机器学习基础

Yunpei Liu 2020.09.08

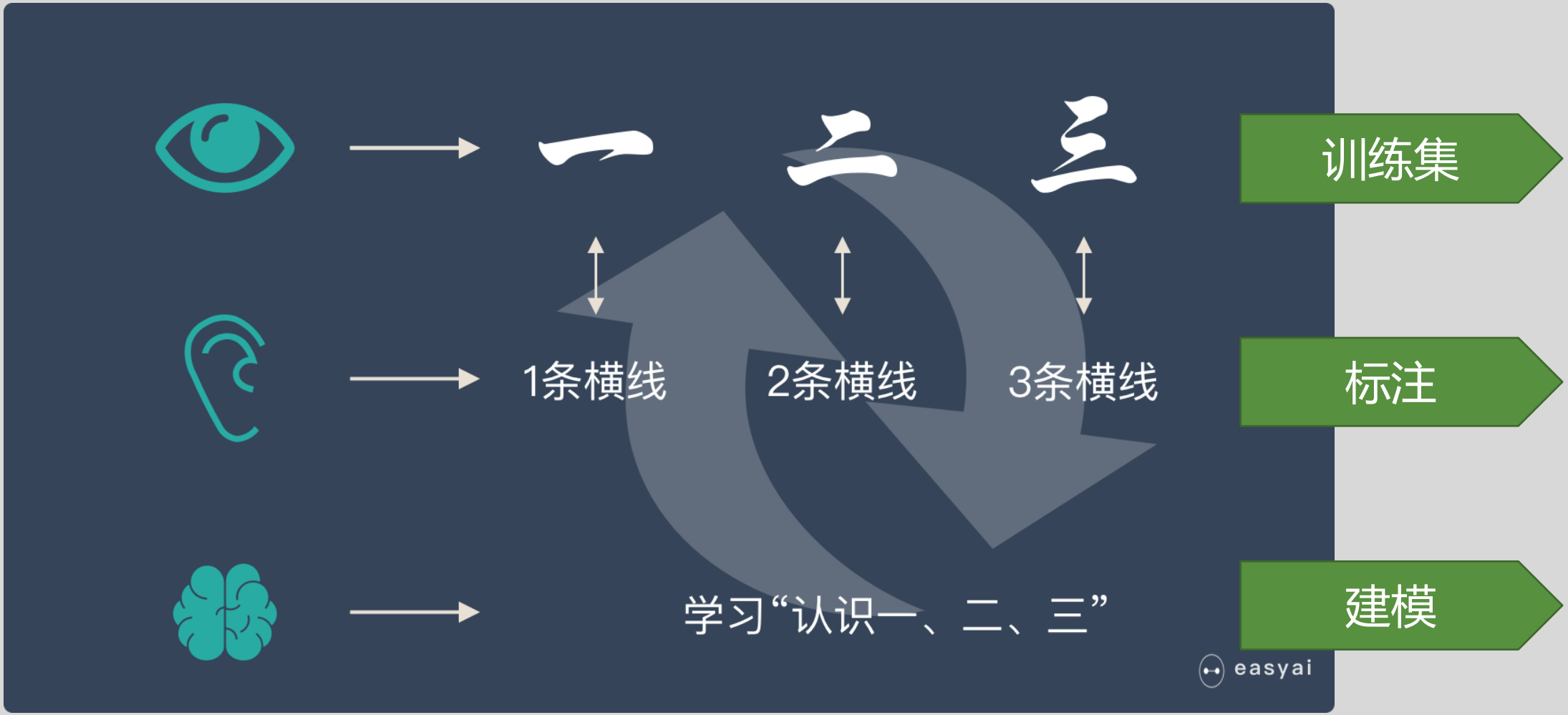
什么是机器学习

机器学习(Machine Learning)

- *Field of study that gives computers the ability to learn without being explicitly programmed.*
  在没有明确设置的情况下，使计算机具有**学习能力**的研究领域。

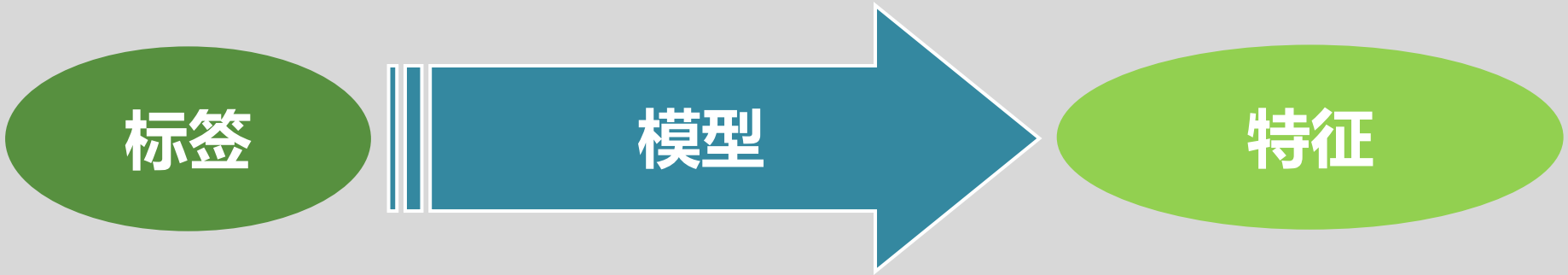
——Arthur Samuel (1956)

