Лекция 1. Введение

Цель курса

- Уметь сформулировать задачу в терминах ML
- Понимать, какие задачи можно решать и что для этого нужно
- Уметь находить подходящий класс решающих алгоритмов по формулировке задачи
- Ориентироваться в области и знать «где посмотреть» существующие решения
- Понимать границы применимости различных методов

Содержание курса

1. Введение
   - Основные используемые понятия и определения
   - Признаки и работа с ними
   - Обучение и оценка качества
   - Функции потерь и метрики качества
2. Модели
   - Линейные модели
   - Метрические методы
   - Решающие деревья
   - Ансамбли моделей
   - Нейронные сети
3. Задачи
   - Кластеризация и понижение размерности
   - Анализ временных рядов
   - Задачи обработки изображений, звука и текста
   - Обучение с подкреплением

Выставление оценок

Дифференцированный зачет:

1. 2-3 домашние работы
2. 5-7 заданий на практике
3. Аппеляция в форме экзамена-беседы (можно изменить оценку в обе стороны)

Материалы

- https://education.yandex.ru/handbook/ml
- http://www.machinelearning.ru
- Stepik, Coursera, et
- R. Tibshirani, J. Friedman «Introduction to Statistical Learning»
- T. Hastie, R. Tibshirani, J. Friedman «The elements of Statistical Learning»
- etc

Машинное обучение: определения

1. Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions [wikipedia].
2. Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач [русская wikipedia].
3. Машинное обучение — это наука, изучающая алгоритмы, автоматически улучшающиеся благодаря опыту [yandex ml handbook].
4. 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 [Tom M. Mitchell].

История области

50–70 гг. Базы знаний, логические системы, распознавание образов, нейронные сети.

70–80 гг. ID3 деревья, разумные практические результаты, VC-оценки, экспертные системы.

80–90 гг. Первые конференции, много практического применения, активное применение кластеризации в анализе.

90–00 гг. Повторное сэмплирование в ML, bagging, boosting, SVM, LASSO, применение в информационном поиске, размежевание ML и DM.

00–10 гг. Расцвет деревьев и ансамблей. Внедрение ML в интернет-поиск, рекламу, рекомендации.

10–20 гг. Deep Learning, Convolutional, Recurrent, GAN.

20–. . . гг. RL, диффузионные модели, трансформеры и LLM.

Индустрия машинного обучения

Романтика кончилась, начался конвеер.

- Data Engineer
- Аналитик данных
- Data scientist
- ML-инженер
- MLOps
- etc

Задача машинного обучения по прецендентам

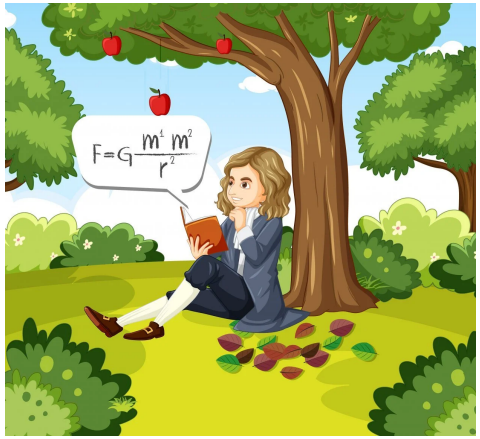
Задано множество *объектов* X, множество *допустимых ответов* Y, и существует *целевая функция* (target function) $y^* : X \to Y$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \ldots, x_\ell\} \subset X$. Пары «объект–ответ» (x_i, y_i) называются *прецедентами*. Совокупность пар $X^\ell = \{(x_i, y_i)\}_{i=1}^\ell$ называется *обучающей выборкой* (training sample).

Задача *обучения по прецедентам* заключается в том, чтобы по выборке X^ℓ *восстановить зависимость* y^*, то есть построить *решающую функцию* (decision function) $a : X \to Y$, которая приближала бы целевую функцию $y^*(x)$, причём не только на объектах обучающей выборки, но и на всём множестве X.

