Введение

Не точные определения, для понимания.

- **Искусственный Интелелкт (ИИ) Artificial Intellect (AI)** - область науки и технолгии, целью которой является создание *"умных машин"*, а так же "умных программ".
- **Большие данные Big data** - методы и подходы для работы с большими массивами данных, которые не поддаются обработке на одной машине.
- **Data science (наука о данных)** - анализ обработка и представление данных.
- **Машинное обучение machine learning** - подраздел ИИ, целью которого является создание алгоритмов, которые способны *"обучаться"*.
- **Искусственные нейронные сети Artificial Neural Networks** - класс алгоритмов машинного обучения.
- **Глубокое обучение deep learning** - раздел машинного обучения изучающий нейронные сети.
- **Классическое машинное обучение, классические алгоритмы машинного обучения, "классика"** - машинное обучение без нейронных сетей.

Как работает машинное обучение

Программу (алгоритм) пишет не человек. Человек выбирает модель и подает модели данные, после чего модель обучается и выдает алгоритм.

Связанные области из матетматики:

- Математическая статистика
- Методы оптимизации
- Линейная алгебра

Типы машинного обучения:

1. Обучение с учителем
2. Обучение без учителя

3. Обучение с подкреплением
4. ...

Обучение с учителем supervised learning

Пусть:

- Генеральная совокупность - $\hat{X}$. Все объекты в мире (может быть бесконечным).
- Все возможные ответы - $\hat{Y}$.

У нас есть:

- $\overline{X}$ - подвыбока реальных объектов. Пусть размер выборки (кол-во объектов в ней): N.
- Y - ответы на нашей выборке. Для каждого элемента из $\overline{X}$ есть соответствующий элемент Y.

Мы хотим разработать алгоритм, вычисляющий ответ для заданного объекта. Предаствим это в виде некоторой функции:

$$f : \hat{X} \to \hat{Y}$$

Обучение без учителя unsupervised learning

У нас нет ответов.
Наша цель найти некую внутренние взаимосвязи между существующими объектами (инсайды).

Признаковое описание объекта

С объектами из реального мира сложно работать без метематического описания. Для этого описания мы используем прзнаки.
Мы можем измерить характеристику j-ю характеристика любого объекта из генеральной совокупности:

$$f_j(\hat{X}_i) = x_{ij}, \text{где } \hat{X}_j \in \hat{X}$$

Измерив M характеристик объектов из нашего набора данных $\overline{X}$ мы получим матрицу X:

$$\begin{bmatrix} f_1(X_1) & f_2(X_1) & f_3(X_1) & \dots & f_M(X_1) \\ f_1(X_2) & f_2(X_2) & f_3(X_2) & \dots & f_M(X_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_1(X_N) & f_2(X_N) & f_3(X_N) & \dots & f_M(X_N) \end{bmatrix}$$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1M} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & x_{N3} & \dots & x_{NM} \end{bmatrix}$$

X называется *матрицей признаков*. Договоримся, что как только у нас есть выборка объектов, то у нас есть и матрица признаков и будем считать. Будем их считать синонимами:

$$\text{выборка объектов} \;=\; \text{матрица признаков}$$

Об объектах мы будем думать как об их признаковом описании .

Типы признаков $f_j : \hat{X}_i \to D_j$:

- $|D_i| = \{0, 1\}$ - *бинарный признак* (пол: мужской/женский)
- $|D_i| < \inf$ - *категориальный признак* (погода: пасмурно, солнечно, ветренно)
- $|D_i| < \inf$ - категориальный признак и при этом есть отношение порядка над D_i , то это *порядковый признак* (воинские звания, уровень образования)
- $|D_i| = \mathbb{R}$ - *прочие числовые признаки*(рост, уровеь давления)

Типы задач (ответов) в обучении с учителем

Классификация
$|Y| < \inf$

- $|Y| = \{+1, -1\}$ – бинарная классификация
- $|Y| = \{1, \dots, K\}$ – классификация на K, причем каждый объект относится к какому-то одному классу
- $|Y| = \{0, 1\}^K$ – классификация на K, причем каждый объект относится к нескольким классам (тегирование)

Регрессия
$|Y| = \mathbb{R}^m$ где $m \geq 1$

Если Y - бесконечное упорядченное множество. Такой тип задач называется ранжированием и может относиться как к обучению с учителем, так и к обучению без учителя.

Основные типы задач в обучении без учителя

Кластеризация - объект делится на непересекающиеся множества, которые называются кластерами. объекты из одного ласса должны "быть похожи" друг на друга, и отличаться от другого кластера.

Нахождение ассоциативных правил - нахождение признаков, которые идут как правило, вместе (пример:составление потребительской корзины).