学习的过程



https://easyai.tech/ai-definition/machine-learning/

机器的学习



- 计算机从被标注的数据集出发，训练模型，预测数据的对应特征。

机器学习的分类

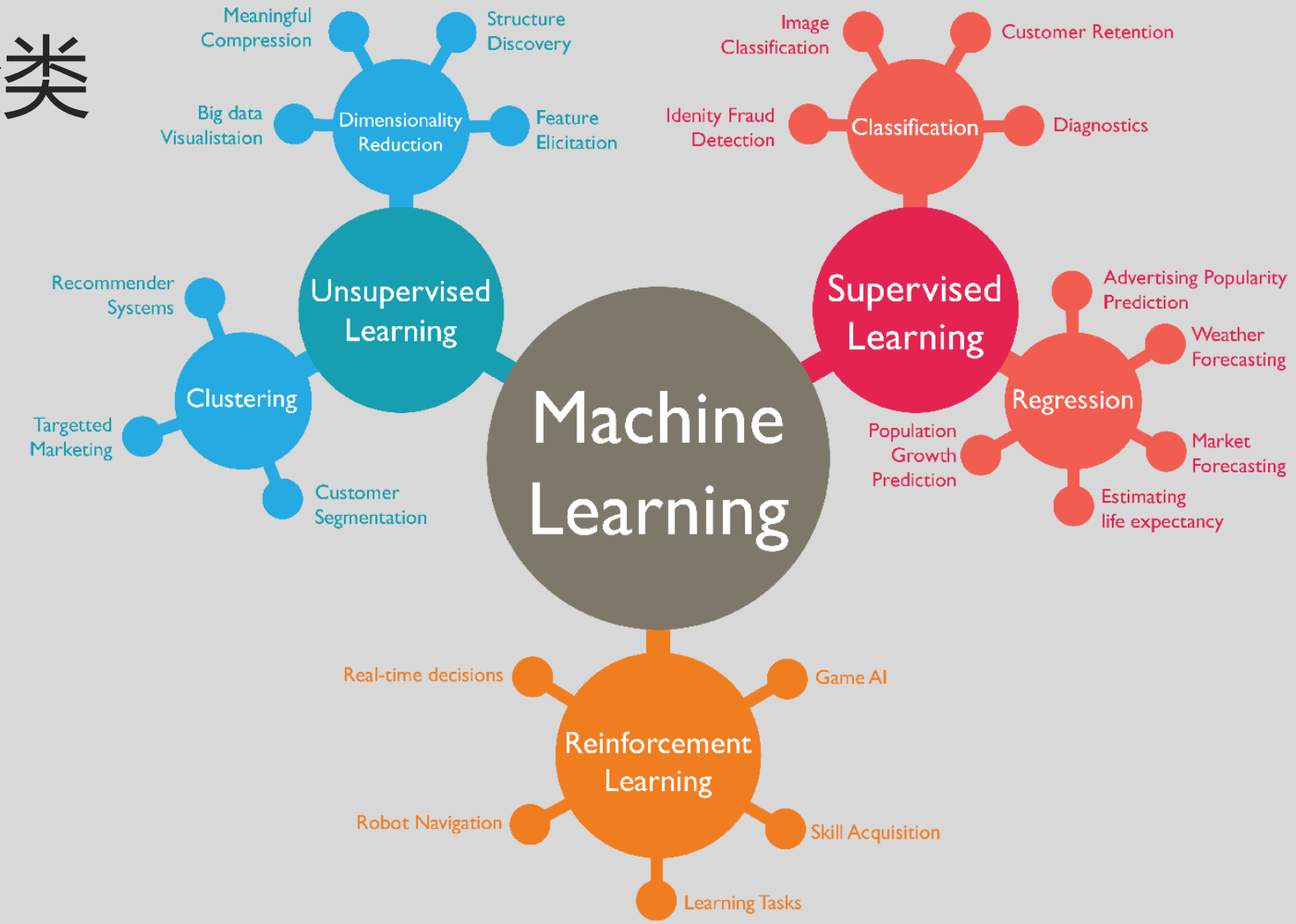
监督学习

- 分类
- 回归

非监督学习

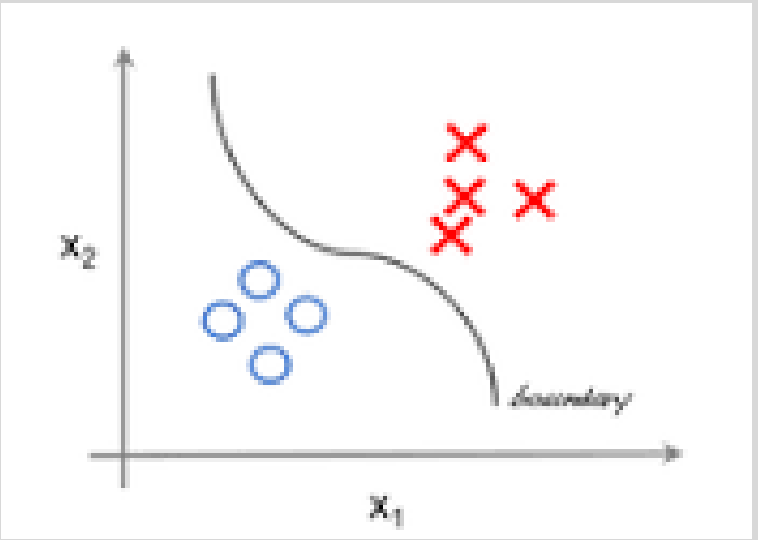
- 聚类
- 降维

强化学习



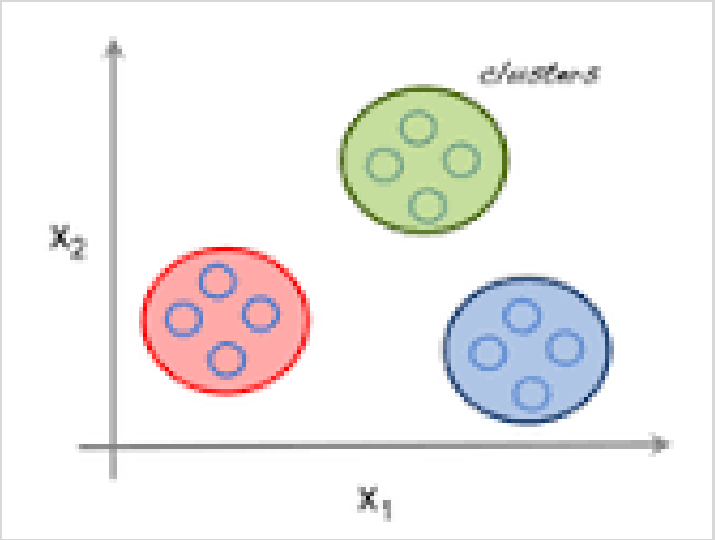
