

# Машинное обучение: начало



**ИГОРЬ КУРАЛЁНОК**

К.Ф.-М.Н., ЯНДЕКС/СПБГУ

# Знакомство: Куралёнок Игорь



30-ка

Компиляторы

ИМ-ПУ СПбГУ

Оценка текстового поиска

Sun Microsystems

Полнотекстовый поиск

JetBrains

Машинное обучение

Яндекс

Обработка сигналов

Руководитель группы модернизации поиска.  
Яндекс.

+7(921)9031911

# Что почитать?



- Википедия (лучше en)
- Т. Hastie, R. Tibshirani, J. Friedman “The elements of Statistical Learning”
- Т. Mitchell “Machine Learning”
- Труды конференций: ICML, KDD, NIPS, CIKM,...
- Журналы: JMLR, JML, JIS, NC
- Видео курс: [www.ml-class.org](http://www.ml-class.org)

# Какие у нас цели?



- Уметь сформулировать задачу в терминах ML
- Найти подходящий класс решающих алгоритмов по формулировке
- Ориентироваться в области и знать «где посмотреть» существующие решения
- Понимать границы применимости

# Что нужно, чтобы понять?



- ТВ и МС
- Линейная алгебра
- Язык программирования

# Как отчитываться?



- К концу обучения сделать 15 минутную презентацию по применению ML в вашей любимой задаче.
- Задачки на Octave
- Ошибки к лекциям и в слайдам :)

# Машинное обучение: определения



*Tom M. Mitchell:* A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

*Webster:* machine learning - The ability of a machine to improve its performance based on previous results.

*Ru.Wikipedia:* Машинное обучение — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

# Немного истории



- 50-60 базы знаний, полнотекстовый поиск, распознавание образов, нейронные сети
- 70-е символьный вывод, Quinlan ID3 деревья, разумные практические результаты, VC-оценки
- 80-е первые конференции, много практического применения, активное применение кластеризации в анализе
- 90-е повторное семплирование в ML, SVM, применение в IR, ML  $\neq$  DM, LASSO, bagging, boosting, CF валидация
- 00-е Compressed sensing, развитие ансамблей,...



# Основные понятия



- Область работы
- Опыт = Data Set = DS
- Целевая функция = Target
- Решающая функция

# Какое бывает обучение



Делить можно по:

- способу генерации DS;
- виду целевой функции;
- классу решающих функций;

# Деление по способу формирования DS/U



- Transductive
- Обычное
- Активное
  - Стохастическая оптимизация
  - Бюджетное
  - Бандиты
- Необычное
  - Online learning
  - Reinforcement learning

# Transductive learning



1. Фиксируем множество примеров
2. Фиксируем рабочее множество
3. Обучаемся на всех/доступных примерах

$$F_0 = \operatorname{argmax}_{F(X,Y)} (T(Y|F))$$

# Обычное обучение



1. Фиксируем множество примеров
2. Определяем генеральную совокупность
3. Обучаемся на доступных примерах

$$F_0 = \operatorname{argmax}_{F(X)} (\mu_{\xi \sim U(\Gamma)} (T(\xi|F)))$$

# Активное обучение



1. Фиксируем множество примеров
2. Определяем генеральную совокупность
3. Обучаемся на доступных примерах
4. Пополняем множество примеров по просьбе алгоритма и переходим к п. 3

$$F_i = \underset{F(X_i = X_{i-1} \cup x_i, x_i \in \Gamma \setminus X_{i-1})}{\operatorname{argmax}} (\mu_{\xi \sim U(\Gamma)} (T(\xi | F)))$$

# Активное обучение



1. Стохастическая оптимизация:  $x_i \sim U(\Gamma)$

2. Бюджетное

$$w : \Gamma \rightarrow \mathbb{R}$$
$$\sum_i w(x_i) < B$$

3. Бандиты

$$\Gamma = \cup_i \chi_i$$
$$x_i \sim \chi_{n(i)}$$

# Деление по целевой функции



- **С учителем**
  - Классификация
  - Аппроксимация (экстраполяция)
  - Metric learning
  - Последовательности
- **Без учителя**
  - Кластеризация
  - Уменьшение размерности
  - Representation Learning
- **Смешанные**
  - Кластеризация с условиями
  - Все те же, что и с учителем
  - Transfer learning



