Машинное обучение

Лекция 1. Введение

Автор: Рустам Азимов

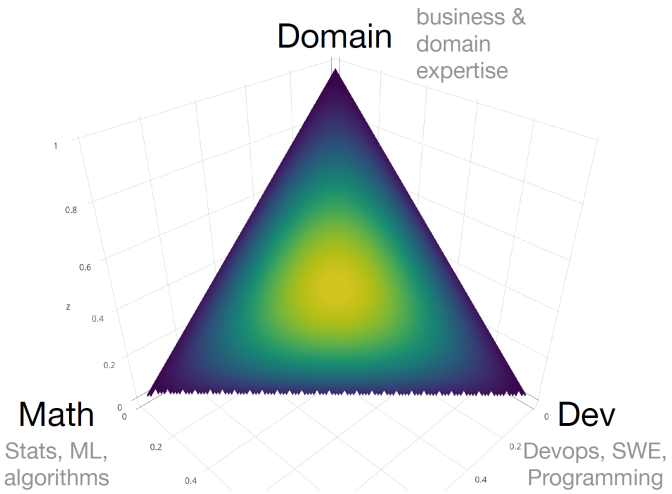
Санкт-Петербургский государственный университет

Санкт-Петербург

Где применяется ML?

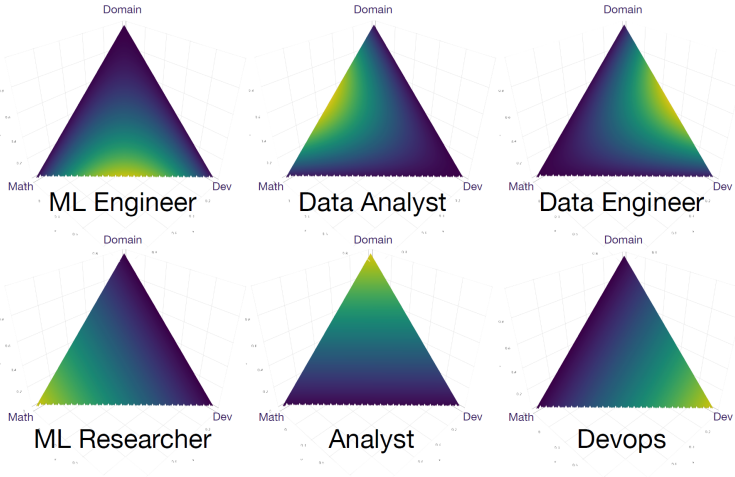
- Рекомендательные системы (соцсети, интернет-магазины, стриминговые сервисы и т.д.)
- Поисковые системы (ранжирование поисковой выдачи)
- Прогнозирование спроса на товары, поведения покупателей
- NLP
- Распознавание речи
- CV
- В общем, машинное обучение применяется для **поиска зависимостей** в данных

ML и Data Science



*Взято из выступления Алексея Натёкина на встрече сообщества ODS

ML и Data Science



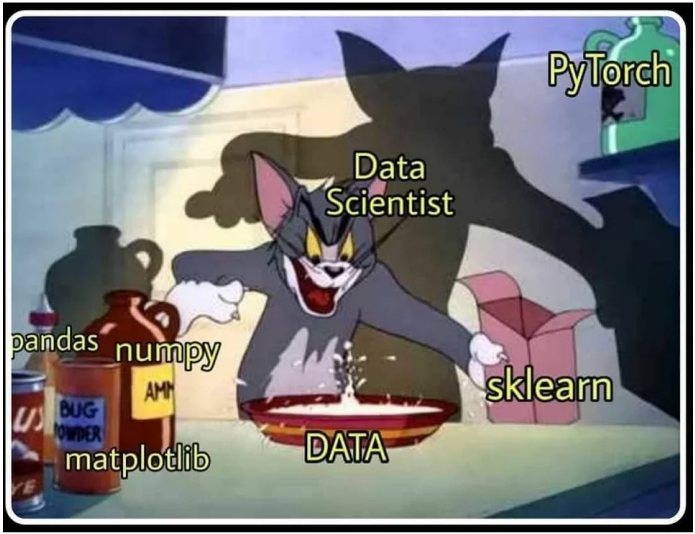
*Взято из выступления Алексея Натёкина на встрече сообщества ODS

DS Lifecycle

Входные знания

- Математическая статистика и теория вероятностей
- Линейная алгебра, некоторые идеи из геометрии
- Язык программирования и необходимые инструменты (Python, Jupyter notebooks)