监督学习

- 在监督学习中，提供给算法的包含所需解决方案的训练数据，成为标签或标记。
- 包括：
  - 分类
  - 回归
- 举例：
  - 垃圾邮件分类
  - 根据肿瘤的尺寸等信息判断良性还是恶性
  - 机器学习势函数



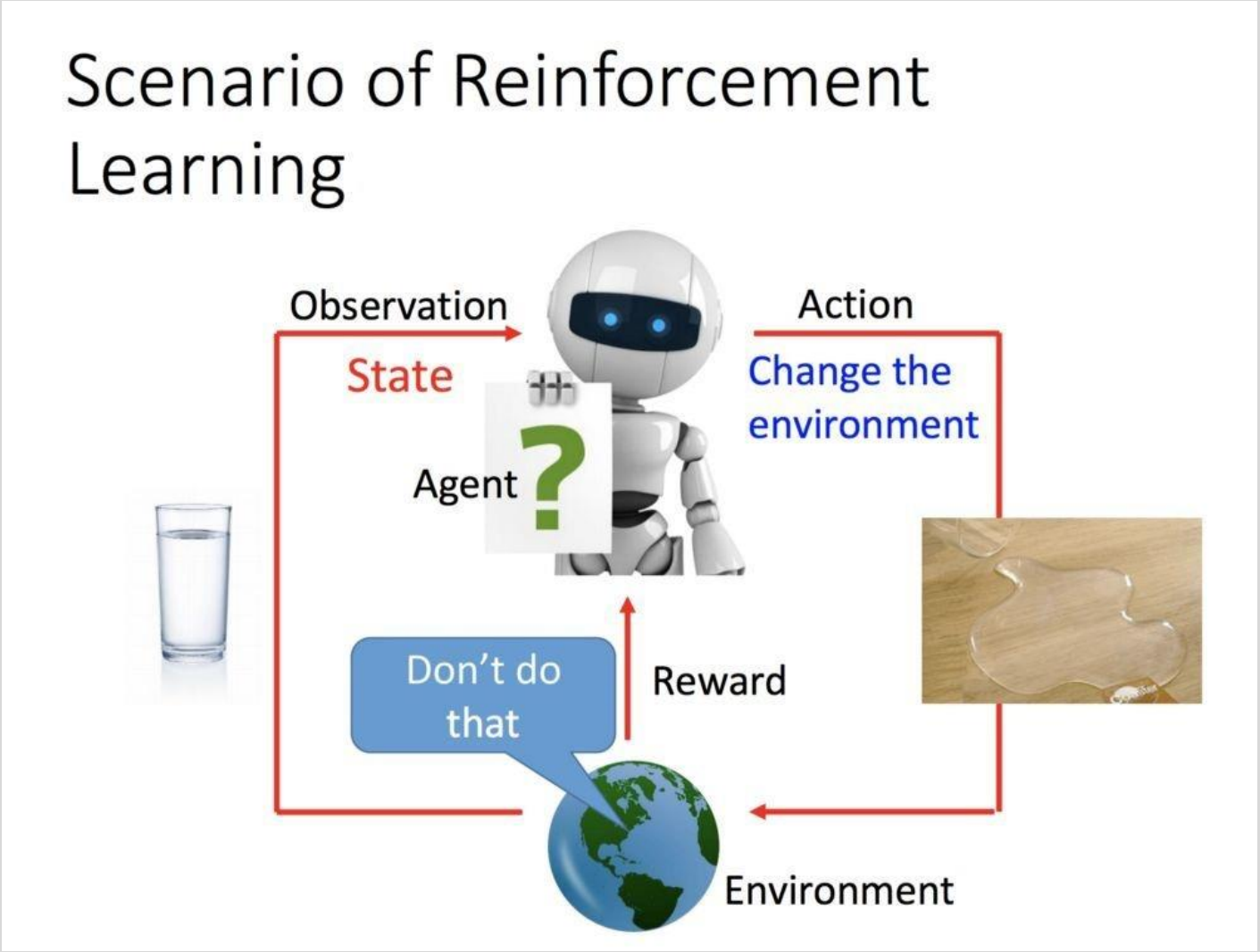
非监督学习

- 非监督学习的训练数据都是未经标记的，算法会在没有指导的情况下自动学习。
- 包括
  - 聚类
  - 降维
- 举例：
  - 大数据推荐
  - 从多个因素中提取一个新的变量表征研究对象的属性



