

+

×

-

÷

Лекция №1

Александр

Дваконов

Форма контроля

Осень

- 5 практические задания
- 5 тестов
- Бонусные баллы

Весна

- $\max = 239$ весна
- $\max = 514\delta$ за 209

Практические

- 30 - 45δ
- 10 - 14 задач
- просрочил \Rightarrow штраф 40%

Тест

- 16 δ
- 3 дня
- один раз

Соревнование

- до 50 δ
- при достижении позиции - баллы

Термины ML

Наука о данных (Data Science)

- представление, извлечение, сбор, обработка, хранение, анализ и использование данных

Анализ данных (Data Mining)

- находжение закономерностей и моделей, которые

- Валидны

Э в действительности

- полезны
- нетривиальны
- понятны/интерпретируемые

их можно объяснить

Математическая статистика

- математические методы систематизации

Машинное обучение (Machine Learning)

- извлечение машиной, которое ранее не было явно запрограммировано

Задача состоит из

- Задания

- меры ошибок

Пример

Задача: распознавать символы

Мера: процент правильного распознавания

Большие данные (Big data)

— сбор, хранение, обработка и анализ данных огромных объемов

Причины:

- упрощение средств хранения
- ускорение средств обработки
- успехи в некоторых разделах ML (DL)
- интерес бизнеса

Искусственный интеллект (Artificial Intelligence)

— наука и технология создания интеллектуальных машин

Проблемы:

- AI в слабом смысле
(Слишком однородивший)
- AI в слишком сильном смысле
(Термитатор)
- Создание
 - Самондектификация
 - Идентификация других
 - Борьба за ресурсы

Основной тип задач

model based reasoning

- можно записать упр-ие

case based reasoning

- на основе прецедентов известка выборка

Лекция №2

Постановка основных задач

Обучение с учителем (Supervised Learning) (с размечанными данными)

$$X_{\text{train}} = \{(x_1, y_1), \dots, (x_m, y_m)\}$$

$y: X \rightarrow Y$ - целевая функция

$$y(x_i) = y_i$$

задача экстраполяции

Y - значение целевого признака

X - пространство объектов

- 1) Восстановить зависимость
- 2) Интерпретация
- 3) Оценка качества

Types supervised learning

Классификация

$$|Y| = k < \infty$$

Бинарная

$$Y = \{0, 1\} \text{ или } Y = \{1, -1\}$$

Скориковая

Бинарная

$$L(x) \in \{0; 1\}$$

Регрессия

$$Y = \mathbb{R}$$

Многомерная

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

Прогнозирования

$$X_{\text{train}} = \{(x_1, t_1), (x_2, t_2), \dots, (x_m, t_m), y_m\}$$

$$t_1 \leq t_2 \leq \dots \leq t_m$$

ка к непересекающимся классам

$$Y = \{1, 2, \dots, k\}$$

ка к пересекающихся классам

$$Y = \{0, 1\}^k$$

Ракурсные

Y-тум

Восстановление за висимости

- Строим гипотезу $\lambda(x)$

$L(y, a)$ — функция ошибок

Ошибки на x

На практике замерка мерой вероятностной ошибки на эмпирическую.

Минимизируем ошибку

Модель — параметрическое семейство ошибок

Однократное — определение параметров

Однократная способность

Кол-во ошибок на train vs на test

Алгоритм:

- Эффективно реализуется на компьютере
- широкий набор ресурсов.

Требования:

- качество
- эффективность | т. обусловлено и использ. т.в.
- надежность | устойчивость к шуму
- масштабируемость | ↑ объема ⇔ ↑ т.
- интеграторируемость | обобщенные раз-ов
- компактность | затраты на хранение

Система решения задачи

1) Постановка задачи

покомандные задачи

2) Сбор + подготовка

Выбор составляющих:

- алгоритм
- контроль
- признаков

3) обуздание

4) проверка качества

5) прошивка леккость

6) отчетность

Обучение без учителя

с частично разметанными данными

- показать структуру пространства объектов X

Как распределены объекты?

Можно ли разделить на подпр-ва?

Обучение с частично разметанными данными

Semi-supervised learning

$$X_{train} = \{(x_1, y_1), \dots, (x_k, y_k), x_{k+1}, \dots, x_m\}$$

Если заранее известна контролируемая выборка x'_1, \dots, x'_q , то это трансдуктивное обучение.

Призвищированное обучение

с частично размет-ми данными \oplus

- есть дополнительные признаки

Лекция №3

Метрические методы

Основываются на расстояниях между объектами

Метод ближайшего центрида

центрид:

$$c_j = \frac{1}{|\{i : y_i = j\}|} \sum_{i:y_i=j} x_i$$

- класс тот, к центриду которого ближе.
 - + определяется только центридом
 - + простой алгоритм
- можно менять метрику

KNN

Упорядочиваем по метрике \Rightarrow к первых
в классе.

Как выбирать k?

$k > 1 \Rightarrow$ маленькие классы не детектируются

- обобщаются на регрессию.

Выходы деревьев - для задачи классификации

Выходы обобщенной регрессии

Какую метрику выбирать?

- зависимость от масштаба
- метрика \rightarrow близость

Проблема проклятия размерности

$\dim \mathbb{R} \Rightarrow$ Все объекты однокаково удалены

Метрические алгоритмы

- не нужно описывать признаки
- легко реализовать
- неконтролируемые
- нет обобщения
- мало гиперпараметров
- можно учитывать контекст.

Регрессия Надара-Ватсона

$$a(x) = \frac{\sum_{i=1}^m w_i(x) y_i}{\sum_{i=1}^m w_i(x)}$$

$$w_i(x) = k \left(\frac{d(x, x_i)}{h} \right)$$

ширина - h

Смысл: решение МНК

Метрики

Рассстояние Махаланобиса

$$x \rightarrow \varphi(x) = \sum^{-\frac{1}{2}} (\mathbf{x} - \boldsymbol{\mu})$$

$$\text{norm}(\boldsymbol{\mu}, \Sigma) \rightarrow \text{norm}(0, I) \sim N(0, I)$$

$$\rho(x, z) = \sqrt{(\varphi(x) - \varphi(z))^T (\varphi(x) - \varphi(z))}$$

Рассстояние Канторва.

$$\frac{1}{n} \sum_{i=1}^n \frac{|x_i - z_i|}{|x_i| + |z_i|}$$

В DL нонлинейно

косинусное расстояние

Рассстояние Дорсактарда

$$\rho(x, y) = 1 - \frac{|x \wedge y|}{|x \vee y|}$$

Метрика временных рядов

Матрица из разностей симкалов

расстояние **рекурсивное**: к элем-ту матрицы прибавляют минимум из предыдущих расстояний (3 соседних).