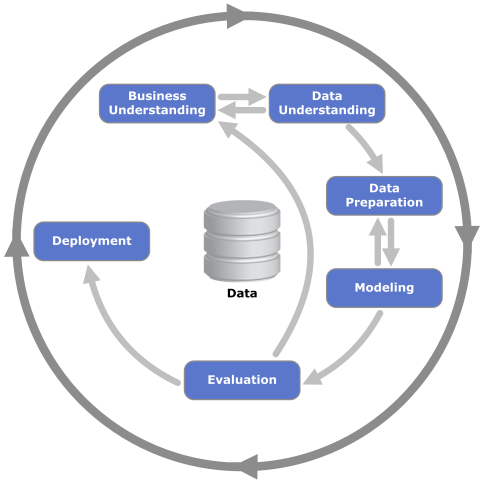
ТЕХНОСФЕРА

Лекция 11
Data Mining в реальных системах

Николай Анохин

30 ноября 2015 г.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

План занятия

Работа с признаками

Алгоритмы машинного обучения

Особенности реальных систем

Что дальше

Работа с признаками

Конструирование признаков

Лог посещения пользователями Интернет-сайтов

```
1432068600.002494
46.148.52.217
22B695259
22/159/5/0/0/6/1/1.000000/16.0
2579437
http://vk.com/ivan_se
http://vk.com/friends?act=find
1152*864
1137*747
1432068601241
Поиск друзей
```

Преобразование признаков

- Дискретизация
- Проекции
  - PCA
  - Random projections
- Заполнение отсутствующих значений
- Удаление шума
- Преобразование категориальных классов в бинарные
  - One-vs-rest
  - One-vs-one
  - Error correcting output codes

```
A: 1 1 1 1 1 1 1
B: 0 0 0 0 1 1 1
C: 0 0 1 1 0 0 1
D: 0 1 0 1 0 1 0
```

Отбор признаков

Как “нерелевантные”, так и “релевантные” признаки могут быть вредными

1. Независимо от алгоритма обучения: backward elimnation, forward selection
   - mutual information
   - коэффициент корреляции
   - линейная модель
   - генетические алгоритмы
2. С использованием алгоритма обучения
   (кросс-валидация, hold-out)

Алгоритмы машинного обучения

Некоторые правила 1

Первое правило машинного обучения

Если нет необходимости, не использовать машинное обучение

Второе правило машинного обучения[1]

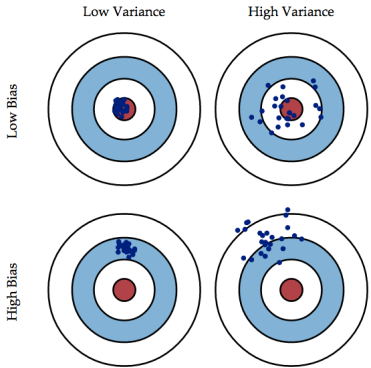
$$LEARNING =$$

$$REPRESENTATION + EVALUATION + OPTIMIZATION$$

[1] A Few Useful Things to Know about Machine Learning

Некоторые правила 2

- Обобщающая способность имеет значение
- Только данных не достаточно
- У переобучения много видов



Некоторые правила 3

- Интуиция подводит в многомерных пространствах
- Больше данных лучше, чем сложный алгоритм
- Обучайте много моделей

Все модели имеют недостатки[2]

Подход	Что хорошо	Что плохо
bayesian learning	хорошо работает на маленьких данных	трудно обосновать априорные распрделения, вычислительно сложные
градиентный спуск	вычислительно эффективен, оптимизируем что нужно	подбор параметров для сходимости, переобучение
kernel	натуральное выражение схожести через ядро	подбор ядра, медленный
деревья решений	быстрый и автоматизированный	крайне нестабильный
boosting	хорошее качество	выбор алгоритма, предположение о weak learner нарушается

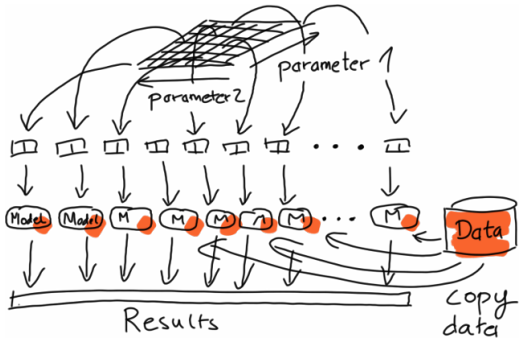
[2] All Models of Learning have Flaws

8 худших алгоритмов[3]