强化学习

- 强化学习是一个非常与众不同的算法，它的学习系统能够观测环境，做出选择，执行操作并获得回报，或者是以负面回报的形式获得惩罚。它必须自行学习什么是最好的策略，从而随着时间推移获得最大的回报。



常见概念解释

数据集

- 训练集（Training set） —— 学习样本数据集，通过匹配一些参数来建立一个模型，主要用来训练模型。
- 验证集（Validation set） —— 对学习出来的模型进行验证，调整模型的参数，如在神经网络中选择隐藏单元数。验证集还用来确定网络结构或者控制模型复杂程度的参数。
- 测试集（Test set） —— 在该数据集上测试训练好的模型的分辨能力。

误差评价

- **均方误差 (MSE)** 指的是每个样本的平均平方损失。要计算 MSE，请求出各个样本的所有平方损失之和，然后除以样本数量：

$$MSE = \frac{1}{N} \sum_{(x,y)\in D} (\hat{y} - y)^2$$

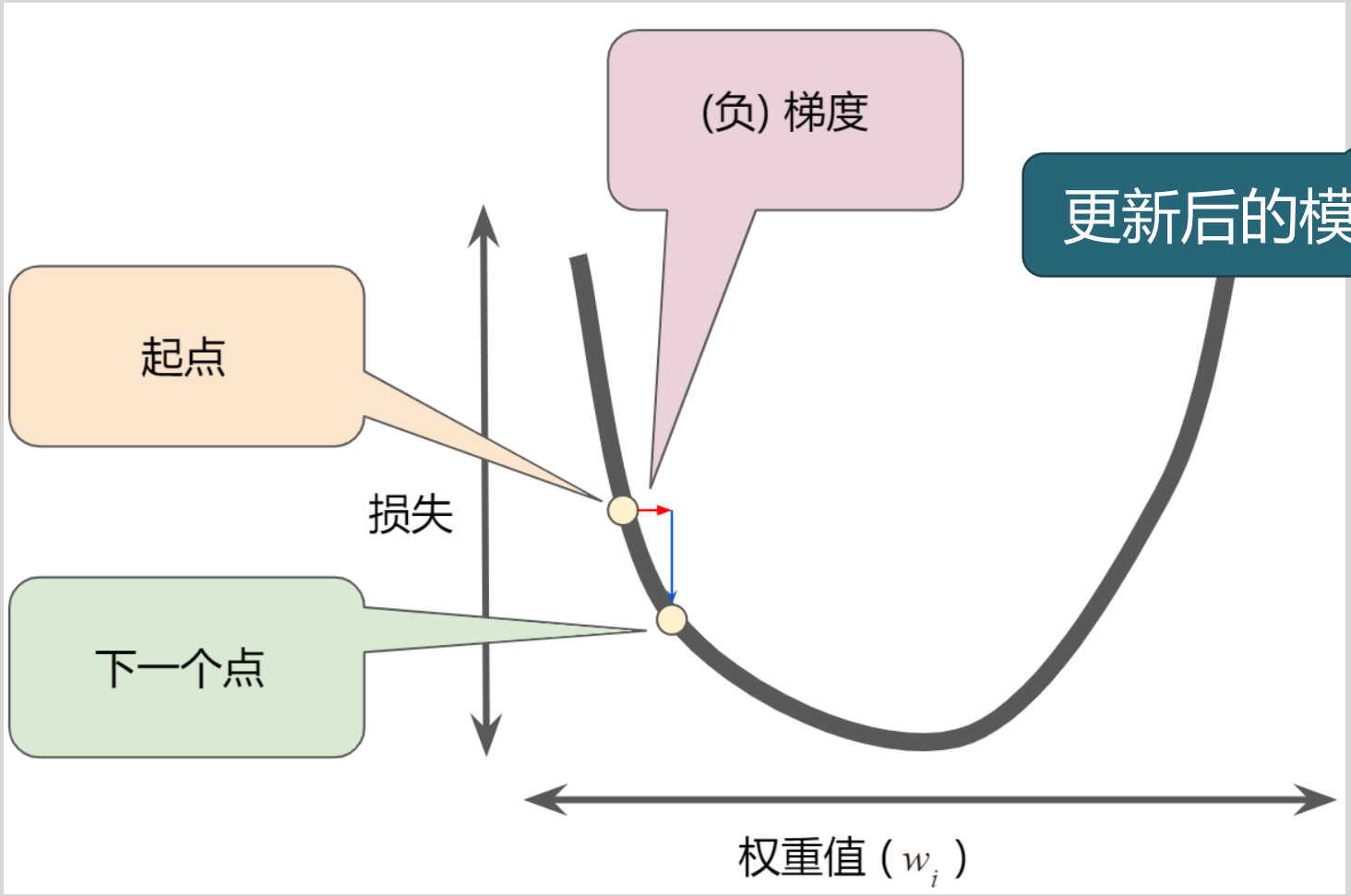
- (x, y)指的是样本，其中 x 指的是模型进行预测时使用的特征集， y指的是样本的标签。
- $\hat{y}$ 指的是从特征集 x 依据模型预测得到的函数值。
- D指的是包含多个有标签样本（即(x, y)）的数据集。
- N指的是D中的样本数量。
- 常用根均方差RMSE来表示误差大小。