Решающая функция a должна допускать эффективную компьютерную реализацию; по этой причине будем называть её *алгоритмом*.

К.В. Воронцов. Вычислительные методы обучения по прецедентам.

Пример



- Физика:
  Исходные данные $\rightarrow$ гипотеза $\rightarrow$ математическая формулировка $\rightarrow$ проверка
- ML:
  Исходные данные $\rightarrow$ алгоритм $\rightarrow$ модель $\rightarrow$ проверка

Составные части задачи

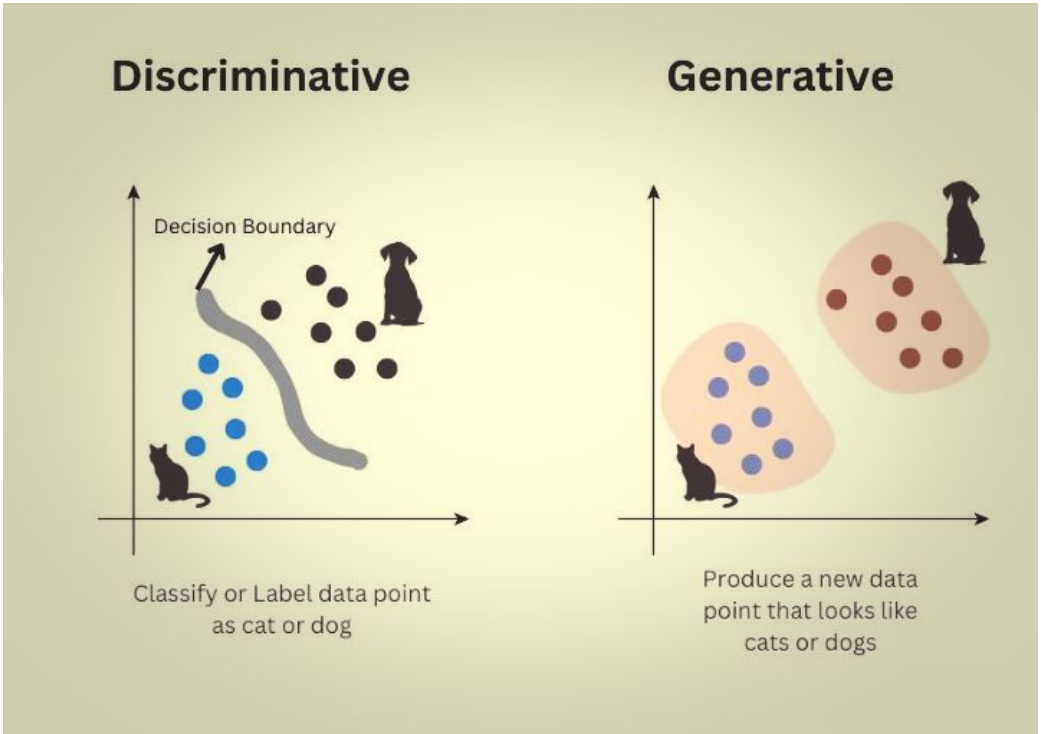
- Данные
- Модель (решающая функция)
- Функция потерь
- Оценка качества

Подходы к решению

- Статистический (байесовский)
- Оптимизационный (минимизация эмпирического риска)

