Машинное обучение. Введение. Обзор идеи методов.

Выполнила: Шамсутдинова Лилия

Машинное обучение

- Машинное обучение — процесс, в результате которого машина (компьютер) способна показывать поведение, которое в нее не было явно заложено (запрограммировано).
- Говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта
- На практике фаза обучения может предшествовать фазе работы алгоритма (например, детектирование лиц на снимке) — batch learning или обучение может проходить в процессе функционирования алгоритма (например, определение почтового спама) — online learning.

Классификация задач машинного обучения

- Дедуктивное обучение (экспертные системы)
- Индуктивное обучение (≈ статистическое обучение)
  - **Обучение с учителем:**
    - ∗ классификация
    - ∗ восстановление регрессии
    - ∗ структурное обучение (structured learning)
    - ∗ . . .
  - **Обучение без учителя:**
    - ∗ кластеризация
    - ∗ визуализация данных
    - ∗ понижение размерности
    - ∗ . . .
  - **Обучение с подкреплением (reinforcement learning)**
  - **Активное обучение**
  - **. . .**

Аппарат

- Линейная алгебра
- Теория вероятностей и математическая статистика
- Методы оптимизации
- Численные методы
- Математический анализ
- Дискретная математика
- и др.

Сферы приложения

- Компьютерное зрение (computer vision)
- Распознавание речи (speech recognition)
- Компьютерная лингвистика и обработка естественных языков (natural language processing)
- Медицинская диагностика
- Биоинформатика
- Техническая диагностика
- Финансовые приложения
- Рубрикация, аннотирование и упрощение текстов
- Информационный поиск
- Интеллектуальные игры
- . . .

Обучение по прецедентам

Множество $\mathscr{X}$ — *объекты, примеры* (samples)

Множество $\mathscr{Y}$ — *ответы, отклики, «метки»* (responses)

Имеется некоторая зависимость (детерминированная или вероятностная), позволяющая по $x \in \mathscr{X}$ предсказать (или оценить вероятность появления) $y \in \mathscr{Y}$.

(в частности, если зависимость детерминированная, то существует функция $f^* : \mathscr{X} \to \mathscr{Y}$)

Зависимость известна только на объектах из *обучающей выборки*:

$$\{(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \ldots,\ (x^{(N)}, y^{(N)})\}$$

Пара $(x^{(i)}, y^{(i)}) \in \mathscr{X} \times \mathscr{Y}$ — *прецедент*.

Задача *обучения по прецедентам*: восстановить зависимость, т. е. научиться по новым объектам $x \in \mathscr{X}$ предсказывать ответы $y \in \mathscr{Y}$.

Признаковые описания

$$x \in \mathscr{X} = Q_1 \times Q_2 \times \ldots \times Q_d,$$

где $Q_j = \mathrm{R}$ или Q_j — конечно

$$x = (x_1, x_2, \ldots, x_d) \in \mathscr{X}$$

x_j — j-й *признак* (*свойство*, *атрибут*) объекта x.

- Если Q_j конечно, то j-й признак — *номинальный* (*категориальный* или *фактор*).

  Можно считать, например, что, $Q_j = \{1, 2, \ldots, s_j\}$.

  Если $|Q_j| = 2$, то признак *бинарный* и можно считать, например, $Q_j = \{0, 1\}$ или $Q_j = \{-1, 1\}$.

- Если Q_j конечно и упорядочено, то признак *порядковый*.

  Например, $Q = \{\mathrm{Beginner}, \mathrm{Elementary}, \mathrm{Intermediate}, \mathrm{Advanced}, \mathrm{Proficiency}\}$ (уровень владения английским языком)

- Если $Q_j = \mathrm{R}$, то признак *количественный*.

Признаковые описания

Аналогично для выходов:
$y \in \mathscr{Y}$, где $\mathscr{Y} = \mathrm{R}$ или $\mathscr{Y}$ — конечно.

x называется *входом*,

y — *выходом*, или *откликом*

Компоненты x_j вектора x так же называют *входами* или *предикатными* (*объясняющими*) *переменными*.

В мат. статистике x_j называют «независимыми» переменными, а y — «зависимой».

Входные переменные и соответствующие им выходы известны для объектов обучающей выборки.

Признаковые описания

Значения признаков объектов из обучающей выборке и соответствующие ответы обычно записывают в матрицы:

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_d^{(N)} \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}$$

$$(\mathbf{X} \mid \mathbf{y}) = \left(\begin{array}{cccc|c} x_1^{(1)} & x_2^{(1)} & \dots & x_d^{(1)} & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_d^{(2)} & y^{(2)} \\ \dots & \dots & \dots & \dots & \dots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_d^{(N)} & y^{(N)} \end{array}\right)$$

Классификация задач обучения с учителем

В зависимости от множества $\mathscr{Y}$ выделяют разные типы задачи обучения.