损失函数

- **损失函数**用来评价模型的**预测值**和**真实值**不一样的程度，损失函数越好，通常模型的性能越好。不同的模型用的损失函数一般也不一样。
- 均方误差就是一种常用的损失函数——平方损失函数。
- 机器学习训练的目标：**降低损失函数的值。**

梯度下降法

- 常用的降低损失函数值的方法之一。



初始模型参数

$$\boldsymbol{\Theta}^1 = \boldsymbol{\Theta}^0 - \alpha \nabla J(\boldsymbol{\Theta})$$

更新后的模型参数

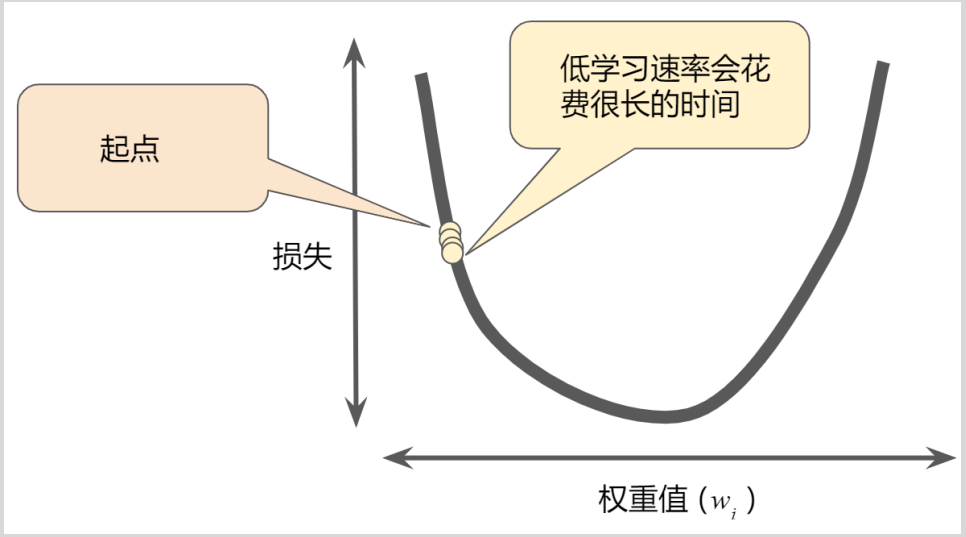
损失函数梯度

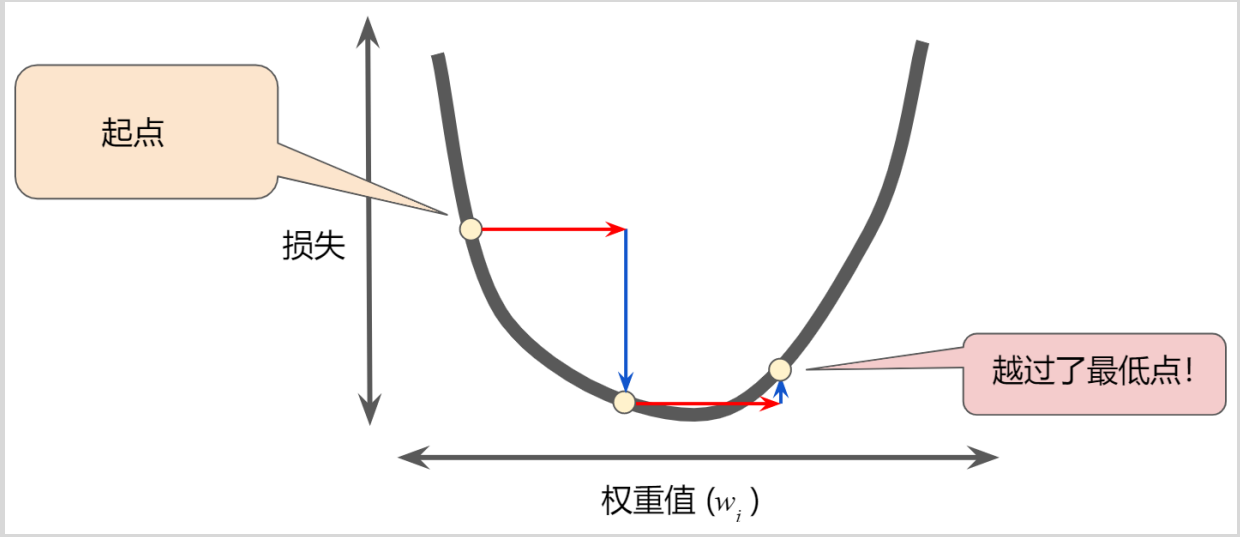
梯度下降法会重复此过程，逐渐接近最低点。

