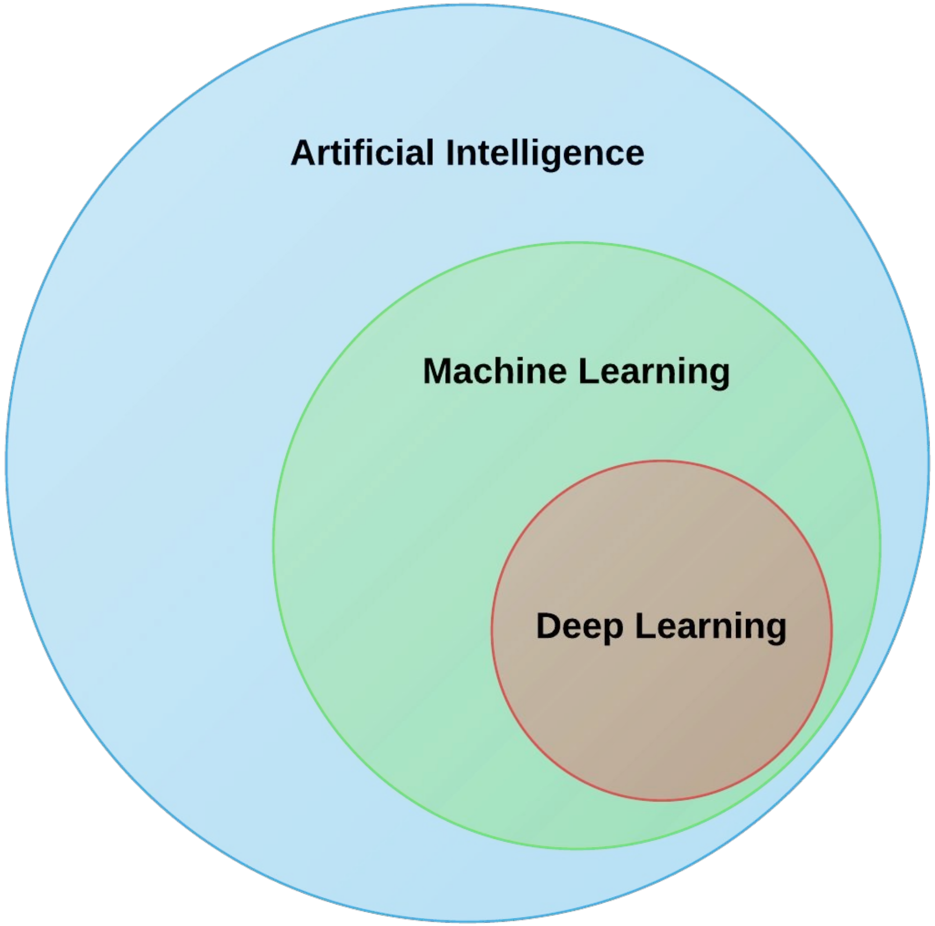
Введение в машинное обучение

Лекция 11

Машинное обучение

Машинное обучение – это наука, изучающая способы извлечения закономерностей из ограниченного количества примеров



Машинное обучение

- $\mathbb{X}$ – пространство объектов
- $\mathbb{Y}$ – пространство ответов
- $X = (\{x_1, y_1\}, \dots, \{x_l, y_l\})$ – обучающая выборка
- $x_1, \dots, x_l$ - обучающие объекты
- $y_1, \dots, y_l$ - ответы
- x – признаковое описание объекта

Признаки

- Бинарные
- Вещественные
- Категориальные
- Ординальные
- Сложные (напр. фотографии, текст, звук)

Виды задач машинного обучения

- Обучение с учителем (supervised learning)
  - $\mathbb{Y} = \mathbb{R}$ - задача регрессии
  - $\mathbb{Y} \in \{0, 1\}$ – бинарная классификация
  - $\mathbb{Y} \in \{1, \dots, K\}$ – многоклассовая классификация
  - $\mathbb{Y} \in \{0, 1\}^{\mathrm{k}}$ - многоклассовая классификация с пересекающимися классами (multi-label classification)
- Обучение без учителя (unsupervised learning)
  - Кластеризация
  - Понижение размерности
  - Оценивание плотности
  - Визуализация