# Обучение с учителем

$$\Gamma = \{\gamma_i | \gamma_i = (x_i, y_i), x_i \in \mathbb{X} \ y_i \in \mathbb{Y}\}$$

- Классификация

$$\mathbb{Y} = \{-1, 1\}$$

$$\mathbb{Y} = \{1, \dots, n\}$$

$$\mathbb{Y} = \{0, 1, \dots, n\}$$

- Аппроксимация (экстраполяция)

$$\mathbb{Y} = \mathbb{R}$$

- Metric learning = Классификация

- Последовательности

$$\mathbb{Y} = \{(b_i, e_i) | b_i < e_i, e_i, b_i \in \mathbb{N}\}$$

$$\mathbb{X} = 2^A$$

# Другое обучение



- **Без учителя**

- Кластеризация
- Уменьшение размерности
- Representation Learning

- **Смешанные**

- Кластеризация с условиями
- Все те же, что и с учителем
- Transfer learning

# Деление по решающей функции



- Линейные решения
- Графы
- Нейронные сети (ANN)
- Параметрические семейства функций
- Instance based learning
- Предикаты
- Ансамбли

# Деление по решающей функции (1)



## ● Линейные решения

- Линейная регрессия, логистическая регрессия
- Скрытый дискриминантный анализ (LDA/QDA\*)
- LASSO
- SVM
- LSI\*

# Деление по решающей функции (2)



- **Графы**
  - Деревья решений
  - Байесовы сети
  - Conditional Random Fields
- **Нейронные сети (ANN)**
  - Персептронные сети
  - Сети Хопфилда
  - Машины Больцмана
  - Сети Кохоннена

# Деление по решающей функции (3)



- **Параметрические семейства функций**
  - Сэмплирование
  - Генетические алгоритмы
  - PLSI/LDA/прочие модели с распределениями (им нет числа)
- **Instance based learning**
  - kNN

# Деление по решающей функции (4)



- **Предикаты**
  - Логические выражения
  - Регулярки/NFA/DFA
- **Ансамбли**
  - Просто ансамбли
  - Bagging
  - Boosting
  - BagBoo/BooBag

Машинное Обучение: Начало



**ОТ СЕБЯ ТЕНА**



# Дедуктивные/индуктивные методы



Индуктивные	Дедуктивные
Полагаются на статистику	Полагаются на prior knowledge
Используют классы элементарных функций	Решающая функция следует из предполагаемой структуры
Работают в любой области	Привязаны к данным
Знание области отражается на составлении target	Понимание области меняет решающую функцию
<i>Логистическая регрессия</i>	<i>LDA</i>
<b>Для вхождения в область, при больших размерностях</b>	<b>Небольшие размерности, «давно тут сидим»</b>

# Data Mining vs. Machine Learning



Data Mining	Machine Learning
Выявление «скрытых данных»	Оптимизация целевой функции
Больше про данные	Больше про методы
<i>«Мы применили такой метод и получили клевые результаты на таких стандартных данных»</i>	<i>«Предложили новый метод, который работает круче чем другие на нескольких датасетах (возможно даже синтетика)»</i>
SIGIR, WSDM, WWWC, ...	ICML, CIKM, ...

# Artificial Intelligence vs. Machine Learning



Artificial Intelligence	Machine Learning
Устройство умных машин	Оптимизация целевой функции
Больше про мат. моделирование	Больше про методы
<i>«Мы придумали как формализовать задачу игры в шахматы, применили такие методы и обыграли человека»</i>	<i>«Предложили новый метод, который работает круче чем другие на нескольких датасетах (возможно даже синтетика)»</i>
AAAI, IJCAI, ...	ICML, CIKM, ...

# Применение ML



- Практически везде (дайте задачку, я попробую придумать применение)
- Есть два больших класса работ

	Академические	Практические
<b>Цели</b>	Существуют ситуации, когда работает хорошо	Обеспечивает измеряемое качество на множестве примеров
<b>Искать</b>	Красивые идеи, хорошую математику	Работающие вещи, много грязных приемов
<b>Смотреть</b>	Конференции	Соревнования