Введение в машинное обучение

Лекция 2

Гапанюк Ю.Е., ИУ-5, 4 семестр

Введение

Машинное обучение (википедия)

- Machine Learning, ML.
- Фактически является аналогом термина «обучение по прецедентам», который использовался в Data Mining.
- Основная задача – предсказание результата на основе предыдущих накопленных данных. Накопление данных называют обучением, поэтому используется термин «машинное обучение».
- Данные могут быть не упорядочены по времени (задачи классификации, регрессии) или упорядочены (прогнозирование временного ряда).
- Является набором наиболее низкоуровневых методов в ИИ. Применяемые алгоритмы очень сильно зависят от набора данных, на разных наборах данных разные алгоритмы могут показывать очень разное качество.
- Фактически основная задача – подобрать алгоритм, который покажет приемлемое качество предсказания на заданном наборе данных и не будет переобучаться.
- Появился лозунг «Data is the new science», то есть накопленные массивы данных определяют характер методов их обработки.

Машинное обучение - методы

- Фактически Data Mining вообще и «машинное обучение» в частности является прикладной дисциплиной, которая базируется на следующих методах:
  - **Теория вероятности и математическая статистика.**
  - **Численные методы и теория оптимизации (к сожалению сейчас не читается на кафедре).**
  - Теория систем и системный анализ, СППР (курс «архитектура АСОИУ»).
  - Теория графов.
  - Другие подходы к ИИ, рассмотренные на предыдущей лекции.

Data Scientist vs Data Engineer

- В курсе мы планируем говорить про машинное обучение, анализ данных. Традиционно эту роль IT-специалиста называют «Data Scientist», то есть специалист по изучению данных и построению моделей, аналитик данных.
- Но данные для анализа нужно где-то хранить, передавать, обрабатывать и т.д. Роль IT-специалиста, которые это обеспечивает называют «Data Engineer».
- Авторы курса http://newprolab.com/ru/dataengineer/ : «*Data engineer – это тот, кто делает всю ту бигдату, про которую вы слышали, возможной*».
- В основном роль IT-специалиста «Data Engineer» сейчас связана именно с обработкой больших данных. (Фактически, такой курс сейчас читается на кафедре в магистратуре).
- Data Engineer это специалист по базам данных, системный администратор (с хорошим знанием виртуализации и Big Data фреймворков), разработчик ETL-процессов (https://ru.wikipedia.org/wiki/ETL).

Машинное обучение – как действовать?

- Не смотря на то, что в курсе мы будем рассматривать «сложившиеся», «относительно стандартные» подходы к решению задач, их не стоит рассматривать из как догму.
- **Корректно поставленная задача** в математике — прикладная задача, математическое решение которой существует, единственно и устойчиво. (Задача или алгоритм решения задачи называются вычислительно неустойчивыми, если малые изменения входных данных приводят к заметным изменениям решения).
- Задача машинного обучения изначально **не является корректно поставленной.**
- Успешное решение задачи машинного обучения на 50% состоит из знания «стандартных» методов и умения их применить и на 50% из смекалки, умения правильно поставить задачу, внимательного анализа исходных данных и выделения признаков.
- Никогда не пренебрегайте своими идеями при решении задач. Их стоит как минимум проверить.