Обучение

- Пусть есть обучающая выборка $\mathrm{X} \in \mathbb{R}^{l \times d}$
- Хотим построить функцию (модель) $a : \mathbb{X} \to \mathbb{Y}$
- Введем функционал ошибки: $Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2$
- Выберем семейство алгоритмов $\mathcal{A}$
- $\mathcal{A} = \{a(x) = w_0 + w_1 x_1 + \ \dots + w_d x_d | w_0, w_1, \dots, w_d \in \mathbb{R} \ \ \}$
- $\frac{1}{l}\sum_{i=1}^{l}\left(w_0 + \sum_{i=1}^{d} w_i x_{ij} - \ y_i\right)^2 \to \min_{w_0, w_1, \dots, w_d}$

Линейная регрессия

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

- w_j - веса или коэффициенты модели
- w_0 - bias или сдвиг
- Сумму можно представить в виде скалярного произведения
$$a(x) = w_0 + \langle w, x \rangle$$
- Если добавить в признаковое описание объектов еще один единичный признак, то
$$a(x) = \langle w, x \rangle$$

Функционалы ошибки

$$MSE(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2$$

Root Mean Squared Error

$$RMSE(a, X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2}$$

Коэффициент детерминации

$$R^2(a, x) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \bar{y})^2}$$

Mean Absolute Error

$$MAE(a, X) = \frac{1}{l}\sum_{i=1}^{l} |a(x_i) - y_i|$$

Mean Absolute Percentage Error

$$MAPE(a, X) = \frac{1}{l}\sum_{i=1}^{l} |\frac{a(x_i) - y_i}{y_i}|$$

Обучение линейной регрессии

Если возьмем MSE:

$$\frac{1}{l}\sum_{i=1}^{l}(\langle w, x_i\rangle - \ y_i)^2 \to \min_w$$

В матричном виде:

$$\frac{1}{l}\left||Xw \ - y|\right|^2 \to \min_w$$

Обучение линейной регрессии

- Точное аналитическое решение для уравнения с предыдущего слайда

$$w = (X^T X)^{-1} X^T y$$

У этого решения есть минусы:

- $O(N^2 D + D^3)$
- Матрица $X^T X$ может быть вырожденной или плохо обусловленной
- Не для всех функционалов ошибки существует точное решение

Нужен другой подход

Градиент

Градиентом функции $f\colon \mathbb{R}^d \to \mathbb{R}$ называется вектор ее частных производных:

$$\nabla f(x_1, \dots, x_d) = \left(\frac{\partial f}{\partial x_j}\right)_{j=1}^{d}$$

- Градиент направлен в сторону наискорейшего роста функции
- Антиградиент $(-\nabla f)$ направлен в сторону наискорейшего убывания

Градиентый спуск

1. Инициализируем начальный набор параметров $w^{(0)}$
2. Сдвигаемся в сторону антиградиента
3. Вычисляем новое значение антиградиента
4. Возвращаемся к шагу 2

$$w^{(k)} = w^{(k-1)} - \alpha \nabla Q(w^{(k-1)})$$

- α – длина шага обучения
- $Q(w)$ - значение функционала ошибки при параметре w
- $O(NDS)$
- Находим локальные минимумы, не обязательно глобальные

Градиентный спуск

Посчитаем градиент MSE

$$\nabla_w L = \frac{2}{l} X^T (Xw \ - y)$$

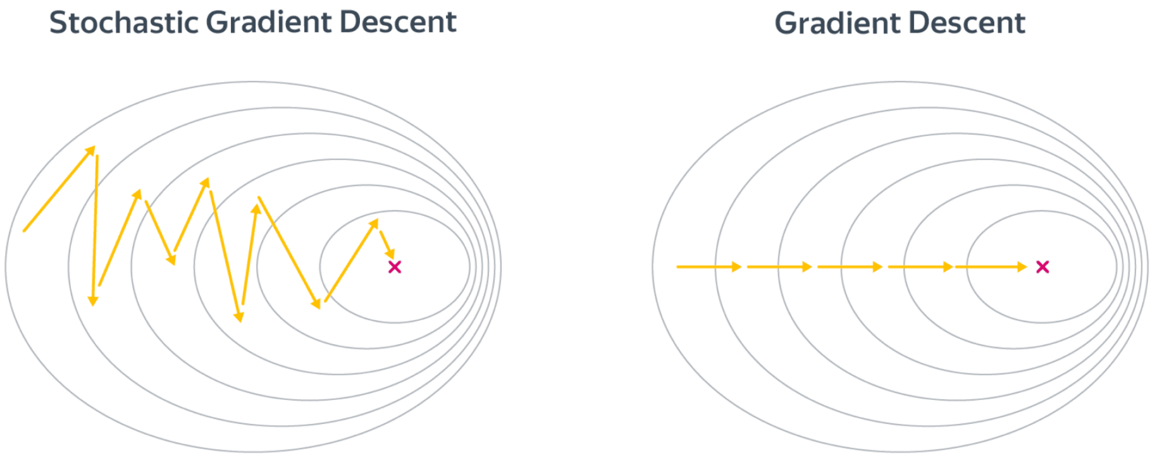
Подставим градиент в алгоритм градиентного спуска