Заполнение пропущенных данных – значения для каких то признаков и даже объектов могут быть пропущены, нам надо их как нибудь заполнить.

Сокращение размерности - исходные данные представлены в виде признаковой матрицы и число признаков может быть большим. Задача: представить эти данные в пространстве меньшей размерности, возможно с потерей информации.

Строгие определения. Обобщающая способность, качество обобщения.

Определения

Модель - параметрическое множество функций

$$\mathbb{A} = \{g(x, \theta) \mid \theta \in \Theta\}\ where\ g : \hat{X} * \Theta \to \hat{Y}$$

Метод обучения - отображение вида:

$$\mu : (\hat{X} * \hat{Y})^l \to \mathbb{A}$$

$(\hat{X} * \hat{Y})^l = (x_i, y_i)_{i=1}^l$ - выборка размера l, sample

$\alpha \in \mathbb{A}$ - алгоритм

В обучении с учителем две фазы:

1. Обучение(Train) - способ обучения μ по объектам $(\hat{X} * \hat{Y})^l$ строит алгоритм $\alpha = \mu((\hat{X} * \hat{Y})^l)$
2. Тестирование(Test) - алгоритм α для (новых) объектов $x \in \hat{X}$ предсказывает ответ(ы) $\alpha(x)$

Функция потерь, (loss function):
$L(\alpha, x)$ - мера ошибки алгоритма $\alpha \in \mathbb{A}$ на объекте $x \in \hat{X}$.

Определение функция y по $x \in X$ возвращает ответ для x из Y.

например для классифицикации:
$L(\alpha, x) = [\alpha(x) \neq y(x)]$ где $[True] = 1$ и $[False] = 0$

Для регрессии:
$L = |\alpha(x) - y(x)|$ - абсолютная ошибка.
$L = (\alpha(x) - y(x))^2$ - квадратичная ошибка.

Эмпирический риск(empirical risk(ER)) - функционал качества алгоритма на выборке:

$$Q(\alpha, \hat{X}^l) = \frac{1}{l} \sum_{i=1}^{l} L(\alpha, x_i)$$

$Q(\alpha, \hat{X}^l) \to min$ - чем меньше, тем лучше

Минимизация эмпирического риска - метод обучения

$$\mu(\hat{X}) = argmin_{\alpha \in \mathbb{A}} Q(\alpha, \hat{X}^l)$$

У нас есть: X, Y - объекты и ответы на них.
Мы выбираем: $\mathbb{A}(\Theta)$
Найти: $\alpha' = \mu(X)$

$$\mu(X) = min_{\alpha \in \mathbb{A}} Q(\alpha, X) =$$

$$min_{\theta \in \Theta} Q(\mathbb{A}(\theta), X) =$$

$$min_{\theta \in \Theta} Q(g(x, \theta), X) =$$

$$min_{\theta \in \Theta} \frac{1}{l} \sum_{i=1}^{l} L(g(x, \theta), x_i)$$

Например, для метода наименьших квандратов, МНК (ordinary least squares, OLS):

$$min_{\theta \in \Theta} \frac{1}{l} \sum_{i=1}^{l} (g(x, \theta) - y_i)^2$$

В результате, для задач машинного обучения мы получаем минимизационную задачу

Обощающая способность, качество обучения

Мы нашли "закон природы" или просто настроились на какой-то частный случай, подогнали $g(x, \theta)$ под заданные параметры? Будет ли полученный алгоритм аппроксиммировать желаемую функцию на всей генеральной совокупности? Хорошо ли алгоритм будет работать на новых данных? Не переобучились (overfit) ли мы?

Пусть $\hat{X}^l$ - обучающая выборка, $\hat{X}^k$ - тестовая выборка. Если

$$Q(\mu(\hat{X}^l), \hat{X}^k) \gg Q(\mu(\hat{X}^l), \hat{X}^l)$$

значит алгоритм переобучен.

Как избежать переобучения?

Проверяем на отложенной выборке (hold-out):

$$HO(\mu, X_{train}, X_{test}) = Q(\mu(X_{train}), X_{test}) \to min$$

Перекрестная проверка (cross-validation):

$$CV(\mu, X^{l+k}) = \frac{1}{|n|} \sum_{i \in n} Q(\mu(X_i^l), X_i^k)$$

Теоретическая оценка эмпирического риска:

$$E\left[\, Q(\mu(X_i^l), X_i^k) \,\right] \to min$$

E - мат ожидание.

Почему мы можем переобучиться?

- большая сложность пространства параметров Θ
- Переобучение будет всегда если мы оптимизируем параметры не для генеральной совокупности.

Что делать? Избавиться совсем нельзя, можно минимизировать.