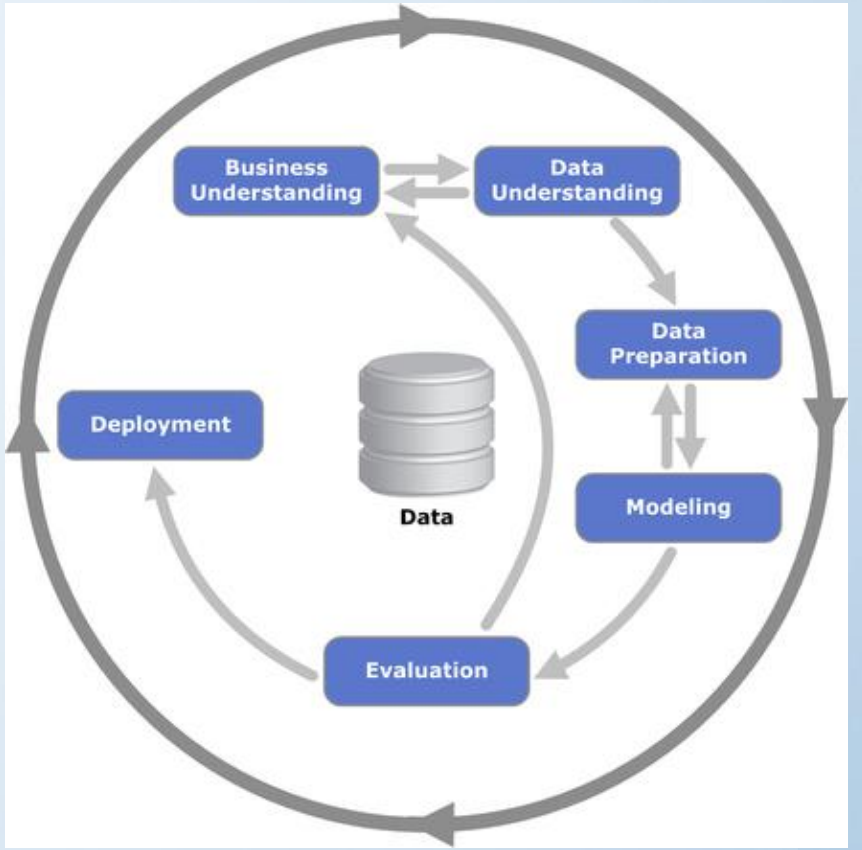
Методологии анализа данных

CRISP-DM

- CRISP-DM (Cross-Industry Standard Process for Data Mining – межотраслевой стандартный процесс для исследования данных) – проверенная в промышленности и наиболее распространённая методология по исследованию данных.

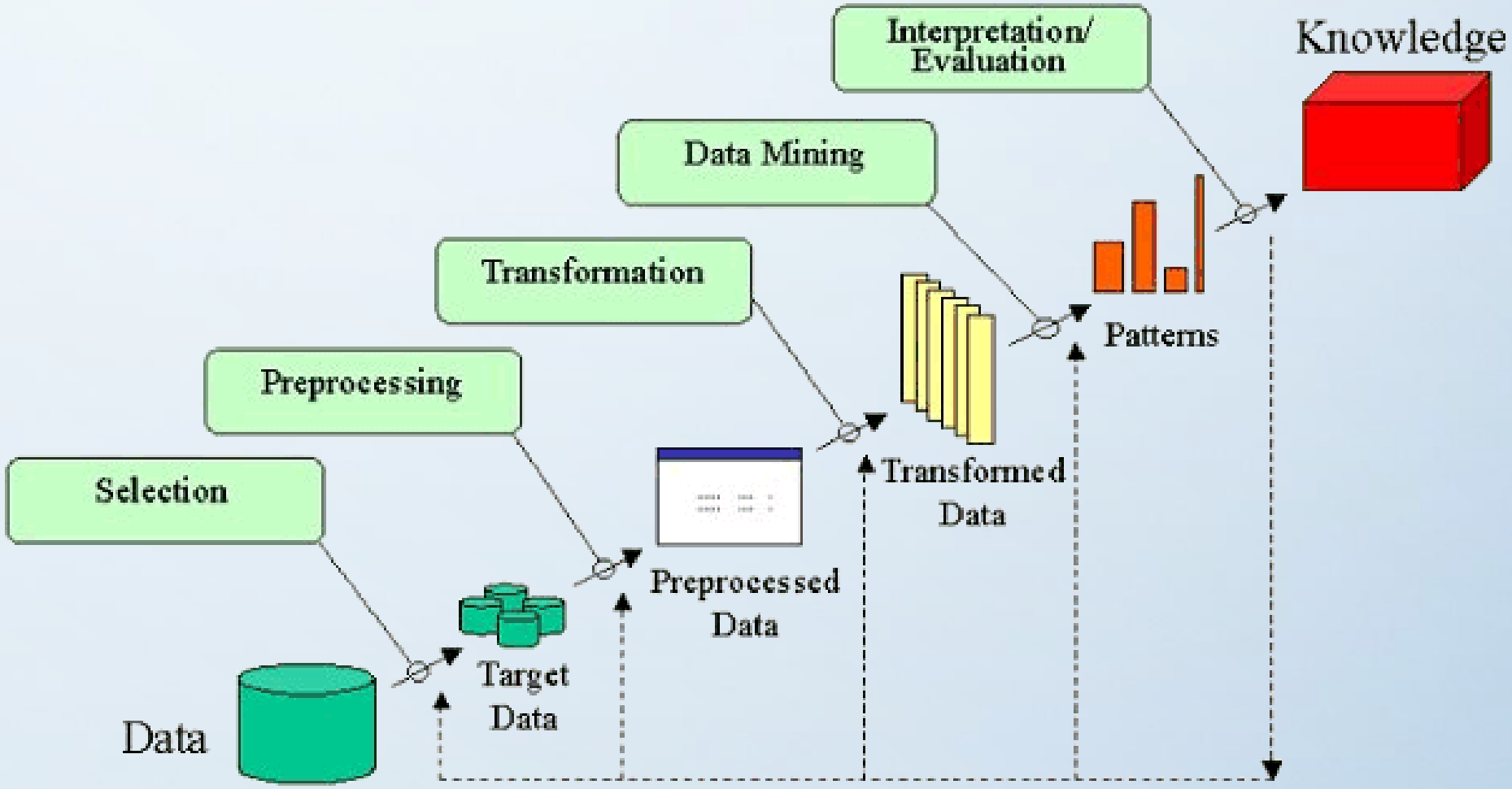
- Модель жизненного цикла исследования данных состоит из шести фаз, а стрелки обозначают наиболее важные и частые зависимости между фазами. Последовательность этих фаз строго не определена. Как правило в большинстве проектов приходится возвращаться к предыдущим этапам, а затем снова двигаться вперед. Описание фаз:

  1. Понимание бизнес-целей (Business Understanding)
  2. Начальное изучение данных (Data Understanding)
  3. Подготовка данных (Data Preparation)
  4. Моделирование (Modeling)
  5. Оценка (Evaluation)
  6. Внедрение (Deployment)



KDD Process

- Методология KDD (Knowledge Discovery in Databases) Process является аналогом CRISP-DM - http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html



Анализ данных и АСОИУ

- На первый взгляд может показаться что анализ данных и «традиционные» информационные системы являются различными подходами. Так ли это?
- Проектирование АСОИУ (существуют различные модели проектирования: каскадная, спиральная):
  1. Определение целей автоматизации, постановка задач.
  2. Изучение предметной области.
  3. Построение модели (схемы) базы данных (с учетом целей автоматизации) – выделение сущностей, связей, атрибутов.
  4. Разработка информационной системы (создание форм, отчетов и т.д.)
  5. Оценка качества разработанной системы (тестирование, проверка работоспособности, моделирование нагрузки).
- Постановка и решение задачи анализа данных:
  1. Понимание бизнес-целей. Определение целей анализа данных.
  2. Начальное изучение данных (первичное изучение набора данных, первичная визуализация данных).
  3. Подготовка данных. Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи:
     - feature extraction – «технический» процесс выделения признаков, например из текстов или изображений.
     - feature engineering – «смысловое» выделение и синтез признаков, которые позволят получить наилучшее качество решения задачи.
     - Кодирование признаков (прежде всего категориальных).
  4. Моделирование. Разработка модели в терминах алгоритмов машинного обучения (применение одного или нескольких алгоритмов).
  5. Оценка. Оценка качества разработанной модели (с помощью методов оценки качества, используемых в машинном обучении).
- При проектировании АСОИУ акцент делается на «накопленные» пользователем бизнес процессы (в каком порядке и какие данные вводятся в формы ввода и сохраняются в БД, какие формируются отчеты и т.д.)
- При решении задачи анализа данных акцент делается на «накопленные» пользователем данные. Как помочь пользователю извлечь пользу из накопленных им данных. Какие нетривиальные зависимости можно найти. Какие решения можно помочь принять. Задачу анализа данных нужно рассматривать как элемент СППР.
- Решение задачи машинного обучения можно рассматривать как частный случай АСОИУ, где мы помогаем пользователю в решении задач, на основе накопленных им данных. Здесь работают как Data Scientist, так и Data Engineer.

CRISP-DM и машинное обучение (анализ датасетов)

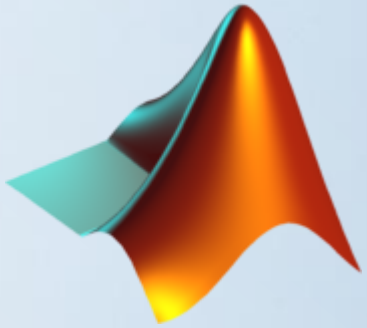
1. Понимание бизнес-целей (Business Understanding) – НЕТ. *Как правило, на этапе решения задачи машинного обучения цель уже задана.*
2. Начальное изучение данных (Data Understanding) – ДА. *Первичное изучение набора данных, первичная визуализация данных.*
3. Подготовка данных (Data Preparation) – ДА. *Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.*
4. Моделирование (Modeling) – ДА. *Разработка модели в терминах алгоритмов машинного обучения.*
5. Оценка (Evaluation) – ДА. *Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.*
6. Внедрение (Deployment) – НЕТ. Оставим эту задачу дата-инженерам.

Языки и инструменты

Языки и инструменты

- Общая особенность всех языков, применяемых для машинного обучения – использование векторизации вычислений. Компиляторы (интерпретаторы) реализуют высокую производительность для векторизованных вычислений.
- Алгоритмы работают не с отдельными ячейками данных, а с многомерными массивами, что увеличивает их производительность.
- В некоторых языках векторизация встроена непосредственно в язык, в некоторых реализована с помощью библиотек. В частности в Python векторизация реализована с помощью библиотеки NumPy.
- В Python также для повышения производительности используют элементы функционального программирования.

MATLAB



- Matrix Laboratory - пакет прикладных программ для решения задач технических вычислений и одноименный язык программирования, используемый в этом пакете. Пакет используют более миллиона инженерных и научных работников, он работает на большинстве современных операционных систем, включая Windows, Linux, Mac OS.
- De facto является пакетом №1 для анализа данных и машинного обучения.
- DSL-язык пакета MATLAB ориентирован на математиков. Не предназначен для разработки полнофункциональных программных систем.
- Пакет является проприетарным и платным.
- Существует свободно-распространяемый аналог - GNU Octave, язык которого в целом совместим с MATLAB, но который содержит меньше библиотек и отличается менее высокой производительностью.

