Машинное обучение

Лекция 1. Введение

Цель

- 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

Материалы

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

Машинное обучение: определения

- 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)**.
- Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач **(wikipedia)**.
- Машинное обучение — это наука, изучающая алгоритмы, автоматически улучшающиеся благодаря опыту (**yandex ml handbook**).
- 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

- 50-70гг — базы знаний, полнотекстовый поиск, распознавание образов, нейронные сети
- 70-80гг —ID3 деревья, разумные практические результаты, VC-оценки
- 80-90гг — первые конференции, много практического применения, активное применение кластеризации в анализе
- 90-00гг — повторное сэмплирование в ML, SVM, применение в IR, ML != DM, LASSO, bootstrap, bagging, boosting
- 00-10гг — Compressed sensing и прочие восстановления сигналов, царство деревьев, развитие ансамблей, . . .
- 10-20гг — Deep Learning, Convolutional, Recurrent, GAN
- 20-… — RL, диффузионные модели, LLM.

Индустрия машинного обучения

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

Постановка задачи МО

Задано множество *объектов* X, множество *допустимых ответов* Y, и существует *целевая функция* (target function) $y^*\colon X \to Y$, 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) $a\colon X \to Y$, 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*.

Пример

Здесь должен быть пример...

Составные части задачи

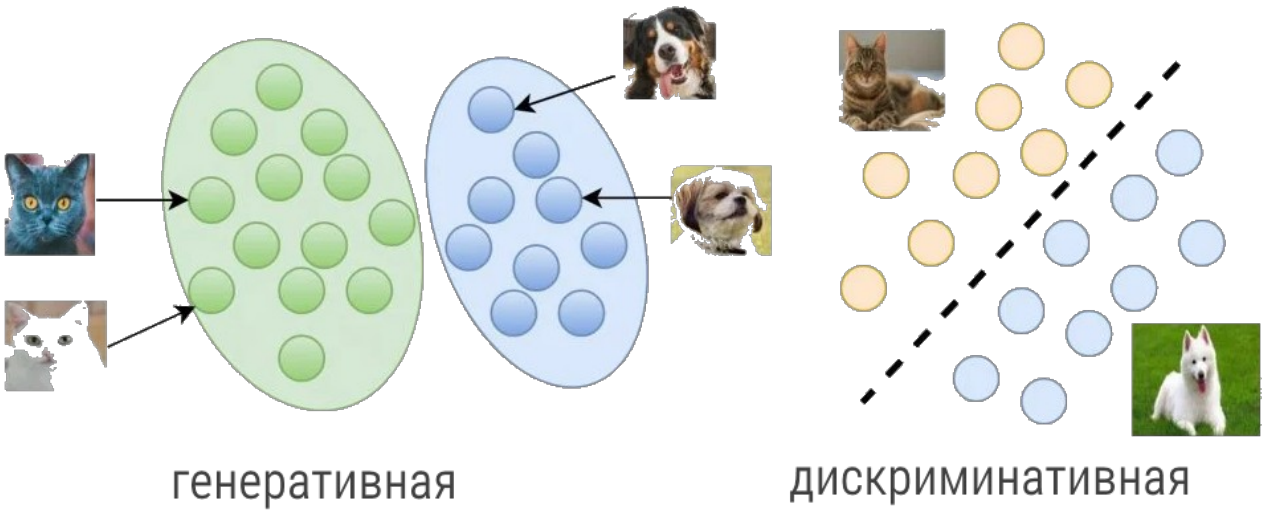
- Данные
- Модель (решающая функция)
- Целевая функция
- Оценка качества

Подходы к решению

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

Классы моделей

- Дискриминативные – оценка P(Y | X)
- 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