学习速率

https://developers.google.com/machine-learning/crash-course/fitter/graph

- 从梯度下降公式可以看出，学习速率α决定了模型参数调整的步长。





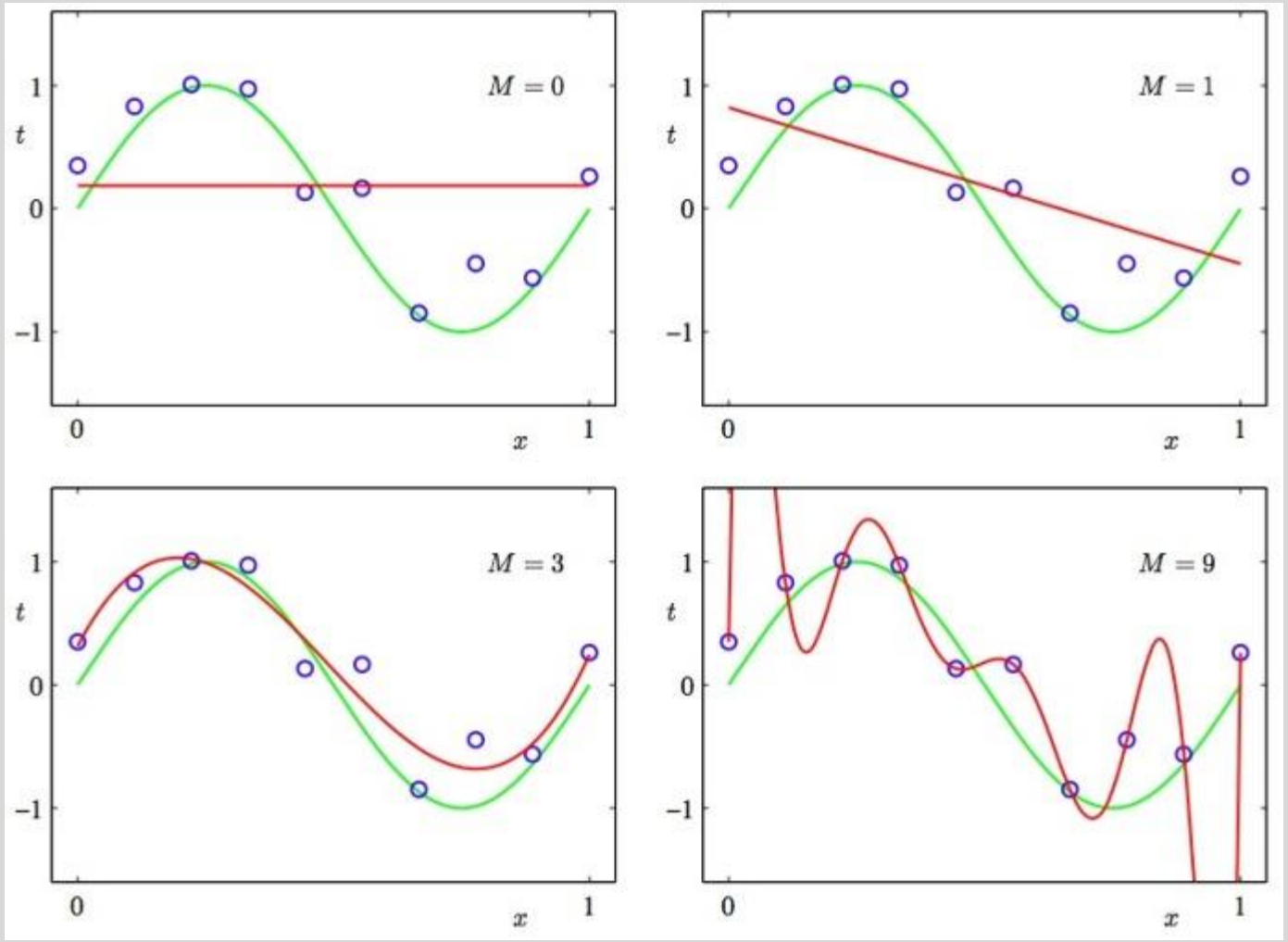
学习速率过小——训练时间延长

学习速率过大——难以到达最小值

因而在训练过程中需要选取适当的学习速率，以较高效地完成训练任务。

异常拟合

- 欠拟合（Underfitting）：模型没有很好地捕捉到数据特征，不能够很好地拟合数据，对训练样本的一般性质尚未掌握。
- 过拟合（Overfitting）：模型对训练样本学习过于精细，可能导致把一些训练样本自身的特性当做了所有潜在样本都有的一般性质，导致泛化能力下降。