- Linear regression
- Traditional decision trees
- Linear discriminant analysis
- K-means clustering
- Neural networks
- Maximum Likelihood estimation
- Density estimation in high dimensions
- Naive Bayes

[3] The 8 worst predictive modeling techniques

Отбор модели занимает очень много времени[4]

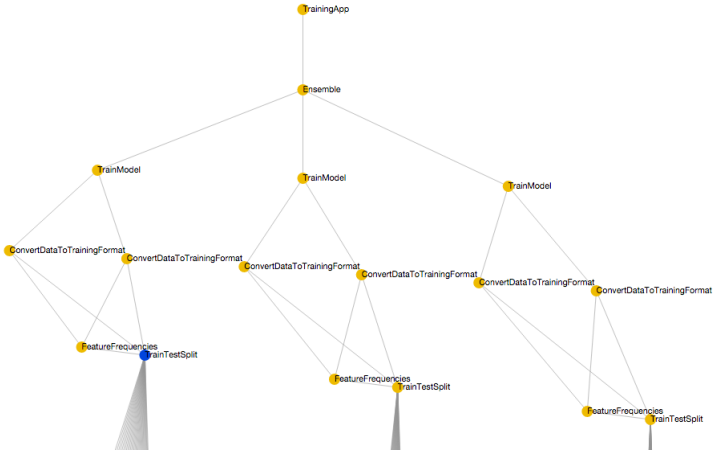


(также, как и работа с признаками)

[4] Three Things About Data Science You Won't Find In the Books

Особенности реальных систем

Пример обучения модели в задаче классификации



Особенности реальной системы

- Очень грязные данные
- Простые модели
- Проверка качества на всех этапах
- Мониторинг и логирование

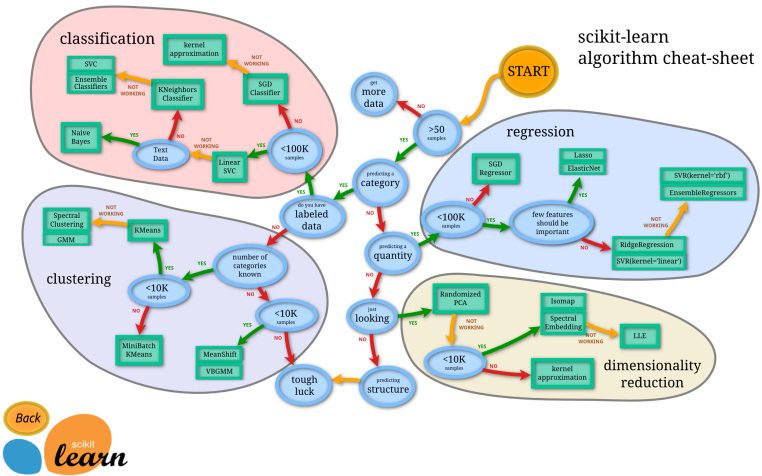
Что дальше

Изучение Data Mining

1. Data Mining II, Hadoop, Инфопоиск
2. Kaggle
3. Литература и статьи
   - Техблог Twitter
   - Техблог Netflix
   - Техблог Spotify
   - Reddit про MachineLearning
   - Подкаст про машинное обучение
   - DataViz

Junior Data Scientist: необходимые навыки

1. Все базовые модели и алгоритмы
2. Знание языка высокого уровня и соответствующие научные библиотеки (R, python, Matlab)
3. Базовые структуры данных и алгоритмы (сортировки, деревья, хэш таблицы и графы)
4. Опыт обработки больших объемов данных точно будет плюсом
5. Умение разбираться с научной литературой


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
classification
kernel approximation
SVC
Ensemble Classifiers
KNeighbors Classifier
SGD Classifier
Naive Bayes
Text Data
Linear SVC
<100K samples
clustering
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
<100K samples
few features should be important
RidgeRegression
SVR(kernel='linear')
dimensionality reduction
Randomized PCA
Isomap
Spectral Embedding
LLE
<10K samples
kernel approximation
NOT WORKING
YES
NO
Back
scikit learn

Вопросы