- минимизировать одну из теоретических оценок
- накладывать ограничения на θ (регуляризация)
- минимизировать HO или CV

Особенности некоторых задач

- Требуется интерпретируемость алгоритма
- Требуется предсказывать вероятность ошибки
- "Большой" размер объектов
- Пропуски в данных
- "Сырые данные", сложности в выборке и построении признаков
- Данные могут быть устаревшими
- Не размеченные и плохо размеченные данные
- Малое количество обучающих примеров
- Типы информации различны:текст + звук + видео
- Нужен высокопроизводительный алгоритм
- Разные функции потерь
- ...

Примеры задач

Задача кредитного скоринга

Выдать или нет кредит по заявке.

Объект: клиент (его заявка на выдачу кредита)

Возможные признаки: пол, возраст, место проживания, профессия, зарплата, стаж работы, сумма кредита, образование, наличие в собственности квартиры, машины, семейное положение.

Задача категоризации новостей

Разбить новостные статьи по категориям (возможна иерархия)

Объект: Статья (текст)

Возможные признаки: издание, автор, частота в тексте(заголовке) того или иного признака.

Задача ранжирования поисковой выдачи (вычитка научных статей)

Выдать последовательность сайтов по запросу. Подходит не подходит документ под запрос, и в какой по порядку.

Объект: пара запрос, документ (веб страница)

Возможные признаки: частота слов запроса в документе, число ссылок на документ, число кликов на документ (просмотров) вообще и по данному запросу.

Задача прогнозируемости стоимости квартиры

Определить стоимость квартиры в конкретный момент.

Объект: квартира

Возможные признаки: район города, число комнат, жилая площадь, общая площадь, высота потолков, расстояние до центра, до метро, этаж, наличие балкона, лифта, мусоропровода, возраст дома, тип постройки (кирпич, панель, монолит...)

Задачи компьютерного зрения, на фото видео спецоборудовании

- Распознавание объекта
- Детектирование объекта
- Определения свойств объекта
- ...

На основе звука

- Те же что и в задачах компьютерного зрения
- Распознавание слов, цифр, текстов, команд
- ...

Задачи медицины и биологии

- Медицинское диагностирование
- Исследования на уровне генома
- ...

Игры

- GO
- На бирже
- ...

Задачи продавцов, магазинов

- Кластеризация покупателей для эффективных рекламных акций.
- Предсказание оттока клиентов.
- Прогнозирование объемов продаж.
- Контекстная реклама
- ...

Хранение больших объемов информации

- поиск наиболее информативных, или интересных признаков
- ...

Линейная регрессия как пример модели алгоритма. Градиентный спуск и аналитическое решение.

Линейная регрессия - это модель алгоритмов которая выглядит как:

$$\mathbb{A} = \{g(x, \theta) \mid \theta \in \Theta\}$$

где $g(x, \theta) = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \ldots + \theta_3 * x_l$ и

$$\theta = [\theta_0, \theta_1, \theta_2, \ldots, \theta_l]$$

Дано:

$X * \theta = Y_{predict}$

Функция потерь $L = (\alpha(x) - y(x))^2$ - квадратичная ошибка.

Минимизация эмперического риска -> методу наименьших квадратов

$Q \to min_\theta$

$\sum(\alpha(x) - y(x))^2 \to min_\theta$

$(Y_{predict} - Y)^T * (Y_{predict} - Y) \to min_\theta$

$(X * \theta - Y)^T * (X * \theta - Y) \to min_\theta$

Аналитическое решение

$$(X^T * X) * \theta = X^T * Y \Rightarrow \theta = (X^T * X)^{-1} * X^T * Y$$

$(X^T * X)^{-1} * X^T$ - псевдообратная матрица

Аналитическое решение сложное по вычислениям.
Поэтому используем **градиентный спуск**
итеративный метод:

$$\theta_k^{(j+1)} := \theta_k^{(j)} - \gamma * \frac{\partial}{\partial \theta_k^{(j)}} Q$$

Для МНК:

$$Q = \frac{1}{2l}\sum_{i=1}^{l}((\alpha(x_i) - y_i)^2$$

$$\frac{\partial}{\partial_k \theta}Q = \frac{1}{l}\sum_{i=1}^{l}((\alpha(x_i) - y_i) * x_{ik}$$

повторяем пока не сойдется:

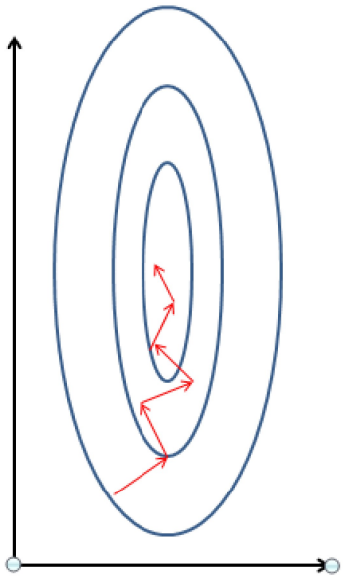
$$\theta_k^{(j+1)} := \theta_k^{(j)} - \gamma * \frac{1}{l} * \sum_{i=1}^{l}((\alpha(x_i) - y_i) * x_{ik}$$