Классы моделей

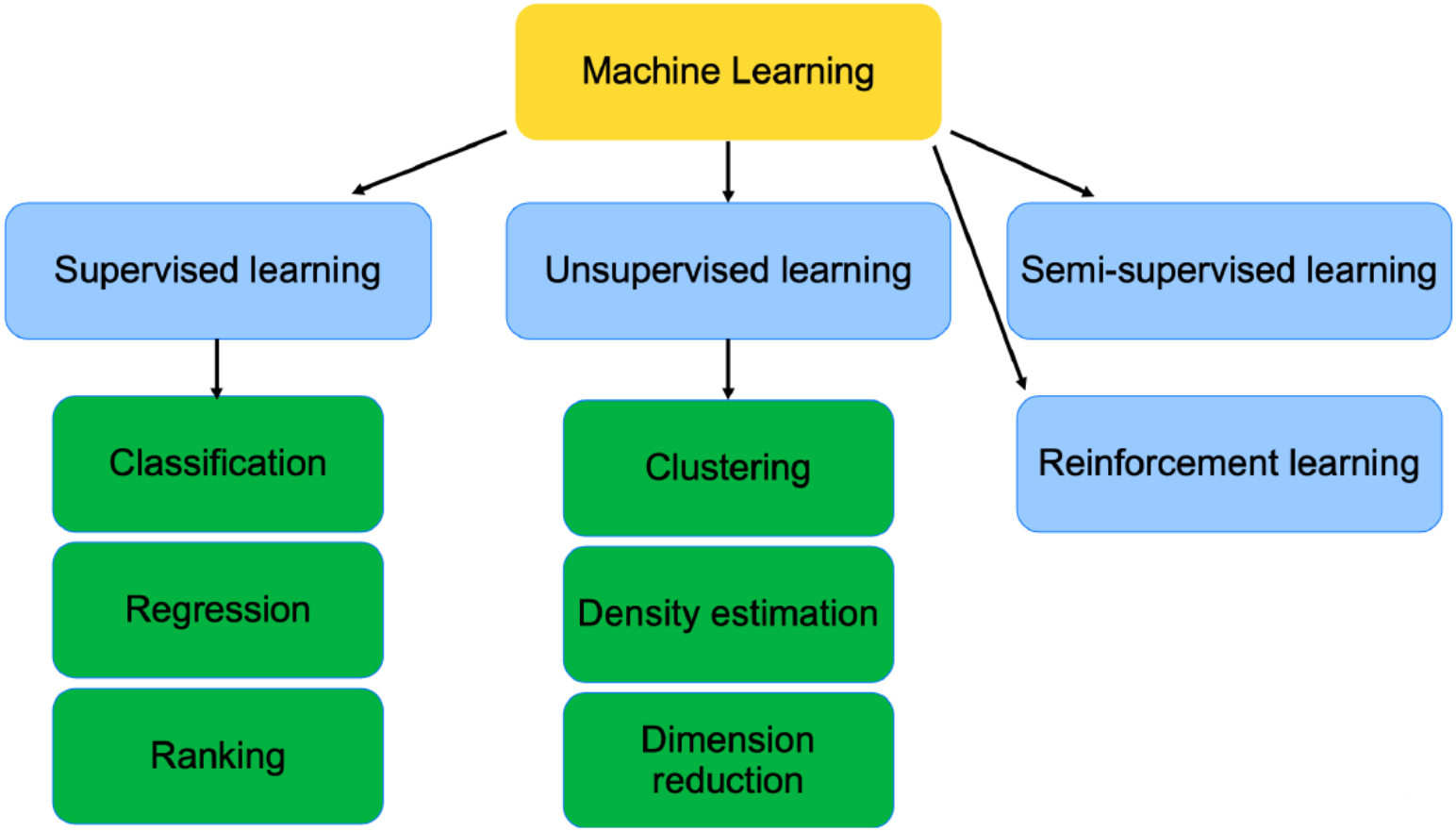
1. Дискриминативные – оценка $P(Y|X)$
2. Генеративные – оценка $P(X, Y) = P(X|Y)P(Y)$



Классификация по способу получения опыта

- Transductive learning
- Обычное обучение
- Активное обучение (active learning)
- Обучение с бюджетом (budget learning)
- Интерактивное обучение (online learning)
- Многорукие бандиты (multi-armed bandits)
- Обучение с подкреплением (reinforcement learning)

Классификация по цели обучения



ML для решения «классических» задач

- Оптимизация и планирование
- Численные методы и симуляции
- Компиляторы и оптимизация кода
- Компрессия данных
- etc