Используемые библиотеки



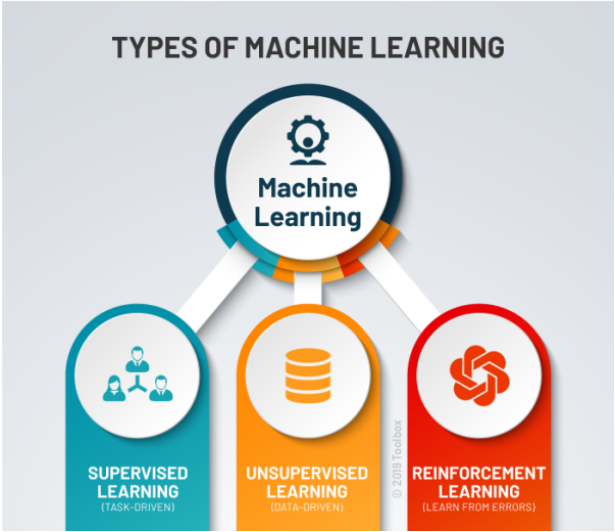
Цели ML

- Найденные зависимости хочется использовать с какими-то целями
  - **Краткосрочные** — автоматизации каких-то процессов, когда результат надо получить быстро
  - **Долгосрочные** — проанализировать действия в некотором процессе на долгосрочный период
- ML для краткосрочных и долгосрочных целей очень отличается
- Зависимости, которые находятся, могут быть в очень разном виде
- Для долгосрочных целей применяются методы с моделями, которые можно расшифровать человеку

Идея ML

- Почти у всех методов ML есть общая идея: модель должна описывать некоторые параметры реально происходящего процесса
- Мы не знаем, как процесс устроен, поэтому построить модель с помощью теории не можем
- Поэтому делаем трюк: запихиваем в модель множество свободных параметров, а потом пытаемся подбирать, чтобы результат совпадал с ожидаемым
- Так работает почти любой ML

Виды ML


TYPES OF MACHINE LEARNING
Machine Learning
SUPERVISED LEARNING
(TASK-DRIVEN)
UNSUPERVISED LEARNING
(DATA-DRIVEN)
REINFORCEMENT LEARNING
(LEARN FROM ERRORS)
© 2019 Toolbox

Обучение с учителем

- Между входными и выходными данными может существовать некоторая зависимость, но она неизвестна
- Известна только конечная совокупность прецедентов (обучающая выборка)
- Прецедент представляет собой пару «объект, ответ»
- Необходимо построить алгоритм, способный для любого нового объекта выдать достаточно точный ответ
- Примеры: задача классификации (classification), задача регрессии (regression)

Обучение без учителя

- Ответы не задаются и требуется искать зависимости между объектами
- Примеры: задача класстеризации (cluster analysis), уменьшение размерности (dimensionality reduction)

Обучение с подкреплением

- Обучение методом проб и ошибок
- Удачные действия поощряются («подкрепляются») средой
- Пример: учимся играть в шахматы

Как подходить к ML задачам?

1. **Цель**: Зачем вообще это делать?
2. **Литература**: Скорее всего проблему, которую вы хотите решить, уже решали
3. **Сбор данных**: Если на этапе сбора имеет смысл пытаться собирать максимальное количество данных, то не нужно пытаться строить модель сразу на всех данных. Начинаем с простых срезов, чтобы не получить белый шум. А потом постепенно расширяем объемы данных и количество признаков

Как подходить к ML задачам?

4. **Построение модели**: Начинаем от простого к сложному, не стоит сразу начинать со сложных моделей (нейросетей и т.д.). Когда выявлены какие-то зависимости, уже можно строить сложные модели
5. **Оценка качества**: Для ML проектов всегда критичный аспект. Имея две модели, нужно уметь оценить, какая из них лучше/хуже
6. **Внедрение**: Для долгосрочных целей — внедрить, значит перейти к конкретным решениям. Для краткосрочных целей — встраивание модели в продукт для автоматизации

Представление данных

- Данные удобно представлять в виде таблиц
- Количество объектов n — **объём выборки**
- У каждого из объектов имеется m **признаков** (**фич**)
- **Фича-инжиниринг** — передача знаний машине от человека, построение признаков с помощью интеллекта (думать надо)
- **Фича-анализ** — анализ результатов, выкидывание фич, которые мало повлияли на результат