R

- R — язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом.
- Изначально был ориентирован на задачи математической статистики, но в настоящее время содержит большое количество пакетов для анализа данных и машинного обучения.
- DSL-язык, ориентированный на математиков. Не предназначен для разработки полнофункциональных программных систем.

Julia

- Высокоуровневый высокопроизводительный свободный язык программирования с динамической типизацией, созданный для математических вычислений. Эффективен также и для написания программ общего назначения.
- Синтаксис языка схож с синтаксисом других математических языков (например, MATLAB и Octave), однако имеет некоторые существенные отличия. Julia написана на Си, C++ и Scheme.
- В стандартный комплект входит JIT-компилятор на основе LLVM, благодаря чему, по утверждению авторов языка, приложения, полностью написанные на языке, практически не уступают в производительности приложениям, написанным на статически компилируемых языках вроде Си или C++. Большая часть стандартной библиотеки языка написана на нём же.
- Также язык имеет встроенную поддержку большого числа команд для распределенных вычислений.
- Преимущества:
  - Производительность;
  - Ориентирован на параллельные вычисления.
- Недостатки:
  - Активно развивается, но пока находится в экспериментальной фазе.

Python

- Высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. Синтаксис ядра Python минималистичен. В то же время стандартная библиотека включает большой объём полезных функций.
- Python поддерживает несколько парадигм программирования, в том числе структурное, объектно-ориентированное, функциональное, императивное и аспектно-ориентированное.
- Как и C++ поддерживает множественное наследование.
- Большинство библиотек является обертками над библиотеками, написанными на C/C++, что обеспечивает хорошую производительность работы библиотек.
- **Как правило, вызов библиотечной функции (написанной на C/C++) намного производительнее аналогичного кода написанного прикладным программистом на Python (особенно если это ML-алгоритм написанный без использования векторизации).**
- Основное преимущество Python состоит в том, что может использоваться и как язык для обработки данных и как язык для разработки приложений (веб-приложений). Это очень облегчает встраивание ML-решений в веб-приложения (в частности, в бакалаврских работах).

Инструменты

- Основной инструмент – Jupyter notebooks (IPython notebooks). Напоминает одноколоночную таблицу Excel. В ячейках можно писать код, который сразу выполняется:
  - 28 Jupyter Notebook tips, tricks, and shortcuts
  - Для документации используется упрощенный язык разметки Markdown. Краткая справка по использованию в ноутбуках.
- Технология Jupyter notebooks (в различных вариациях) применяется в нескольких дистрибутивах:
  - Conda (оффлайновый дистрибутив, устанавливается локально на компьютер). **Используем Python 3.6 (не 2.7).**
  - https://datalore.io/ – онлайновый сервис от JetBrains (бета-версия).
  - PyCharm (среда разработки для Python) – предназначен в большей степени для веб-разработки (используется в курсе по РИП). Позволяет работать с ноутбуками в среде разработки.

