第二章 回归算法

在实际任务中经常预测一个值："指标"，"价格"，"金额"，"分数"

↙ 高考分数　↓ 房价　↓ 贷款·信用卡　↘ 考研分数

2.1 线性回归算法

提出者：英国生物学家·统计学家：高尔顿，父代与子代之间的高关系。

线性回归最简单，实用的算法之一。机器学习，掌握思路

假设有一场选秀大赛，组织会先了解某个选手的个人信息，例如：身高，体重，颜值，依据这些指标，决定一个选手是否能"晋级"，从而确定选手的"综合分数"

组织如何进行评估的？？

是 否 ⇒ 分类　　综合分数 ⇒ 回归

选秀评估系统的建模过程：

DataSet（数据集）

ID	X_1 身高(m)	X_2 体重(kg)	y 综合得分
001	1.73	60	82
002	1.76	55	90
003	1.82	65	87
004	1.77	53	92
005	1.71	65	75

DataSet ⇓ pandas

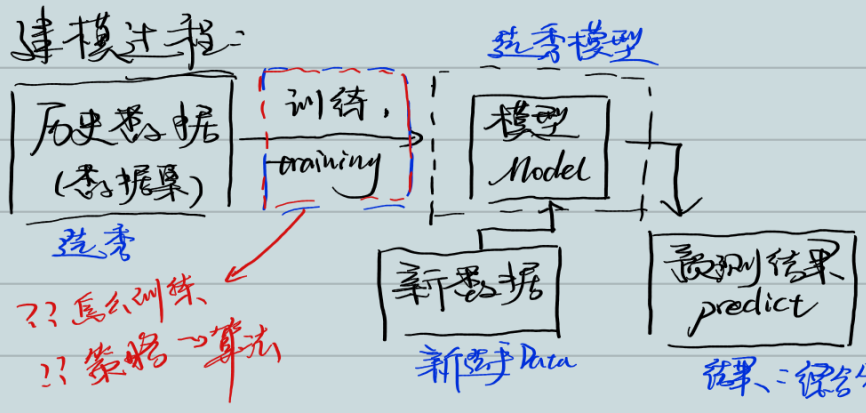
X_1 = 身高 ··· → "特征"、"属性"

X_2 = 体重 ··· → "　"、"　"　} ⇒ 综合得分（标签）

feature ——————→ Label

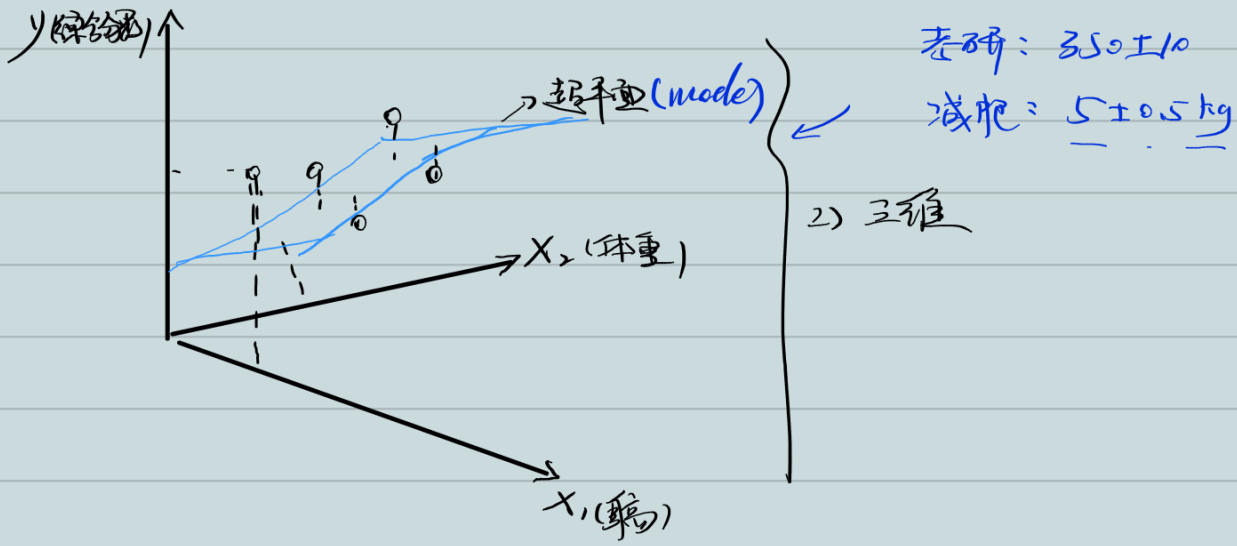
目的：得到一个计算 x_1、x_2 与 y 之间的关系。

机器学习建模过程：



2.2. 线性回归方程.

两个属性 ---> 值 影响



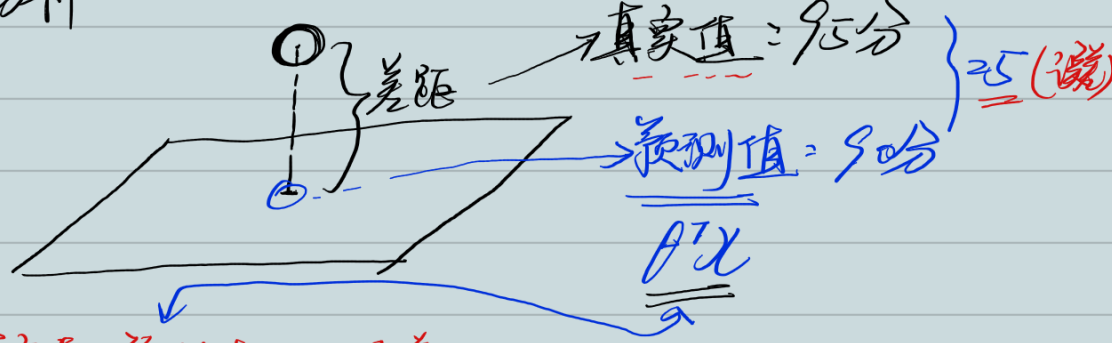
圆点表示真实的得分值，平面表示预测值，可以观察到，综合得分受 X_1 身高和 X_2 体重共同影响的. 每个特征(X_1、X_2)对y的影响有多大. 预测时每个特征加入参考 -> θ. 公式：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

偏置 影参 身高 体参 体重

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \rightarrow \text{New} \rightarrow \text{☆} = 95$$

2.3. 误差项分析



结论：真实值 = 预测值 + 误差

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$