$f_1 \dots f_d$ — признаки

$x_1 \dots x_n$ — объекты $\in \mathbb{R}^d$

$y_1 \dots y_n$ — таргеты

общ. с учит.:
- Регрессия $y_i \in \mathbb{R}$
- Классификация $y_i \in \{1 \dots k\}$ (k маленькое < 1000)

Лин. регр.: $y(x) = \theta^T x$ ← мн-во моделей

МНК: $\|Y - X\theta\|^2 \to \min_\theta$ ← функционал ошибки

$Y_{n\times 1} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $X_{n\times d} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

$$\hat{\theta} = \underset{d\times n}{(X^T X)^{-1}} \underset{n\times d}{X^T} \underset{d\times n}{} \underset{n\times 1}{Y}$$

Ridge: $\|Y - X\Theta\|^2 + \lambda \|\Theta\|^2 \to \min_{\Theta}$

$F(\Theta)$

Град. спуск: $\Theta_{t+1} = \Theta_t - \eta \nabla F(\Theta_t)$

F функционал ошибки

Классификация:

тоже что-то было

Обучение без учителя — нет таргетов

Задача кластеризации:

$X = (x_1 \dots x_n)$ — выборка объектов

Хотим выявить K кластеров

Кластер может быть:

- Подвыборкой $f: X \to \{1 \dots k\}$
- Область пр-ва $f: \mathbb{R}^d \to \{1 \dots k\}$
- Нечетким $f(x) = (p_1 \dots p_k) \quad \sum p_i = 1$

p_i – вероятность попадения $(\cdot)$ в i-ый кластер

Цели кластеризации:

- Разбить на группы, работать с группами отдельно
- Сократить объем данных
- Выделить нетипичных объектов
- Выделить схожих по поведению объектов

Требования к форме кластеров:

- понимать какую форму даст метод
- понимать какая форма нужна в задаче
- обычно кластеры выпуклые

Масштабирование и расстояния

- кластеры зависят от метрики
- надо уметь интерпретировать метрики
- стандартизация
- учитывать степень важности признака

(сделать стандартизацию и указать веса признаков)

$$\sqrt{(x_1 - x_2)^2 + 5(y_1 - y_2)^2}$$

Метрики для кластеров

- среднее внутрикластерное расстояние = F_1

  опт. — точка = кластер

- среднее межкластерное расстояние = F_2

  опт. — 1 кластер

- F_1 / F_2

— силуэт

a_x = среднее расстояние до (·) его кластера

b_x = среднее расстояние до всех объектов ближайшего другого кластера

$$\text{силуэт} = \frac{b_x - a_x}{\max(b_x, a_x)}$$

силуэт выборки = средний силуэт

Практика:

- внутренние метрики не скажешь заказчику
- нужно убедить его, что кластеры хорошие

  для этого можно нарисовать графики
- кластеризация — вспомогательная задача

  можно оценивать кластеризацию как часть формулы ошибки

K – means

$X = (x_1 \dots x_n)$ — K — фиксировано

$f: \mathcal{X} \to \{1 \dots K\}$

1. Начальные приближения центров $\mu_1 \dots \mu_k$
2. Повторять
   - отнести объект к ближайшему центру
   - новые центры μ_i = центр масс

   Пока μ_i меняются

Утв. K-means оптимизирует сумму квадратов расстояний до соотв. центров

▲ 1) При отнесении (·) к ближ. центру мы не увеличиваемся

2) При смещении центра мы выбираем его опт. позицию

$\Rightarrow$ не возрастаем

т.к. у нас конечное кол-во расположений центров

Особенности:

- сходится к локальному оптимуму
- кластеры — выпуклые
-

K-means ++

EM - алгоритм

DBScan

$X = (x_1 \dots x_n)$ $x_i \in \mathbb{R}^d$ K – не фикс.

гиперпараметры ε, m

для $(\cdot)$ рассм $B_\varepsilon(\cdot)$

если возле точки $\geq m$ $(\cdot)$, то она осн.

если возле точки есть осн., то она погр.

иначе она шумовая

1 Берём не помеченный x

2 $|U_\varepsilon(x)| < m$ — ? шум

3 $|U_\varepsilon(x)| \geq m$ — создаём кластер

добавляем в него все основные и потом пограничные

повторяем

Иерархическая кластеризация