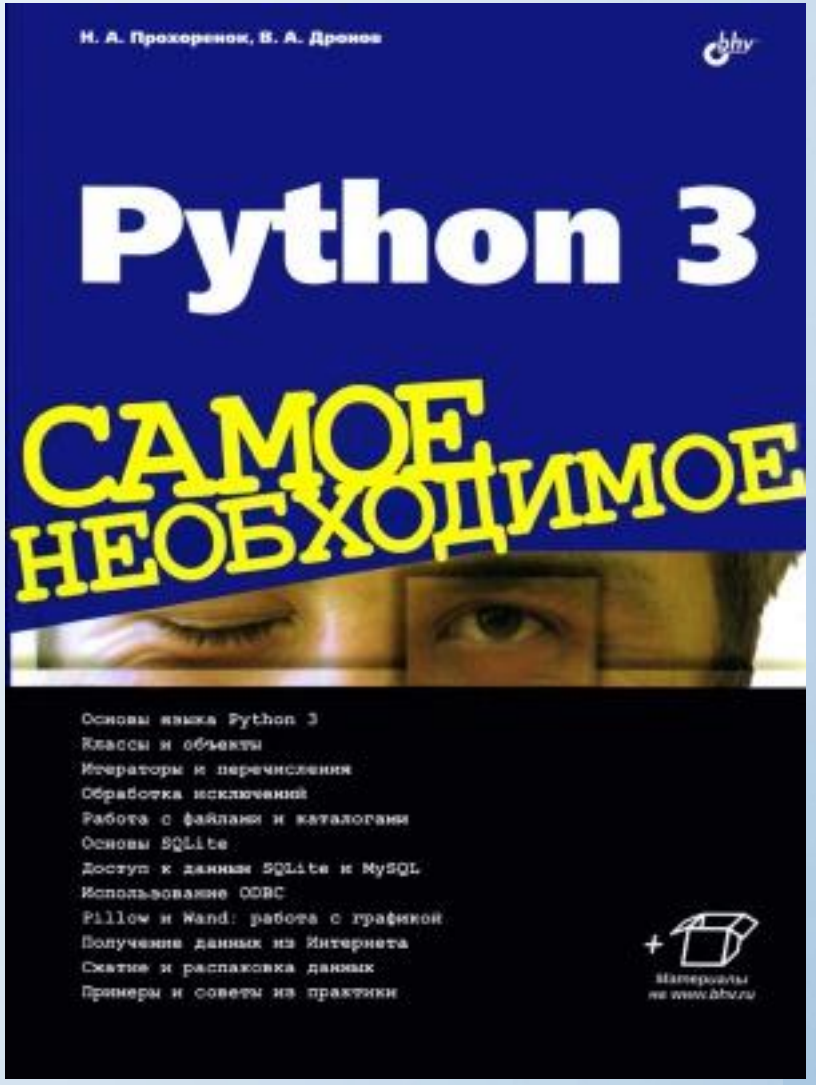
Курсы и книги по машинному обучению

Курсы по машинному обучению

- https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie - краткий курс
- https://www.coursera.org/specializations/machine-learning-data-analysis - специализация из 6 курсов
- Видеолекции курса профессора К.В.Воронцова в ШАД - https://yandexdataschool.ru/edu-process/courses/machine-learning
- Курс лекций профессора К.В.Воронцова - http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%BE
- Курс лекций профессора Н.Ю.Золотых - http://www.uic.unn.ru/~zny/ml/

Книги

- Хорошая вводная книга непосредственно по Python.



Книги

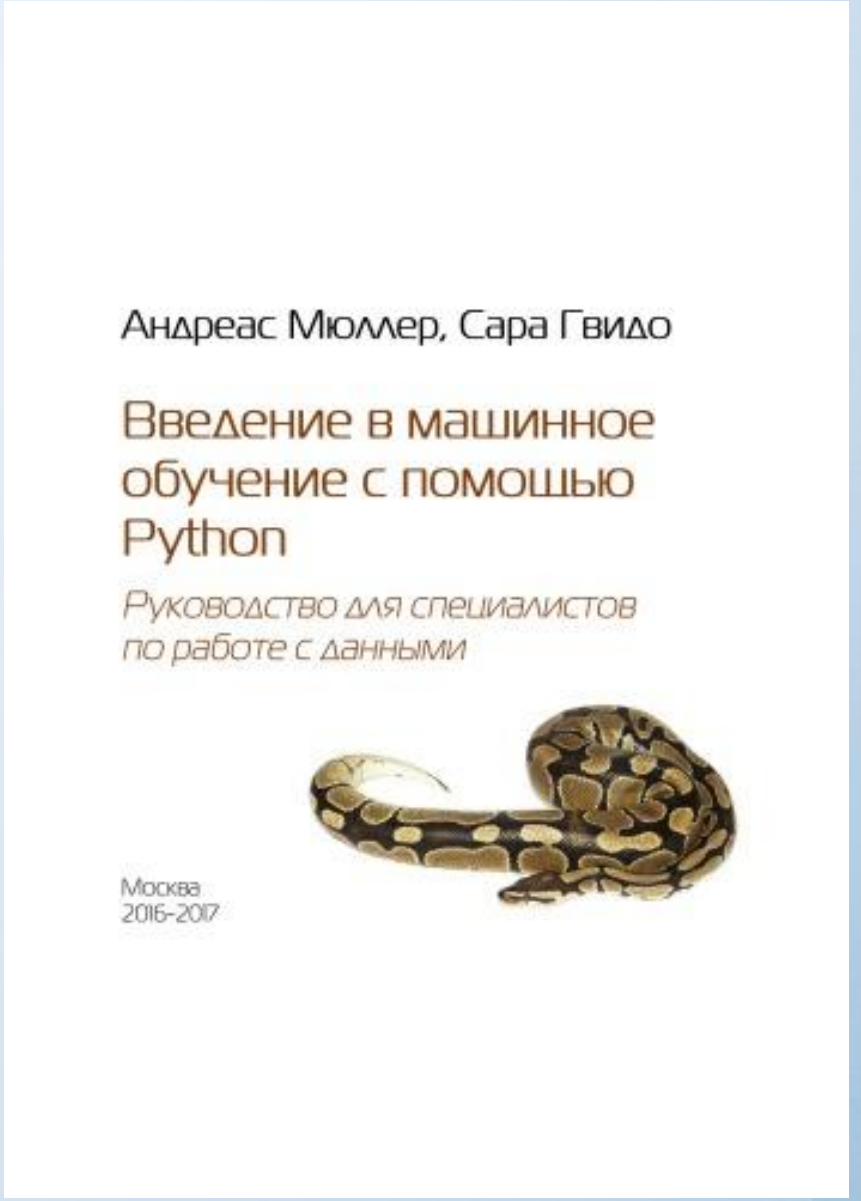
- Хорошая практическая книга с описанием библиотек Python. Меньше примеров задач, больше описания библиотек.

Книги

- Хорошая практическая книга. Исторически была переведена первой.
- Содержит в основном примеры решения задач. В меньшей степени содержит описание библиотек.