Объект	Признак 1	Признак 2	...	Признак m
X_1	$P_{1,1}$	$P_{1,2}$	...	$P_{1,m}$
...	...	...	...	...
X_n	$P_{n,1}$	$P_{n,2}$	...	$P_{n,m}$

Типы признаков

- **Количественные** (числовые) признаки, у которых область значений — вещественные числа (сам признак имеет числовую природу)
- **Порядковые** признаки задают порядок на объектах
- **Номинальные** (категориальные) признаки не имеют числовой природы и (как правило) число их возможных значений конечно
  - В частности, **бинарные** признаки — это номинальные признаки с двумя возможными значениями

Пример

- Средний балл — это вещественное число (количественный признак)
- Пол — это бинарный признак
- Место в рейтинге — это порядковый признак

Студент	Пол	Средний балл	Место в рейтинге
Иванов	1	4.5	2
Сидорова	0	5.0	1
Серов	1	3.5	3

Анализ признаков

- Берем соотвествующий признаку столбец в таблице, получаем вектор $P = (p_1, \ldots, p_n)$
- Находим важные нам характеристики полученной выборки
- **max, min**
- **Среднее** $mean = p = (\sum_{i=1}^{n} p_i)/n$
- **Медиана** — такое число, что ровно половина из элементов p_i больше него, а другая половина меньше него
- Не путать медиану со средним: для вектора $(0, 1, 2, 18, 19)$ среднее значение 8, а медиана 2

Медиана и среднее

- Значение медианы не так сильно (как среднее) зависит от попадания в выборку аномально больших и аномально малых значений признака
- Если медиана и среднее близки друг к другу, то выборка называется **симметричной**

Мода

- **Мода** — значение, которое встречается наиболее часто в выборке
- Например, модой для вектора $(1, 0, 5, 1, 0, 3, 0, 0)$ является 0
- Мода не всегда определена однозначно
- Мода (в отличие от среднего и от медианы) имеет смысл и для номинальных признаков

Среднеквадратическое отклонение

- Среднее и медиана могут быть одинаковыми у совершенно разных выборок
- Например, выборки $(0, 0, 0, 0, 0)$ и $(-2, -1, 0, 1, 2)$
- Поэтому для адекватного описания выборки необходимо определить разброс значений
- Для этого считают среднеквадратическое отклонение

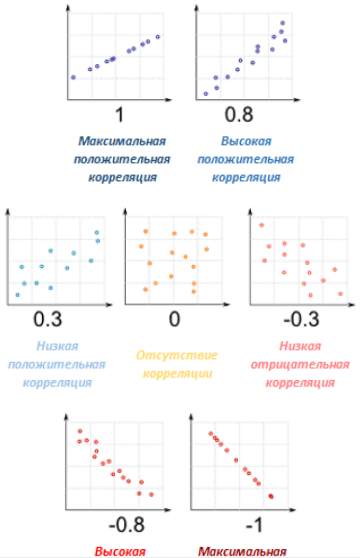
$$s_p = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^2}$$

Коэффициент корреляции

- Коэффициент корреляции показывает как значения одного признака определяют значения другого признака
- Сильная зависимость между годом поступления и годом рождения
- Слабая зависимость между этими признаками и полом

Студент	Пол	Год поступления	Год рождения
Иванов	1	2020	2002
Сидорова	0	2019	2001
Серов	1	2018	2000

Коэффициент корреляции



Коэффициент корреляции

- Пусть $P = (p_1, \ldots, p_n)$ и $Q = (q_1, \ldots, q_n)$ — интересующие нас признаки
- Тогда коэффициент корреляции считается по формуле:

$$r(P, Q) = \frac{(\sum_{i=1}^{n} p_i q_i) - n\bar{p}\bar{q}}{(n-1)s_p s_q}$$

- $r(P, Q)$ — число от -1 до 1
- Если $r(P, Q) = 0$, то очевидной зависимости нету
- Если $r(P, Q) = 1$, то между признаками P и Q существует линейная зависимость

Дополнительные источники

- machinelearning.ru
- mlcourse.ai, dlcourse.ai
- kaggle
- T. Hastie, R. Tibshirani, J. Friedman "The elements of Statistical Learning"
- T. Mitchell "Machine Learning"
- Труды конференций ICML, NIPS, CIKM, KDD, и т.д.
- Журналы JML, JMLR, JIS, NC и т.д.