$$w^{(k)} = w^{(k-1)} - \alpha X^T (Xw \ - y)$$

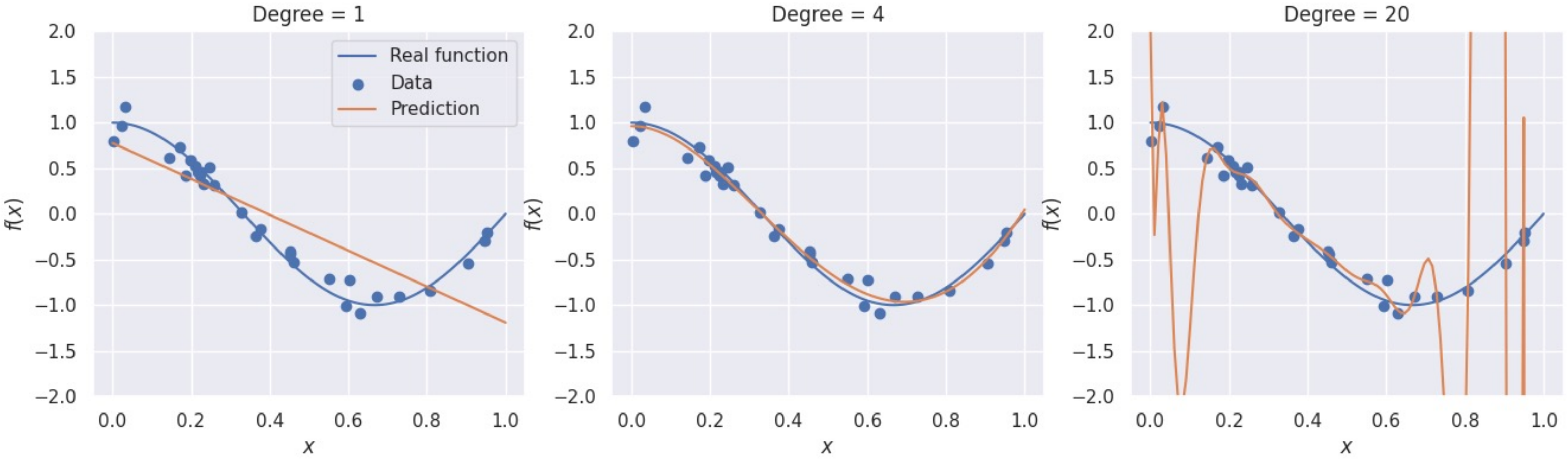
Стохастический градиентный спуск

Градиентный спуск можно ускорить, если вместо вычисления градиента по всей выборке, вычислять градиент на небольшой подвыборке (батче)

- $O(NDE)$
- Случайно перемешиваем выборку
- Линейным проходом набираем батчи



Переобучение и недообучение



Регуляризация

- Если признаки мультиколлинеарны, то решение задачи регрессии не всегда единственно $\Longrightarrow$ веса могут быть очень большими, что приводит к вычислительным сложностям
- Чтобы этого избежать, используют регуляризацию

$$l1\colon \frac{1}{l}\left||Xw - y|\right|^2 + \lambda|w|_1^1 \to \min_w$$

$$l2\colon \frac{1}{l}\left||Xw - y|\right|^2 + \lambda|w|_2^2 \to \min_w$$

Где λ – коэффициент регуляризации