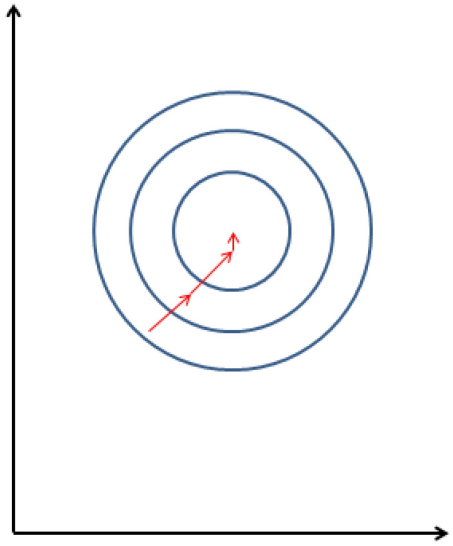
в матричной форме:

$$\theta^{(j+1)} := \theta^{(j)} - \gamma * \frac{1}{l} * (X^T * (X * \theta^{(j)} - Y))$$

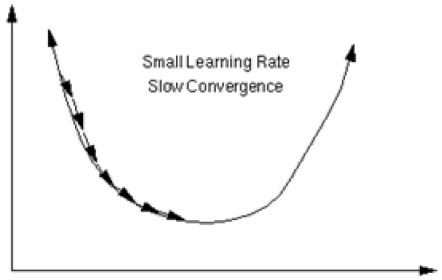
Особенности градиентного спуска

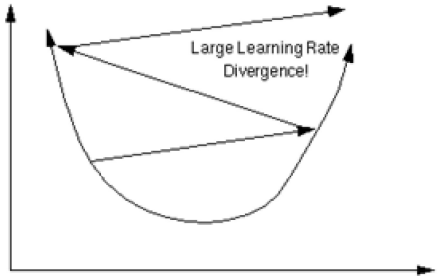
normalize features



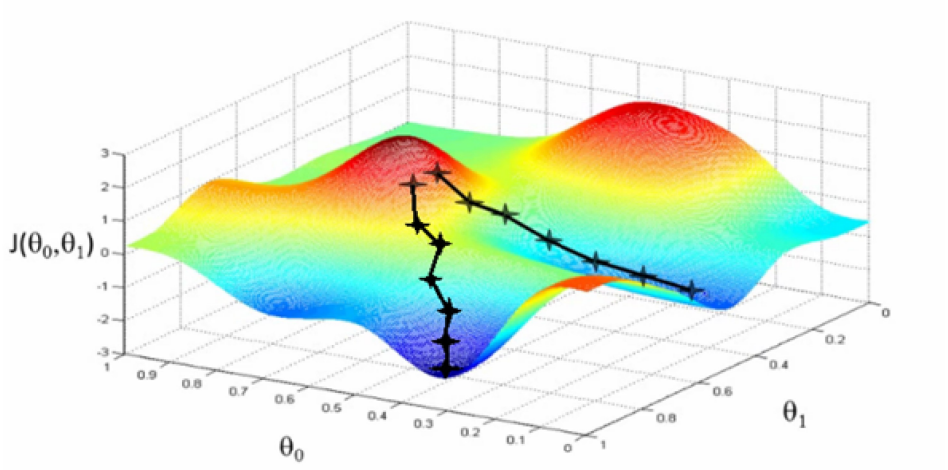


choosing learning rate





several minimus



градиентный спуск

- подобрать параметры
  - learning rate
  - максимальное число итераций
  - условие сходимости
  - стартовые параметры (сетка)
- легко реализуем
- хорошо работает на больших выборках
- нет уверенности в том что выдал minimum функции

Подходит для большинства моделей.

Используется (с некоторыми усовершенствованиями) в том числе нейронных сетях.

Резюме

- Основные понятия машинного обучения: Обучение с учителем, без учителя, объект, ответ, признак, алгоритм, модель алгоритмов, метод обучения, эмпирический риск, переобучение, кластеризация, классификация, восстановление регрессии, ранжирование.

Этапы решения задач машинного обучения:

- понимание задачи и данных;
- предобработка данных и изобретение признаков;
- !построение модели;
- !сведение обучения к оптимизации;
- !решение проблем переобучения и эффективности;
- !оценивание качества;
- внедрение и эксплуатация.

Градиентный спуск.