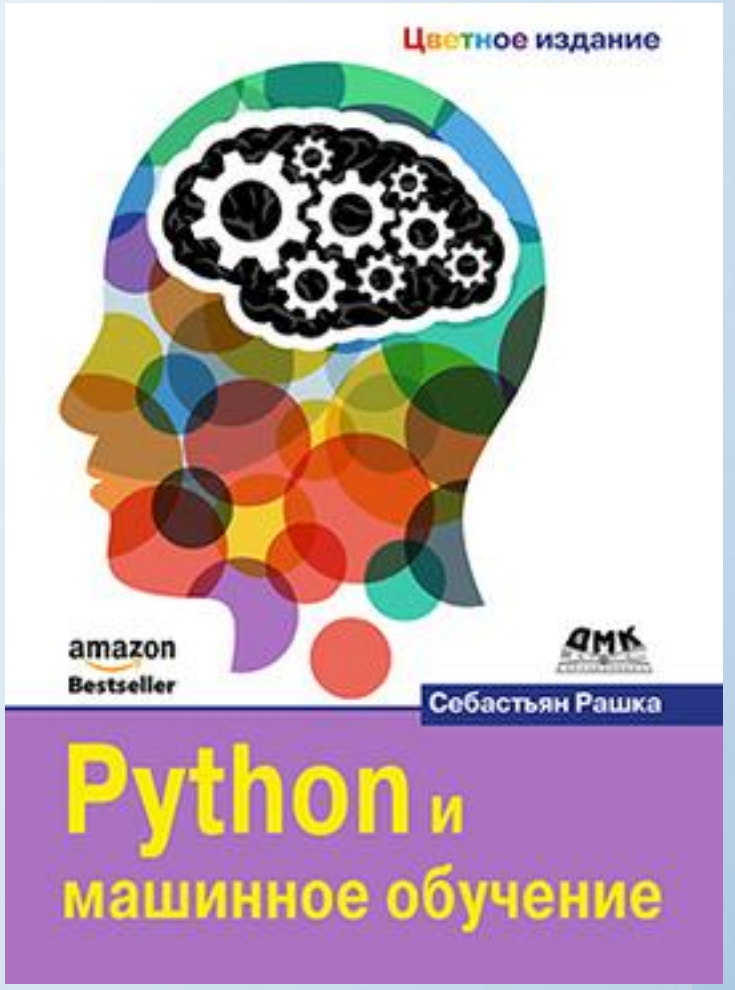
Книги

- Рассматриваются как задачи обучения с учителем, так и задачи обучения без учителя.
- Рассматривается весь жизненный цикл анализа данных – выделение признаков, оценка качества моделей.



Книги

- Рассматривается весь жизненный цикл анализа данных – выделение признаков, оценка качества моделей.
- Рассматриваются много различных задач машинного обучения, в том числе довольно специфических.