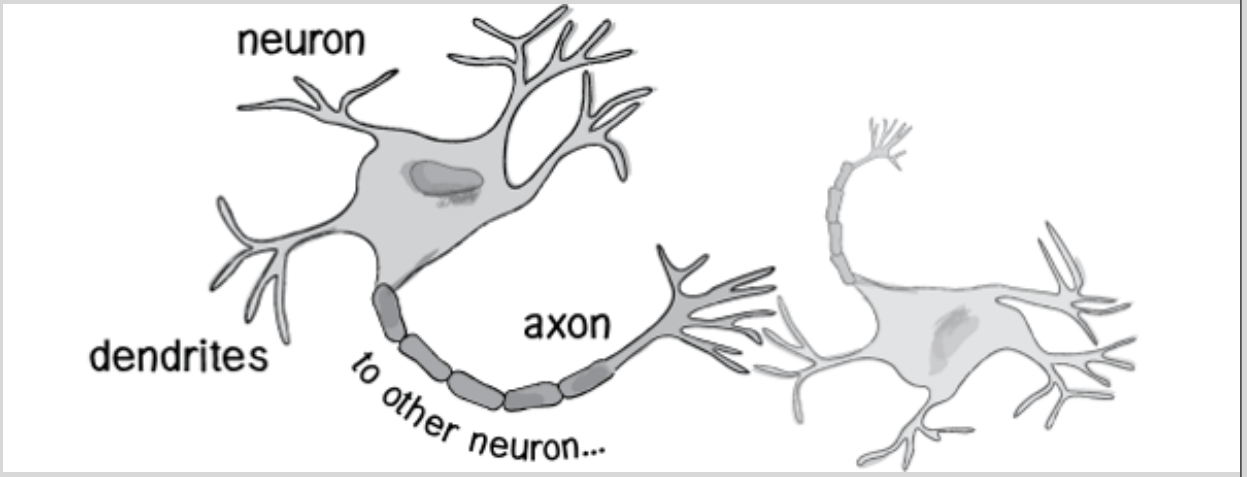


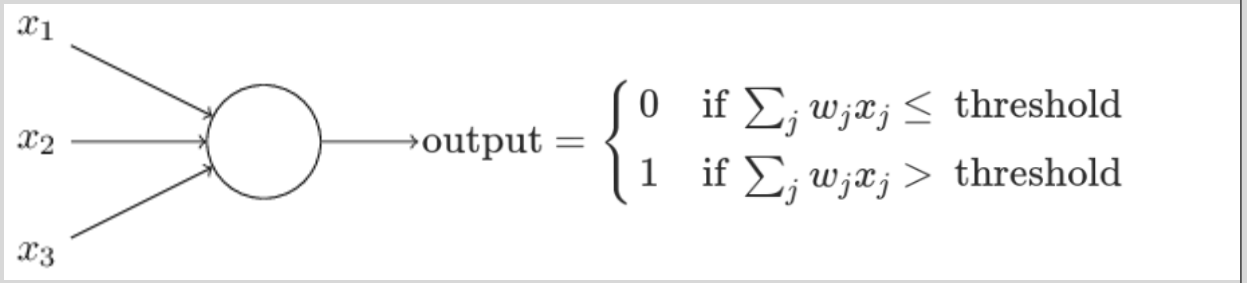
神经网络

神经元与感知器

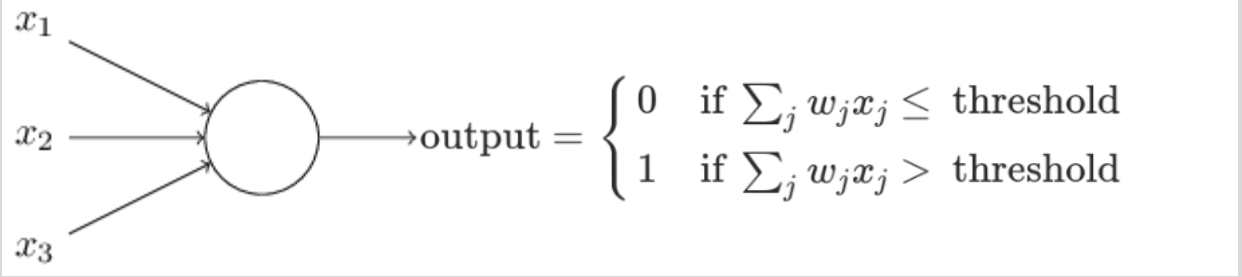
- 输入: x_1, x_2, x_3
- 权重: w_1, w_2, w_3
- 阈值: threshold