- $\mathscr{Y}$ конечно, например, $\mathscr{Y} = \{1, 2, \ldots, K\}$, — *задача классификации* (или *задача распознавания образов*):

  $\mathscr{X}$ разбивается на K классов

  $$\mathscr{X}_k = \{x \in \mathscr{X} : f(x) = k\} \qquad (k = 1, 2, \ldots, K).$$

  По x требуется предсказать, какому классу он принадлежит.

- $\mathscr{Y} = \mathrm{R}$ — *задача восстановления регрессии.*

  Требуется найти функцию f из определенного класса, которая аппроксимирует неизвестную зависимость.

Ситуация, когда y — вектор, сводится к несколькими задачам со скалярным (атомарным) выходом.

y может быть чем-то более хитрым, например, графом, деревом, цепочкой символов (нефиксированной длины) — *структурное машинное обучение* (structured learning)

Обучение без учителя

Обучение по прецедентам — это *обучение с учителем*

Такое обучение можно рассматривать как игру двух лиц: ученика, который должен восстановить зависимость, и учителя, который для объектов из обучающей выборки указывает ученику соответствующий им выход.

Иногда можно считать, что объекты из обучающей выборки предъявляются средой, а иногда — их выбирает сам учитель, в некоторых случаях их выбирает ученик (*активное обучение*).

Рассматривается также *обучение без учителя*.

В этом случае нет учителя и «обучающая выборка» состоит только из объектов.

Ученик, имея только список объектов $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$, должен определить, как объекты связаны друг с другом.

Например, разбить объекты на группы (*кластеры*), так, чтобы в одном кластере оказались близкие друг к другу объекты, а в разных кластерах объекты были существенно различные.

Частичное обучение

- Каждый прецедент представляет собой пару «объект, ответ», но ответы известны только на части прецедентов.

Трансдуктивное обучение

- Дана конечная обучающая выборка прецедентов. Требуется по этим *частным* данным сделать предсказания относительно других *частных* данных — тестовой выборки. В отличие от стандартной постановки, здесь не требуется выявлять *общую* закономерность, поскольку известно, что новых тестовых прецедентов не будет. С другой стороны, появляется возможность улучшить качество предсказаний за счёт анализа всей тестовой выборки целиком, например, путём её кластеризации. Во многих приложениях трансдуктивное обучение практически не отличается от частичного обучения.

Обучение с подкреплением

- Роль объектов играют пары «ситуация, принятое решение», ответами являются значения функционала качества, характеризующего правильность принятых решений (реакцию среды).

Обозначения

d	число входных признаков
N	длина обучающей выборки
$\mathscr{X}$	множество объектов
$\mathscr{Y}$	множество ответов (выходов)
$x^{(1)}, x^{(2)}, \ldots, x^{(N)}$	объекты обучающей выборки, $x^{(i)} \in \mathscr{X}$ $(i = 1, 2, \ldots, N)$
$y^{(1)}, y^{(2)}, \ldots, y^{(N)}$	выходы для объектов из обучающей выборки, $y^{(i)} \in \mathscr{Y}$
K	количество классов (в задачах классификации)
$\mathsf{Pr}\, A$	вероятность события A
$\mathsf{Pr}\,(A \,\vert\, B)$	вероятность события A при условии, что наступило событие B
$P_X(x)$	интегральная функция распределения: $P_X(x) = \mathsf{Pr}\,\{X \leq x\}$
$p_X(x)$	плотность вероятности непрерывной случайной величины X
$P(y \,\vert\, x)$	условная интегральная функция распределения
$p(y \,\vert\, x)$	условная плотность вероятности
$\mathsf{E}\, X$	математическое ожидание случайной величины X
$\mathsf{D}\, X$ или $\mathsf{Var}\, X$	дисперсия случайной величины X
σX	среднее квадратическое отклонение: $\sigma X = \sqrt{\mathsf{D}\, X}$

Вероятностная постановка задачи

$\mathscr{X} = \mathrm{R}^d$ — множество объектов (входов) (точнее: множество их описаний)

$\mathscr{Y} = \mathrm{R}$ — множество ответов (выходов)

Будем рассматривать пары (x, y) как реализации $(d+1)$-мерной случайной величины (X, Y), заданной на вероятностном пространстве

$$(\mathscr{X} \times \mathscr{Y},\, \mathbf{A},\, \mathsf{Pr}), \qquad X \in \mathrm{R}^d,\ Y \in \mathrm{R}.$$

j-й признак — бинарный, номинальный или порядковый $\Leftrightarrow X_j$ — дискретная с. в.
j-й признак — количественный $\Leftrightarrow X_j$ — непрерывная с. в.