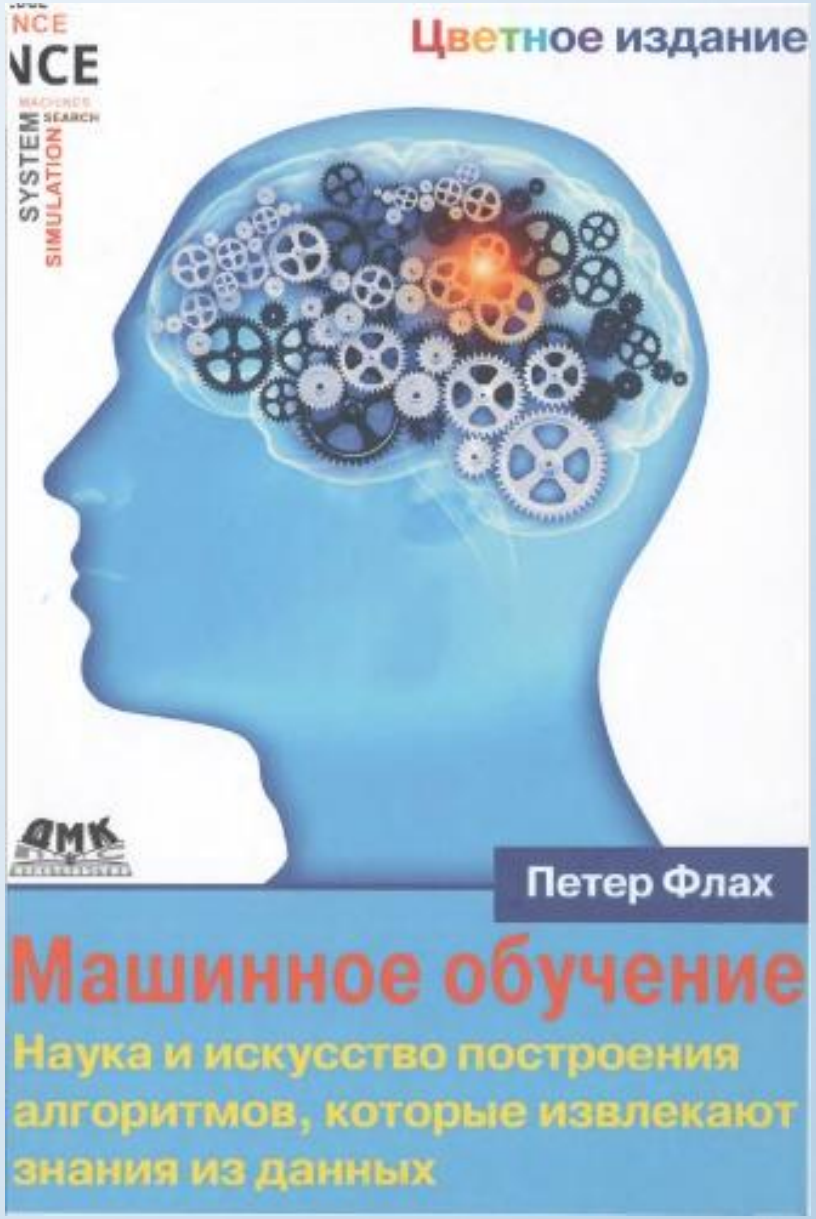


Книги

- Хорошая вводная книга в основном по методам машинного обучения. Методы разбираются достаточно детально.
- Примеров кода относительно немного.

Книги

- Теоретический учебник. Разбираются теоретические основы машинного обучения на основе большого количества примеров.
- Не привязан к конкретному языку программирования.



Классификация и постановки задач машинного обучения

Классификация задач ML

- Обучение с учителем (supervised learning)
  - Классификация
  - Регрессия
  - Прогнозирование временных рядов
- Обучение без учителя (unsupervised learning)
  - Кластеризация
  - Методы понижения размерности
- Обучение с подкреплением (reinforcement learning)

Некоторые типы шкал измерений

- Номинальная шкала (шкала наименований, классификационная шкала), по которой объектам дается некоторый признак (производится классификация объектов по этому признаку). Название «номинальный» объясняется тем, что такой признак дает лишь ничем не связанные имена объектам. Примерами измерений в номинальном типе шкал могут служить номера автомашин, телефонов, коды городов, объектов и т.д.
  - Частный случай – бинарная шкала {0, 1}, {False, True}.
- Шкала называется ранговой (шкала порядка), если множество ее значений состоит из монотонно возрастающих чисел. При этом нет метрики, по которой можно сказать насколько одно значение больше ил Примером шкалы порядка может служить шкала твердости минералов (предложенная в 1811 г. немецким ученым Ф. Моосом), шкала силы ветра, сортности товаров в торговле, различные социологические шкалы и т.д.
- **Количественный (действительный) признак, который является действительным числом. Основной вид шкалы, к которому пытаются свести все остальные.**

CRISP-DM и машинное обучение (анализ датасетов)

1. Понимание бизнес-целей (Business Understanding) – НЕТ. *Как правило, на этапе решения задачи машинного обучения цель уже задана.*
2. Начальное изучение данных (Data Understanding) – ДА. *Первичное изучение набора данных, первичная визуализация данных.*
3. **Подготовка данных (Data Preparation) – ДА. *Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.***
4. Моделирование (Modeling) – ДА. *Разработка модели в терминах алгоритмов машинного обучения.*
5. Оценка (Evaluation) – ДА. *Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.*
6. Внедрение (Deployment) – НЕТ. Оставим эту задачу дата-инженерам.

Данные и признаки

Табличные данные (объекты-атрибуты)

Атрибуты (свойства, поля данных)

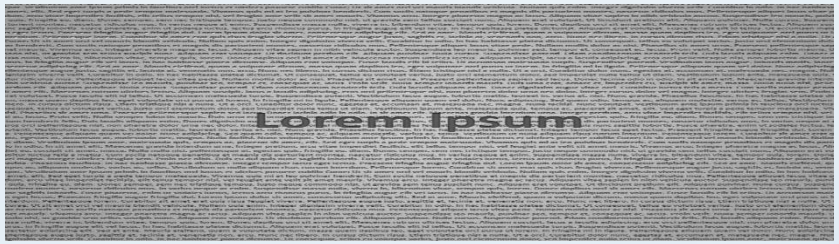
Город	Год рождения	Доход	Пол	
Москва	1990	100,00	Ж	*Объекты*
Курск	1975	85,3	М	
Москва	1983	40,5	Ж	
Брянск	1960	90,5	М	
номинальная шкала	шкала порядка	действительный признак	бинарная шкала	

Признаки

Город	Год рождения	Доход	Пол	
1	1990	100,00	0	*Объекты*
2	1975	85,3	1	
1	1983	40,5	0	
3	1960	90,5	1	

- Матрица объекты-признаки (feature data)
- Эту матрицу традиционно обозначают буквой X.

