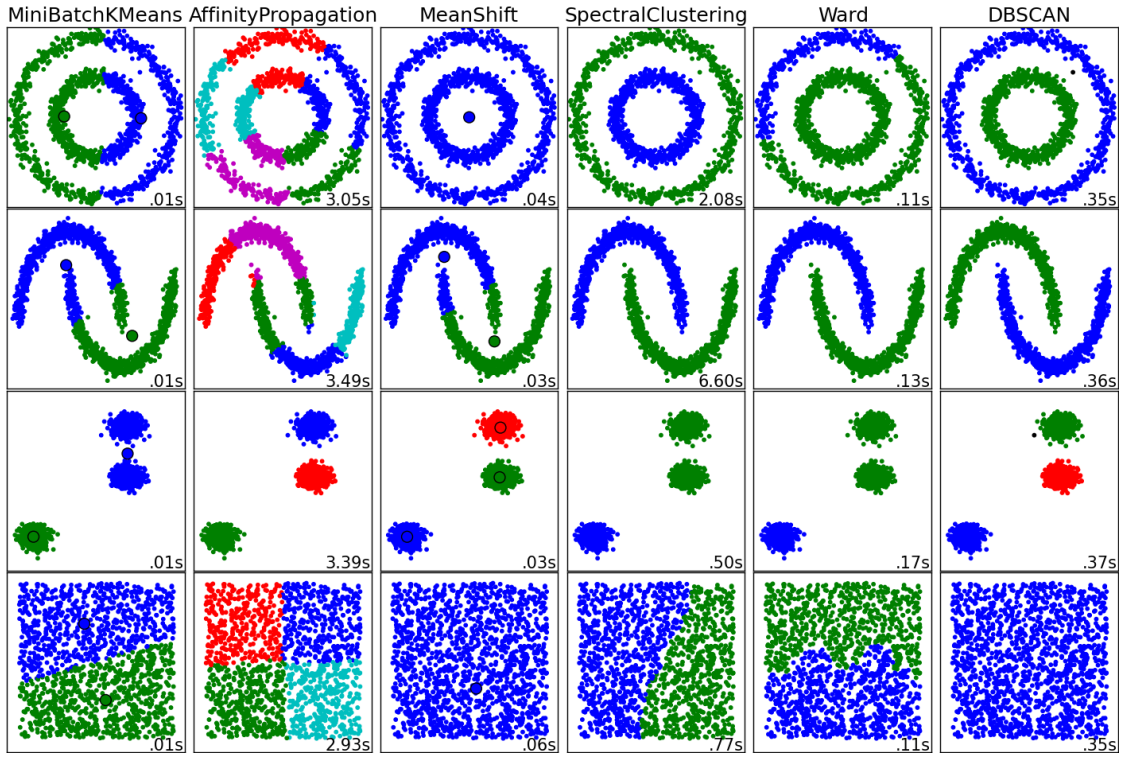


- Forward-selection
- Backward-elimination

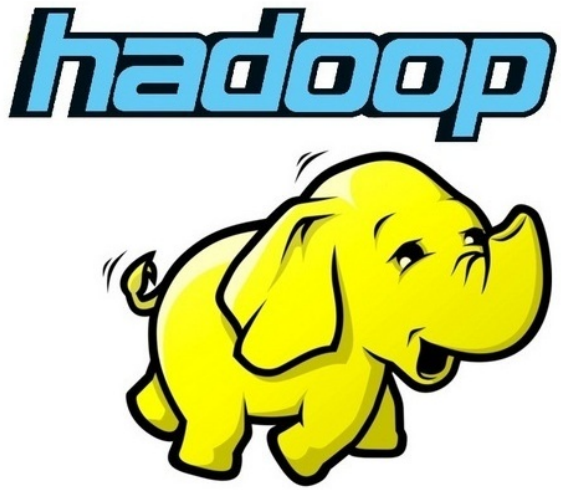
Что мы рассмотрели: классификация



Что мы рассмотрели: кластеризация

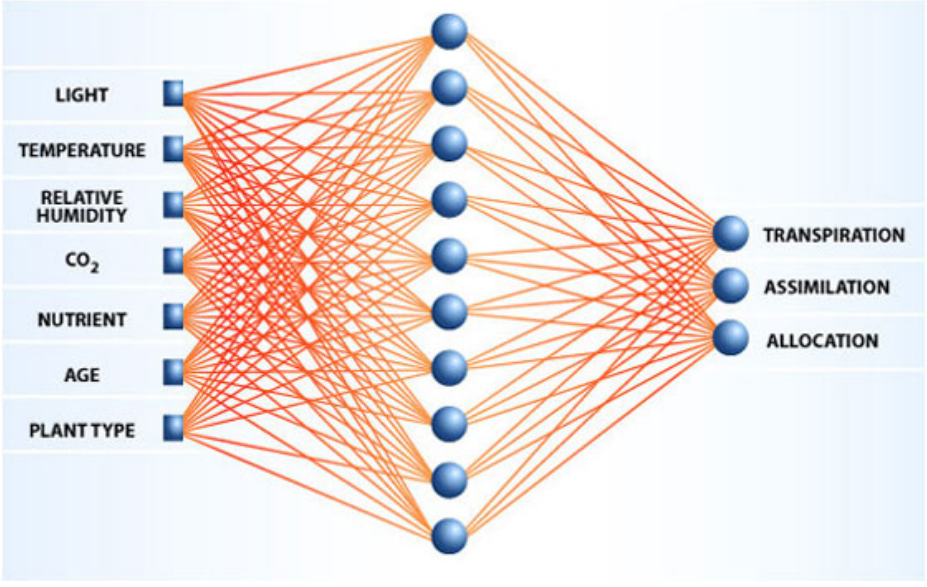


Что мы рассмотрели: технологии



Что мы не рассмотрели

- neural networks
- genetic algorithms
- dimensionality reduction
- semi-supervised learning
- reinforcement learning
- NLP, SNA
- и еще много чего



Что делать дальше

- Kaggle `http://blog.kaggle.com/`
- Hilary Mason `http://www.hilarymason.com/`
- Alex Holmes `http://grepalex.com/`
- Cloudera `http://blog.cloudera.com/`
- Coursera
- Аспирантура (+PhD)
- Трудоустройство
- Собственный проект

m.firulik@corp.mail.ru n.anokhin@corp.mail.ru

На самом деле, еще не совсем все

Результаты (17 июня 00.00)

- Код на bb
- Проклассифицированные пользователи

Презентация (17 июня 09.30)

- Использованные признаки
- Выбранная модель
- Результаты классификации

Время: 10 + 5 мин