Интегральный закон распределения $P_{X,Y}(x, y)$ не известен, однако известна *обучающая выборка*

$$\left\{(x^{(1)}, y^{(1)}),\ (x^{(2)}, y^{(2)}),\ \ldots,\ (x^{(N)}, y^{(N)})\right\},$$

где $(x^{(i)}, y^{(i)})$ являются независимыми реализациями случайной величины (X, Y).

Требуется найти функцию $f: \ \mathscr{X} \to \mathscr{Y}$, которая по x предсказывает y, $f \in \mathscr{F}$

f называется *решающей функцией* или *решающим правилом*, а также *классификатором* (в случае задачи классификации).

Построение f называют *обучением*, *настройкой модели* и т. п.

Метод ближайших соседей

Будем, как и в задаче восстановления регрессии, для аппроксимации $\Pr(y|x)$ использовать k ближайших (по некоторому, например, евклидову расстоянию) объектов из обучающей выборки. Получаем метод k ближайших соседей для задачи классификации.

Пусть $N_k(x)$ — множество из k ближайших к x (по евклидову расстоянию) точек из обучающей выборки,

$I_k(x, y)$ — множество тех точек $x^{(i)}$ из $N_k(x)$, для которых $y^{(i)} = y$.

Согласно *методу k ближайших соседей* (kNN — k nearest neighbours) в качестве $f(x)$ берем результат голосования по всем точка из $I_k(x, y)$:

$$f(x) = \underset{y}{\operatorname{argmax}} |I_k(x, y)|,$$

Частным случаем является *метод (одного) ближайшего соседа*, в котором $f(x) = y^{(i)}$, где $x^{(i)}$ — ближайший к x объект из обучающей выборки.

В этом случае D_y представляют собой *области Вороного*

Наивный Байесовский классификатор. Предпосылки

- Теорема Байеса

$$P(W|Q) = \frac{P(Q|W)P(W)}{P(Q)} = \frac{P(Q|W)P(W)}{P(Q|W)P(W) + P(Q|M)P(M)}$$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

- Нормальное распределение

$$P(B|A) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right)$$

Практика

- Тренировочная выборка

Пол	Рост	Вес	Размер ноги
муж	6	180	12
муж	5.92	190	11
муж	5.58	170	12
муж	5.92	165	10
жен	5	100	6
жен	5.5	150	8
жен	5.42	130	7
жен	5.75	150	9

- Требуется классифицировать

Пол	Рост	Вес	Размер ноги
????	6	130	8

Наивный Байесовский классификатор. Алгоритм

- Шаг 1: вычисление параметров модели (мат. ожидание и дисперсия)

Пол	Мат ожидание (рост)	Диспресия (рост)	Мат ожидание (вес)	Дисперсия (вес)	Мат ожидание (размер ноги)	Дисперсия (размер ноги)
муж						
жен						

$$\bar{x}_0 = \frac{\sum_{i=1}^{m} x_i}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x}_0)^2}{N}$$

- Шаг 2: вычисление вероятностей для каждого параметра

$$p(\text{рост} \mid \text{муж}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\text{x} - \mu)^2}{2\sigma^2}\right)$$

- Шаг 3: вычисление результирующих вероятностей для каждого класса

$$posterior(\text{муж}) = \frac{P(\text{муж})\, p(\text{рост} \mid \text{муж})\, p(\text{вес} \mid \text{муж})\, p(\text{размер ноги} \mid \text{муж})}{evidence}$$

- Шаг 4: определение класса классифицируемого примера (наибольшая результирующая вероятность)

Дополнительно

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- $p(\theta|D)$— это то, что мы хотим найти, распределение вероятностей параметров модели *после* того, как мы приняли во внимание данные; это называется *апостериорной вероятностью*(posterior probability). Эту вероятность, как правило, напрямую не найти, и здесь как раз и нужна теорема Байеса. $p(D|\theta)$ – это так называемое *правдоподобие* (likelihood), вероятность данных при условии зафиксированных параметров модели; это как раз найти обычно легко, собственно, конструкция модели обычно в том и состоит, чтобы задать функцию правдоподобия. А $p(\theta)$ — *априорная вероятность* (prior probability), она является математической формализацией нашей интуиции о предмете, формализацией того, что мы знали раньше, ещё до всяких экспериментов.