- 根据每个输入及其权重得到输出，根据阈值确定输出0还是1，进行决策。





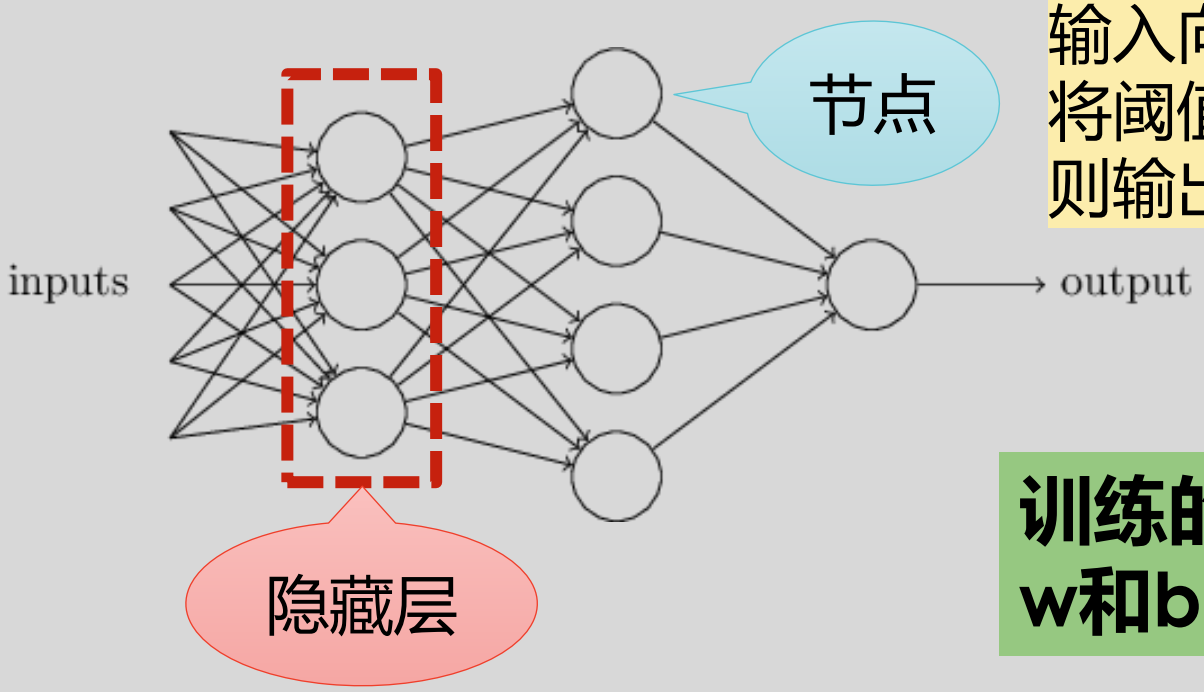
举个例子



- 假如你是李华，这周末你计划去天竺山登山，但周末能否成行取决于以下因素。
  1. 天气：周末是否是晴天?
  2. 同伴：能否找到人一起去?
  3. 交通：前往天竺山的道路是否拥堵?
- 综合考虑以上因素的过程便类似于感知器进行决策的过程，即根据每种因素及其所占的权重，若达到某一阈值，输出1（去），否则输出0（不去）。

多层感知器结构——人工神经网络

- 将多个感知器连结起来形成网络，即构成了神经网络 (Neural Network)。
- 每个节点的输入来自于上一层的输出，而输出又成为下一层的输入。

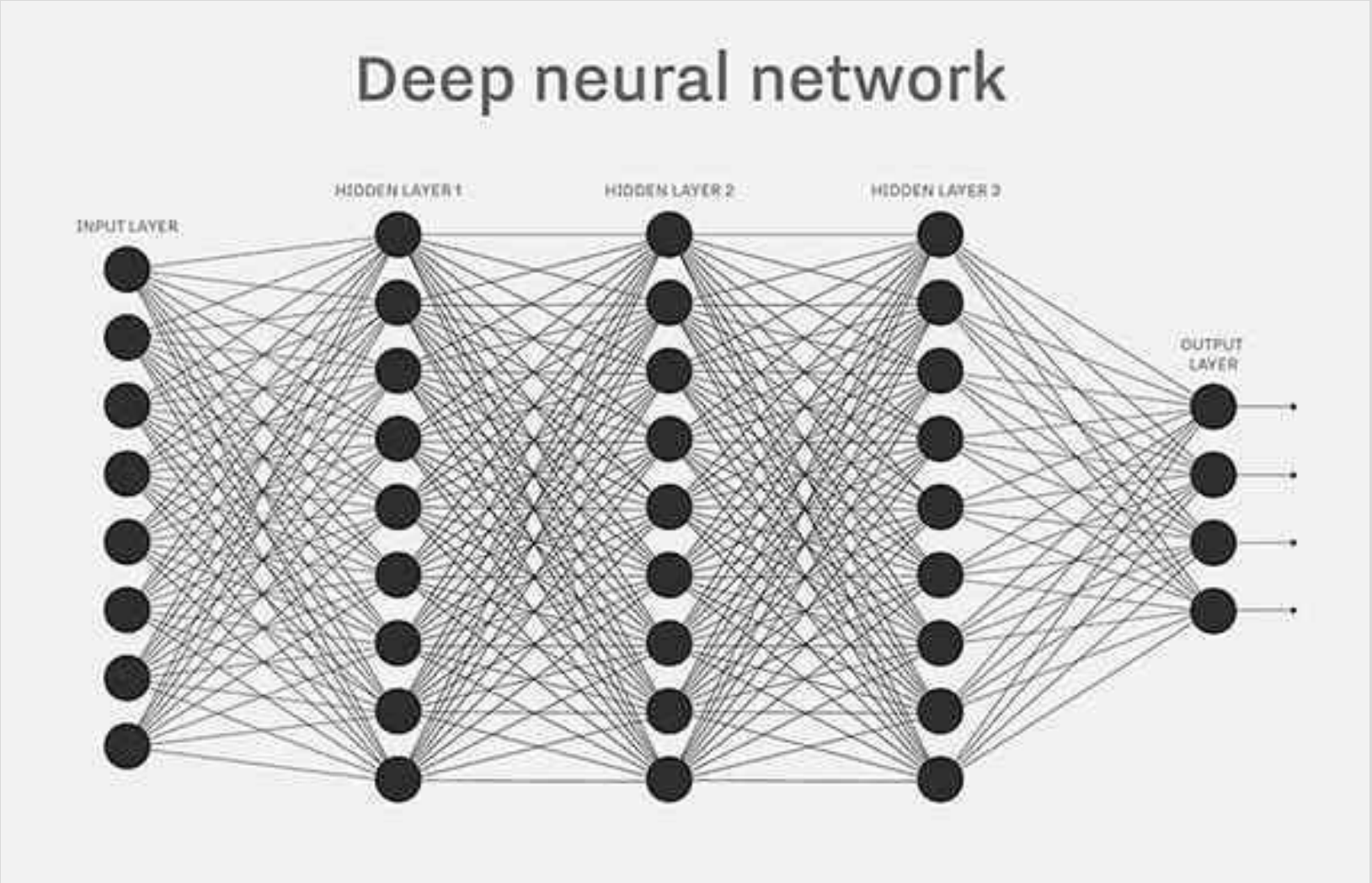


向量化：
输入向量定义为$\boldsymbol{x}$，权重向量定义为$\boldsymbol{w}$，
将阈值移项并定义$b = -threshold$，
则输出为$\boldsymbol{z} = \boldsymbol{w} \cdot \boldsymbol{x} + \boldsymbol{b}$

去线性化：引入**激活函数**

训练的过程即是调整每个隐藏层的w和b，使得损失函数降到最低

深度神经网络



https://playground.tensorflow.org/

谢谢