Текстовые данные (тексты-слова)



Изображения (изображения-пиксели)

CRISP-DM и машинное обучение (анализ датасетов)

1. Понимание бизнес-целей (Business Understanding) – НЕТ. *Как правило, на этапе решения задачи машинного обучения цель уже задана.*
2. Начальное изучение данных (Data Understanding) – ДА. *Первичное изучение набора данных, первичная визуализация данных.*
3. Подготовка данных (Data Preparation) – ДА. *Очистка данных, удаление аномалий. Выделение из исходных данных признаков (features) для решения задачи.*
4. **Моделирование (Modeling) – ДА. *Разработка модели в терминах алгоритмов машинного обучения.***
5. **Оценка (Evaluation) – ДА. *Оценка качества разработанной модели с помощью методов оценки качества, используемых в машинном обучении.***
6. Внедрение (Deployment) – НЕТ. Оставим эту задачу дата-инженерам.

Обучение с учителем (на примере регрессии)

- Каждой строке матрицы X ставится в соответствие значение столбца ответов Y. Y-действительный признак.

Признаки (X)					*Ответы (Y)*
Город	**Год рождения**	**Доход**	**Пол**		**Доход в будущем периоде**
1	1990	100,00	0	*Объекты*	120,05
2	1975	85,3	1		87,30
1	1983	40,5	0		55,20
3	1960	90,5	1		87,40
обучающая выборка					
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ
тестовая выборка					

- Ответы на тестовой выборке могут быть известны, но аналитику данных их не дают, заказчик может использовать их для итогового тестирования.
- **Признаки на обучающей и тестовой выборке должны быть одинаково закодированы.**
- Обучение с учителем происходит в две фазы:

1. Собственно обучение. $M=Alg.fit(X_{обуч}, Y_{обуч}, H)$. Используемый нами алгоритм Alg строит модель соответствия M между $X_{обуч}$ и $Y_{обуч}$ с учетом гиперпараметров алгоритма H.
2. Предсказание. $Y_{тест} = Alg.predict(M, X_{тест})$.

- Гиперпараметры алгоритма – параметры, значение которых задается до начала обучения (значение остальных параметров настраивается в процессе обучения). У каждого алгоритма гиперпараметры свои, для их правильной настройки используются специальные методы, в частности перебор по сетке (grid search).
- Модель соответствия M можно рассматривать как функцию f: Y=f(X). Но в более общем виде стоит рассматривать M как морфизм из теории категорий (введение в теорию категорий).

Оценка качества (на примере регрессии)

- Идея всех методов оценки качества состоит в том, чтобы понять насколько велика ошибка предсказания алгоритма, насколько хорошо или плохо он предсказывает. Разница только в используемых метриках.
- M=Alg.fit($X_{обуч}$, $Y_{обуч}$, H). $\hat{Y}_{обуч}$ = Alg.predict(M, $X_{обуч}$). $\hat{Y}_{обуч}$ – результат работы алгоритма на обучающей выборке.
- При оценке качества стараются учесть возможное переобучение модели.
- Наиболее простая метрика – среднеквадратичная ошибка:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Классификация

- Бинарная классификация (Y-значение по бинарной шкале)

Признаки (X)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		Переедет в другой город?
1	1990	100,00	0	*Объекты*	1 (Да)
2	1975	85,3	1		0 (Нет)
1	1983	40,5	0		0 (Нет)
3	1960	90,5	1		1 (Да)
обучающая выборка					
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ
тестовая выборка					

- Классификация (Y-значение по номинальной шкале)

Признаки (X)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		В какой город переедет?
1	1990	100,00	0	*Объекты*	2
2	1975	85,3	1		2
1	1983	40,5	0		1
3	1960	90,5	1		1
обучающая выборка					
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ
тестовая выборка					

- Многоклассовая классификация (Y-множество значений по номинальной шкале)

Признаки (X)					*Ответы (Y)*
Город	Год рождения	Доход	Пол		В какой город переедет?
1	1990	100,00	0	*Объекты*	2, 3
2	1975	85,3	1		2
1	1983	40,5	0		1
3	1960	90,5	1		1, 3
обучающая выборка					
2	1965	97,5	1		НУЖНО ПРЕДСКАЗАТЬ
тестовая выборка					

- Пример метрики качества – точность (accuracy) – доля правильно предсказанных меток классов.

Обучение без учителя (на примере кластеризации)

- Обучающей выборки нет.
- Для каждой строки матрицы X алгоритм пытается предсказать значение метки (номера) кластера Y.
- Y=Alg.fit_predict(X, H). Используется алгоритм Alg с набором гиперпараметров H.
- Метрики оценки качества базируются на оценке расстояний между получившимися кластерами.
- Одним из наиболее сложных и интересных методов обучения без учителя являются самоорганизующиеся карты Кохонена.

Обучение с подкреплением

- Обучение с обратной связью, с опосредованным учителем.
- Алгоритм обучается, взаимодействуя с некоторой средой. Откликом среды являются сигналы подкрепления, поэтому такое обучение является частным случаем обучения с учителем, но учителем является среда или её модель.
- Частным случаем обучения с подкреплением является Q-обучение.