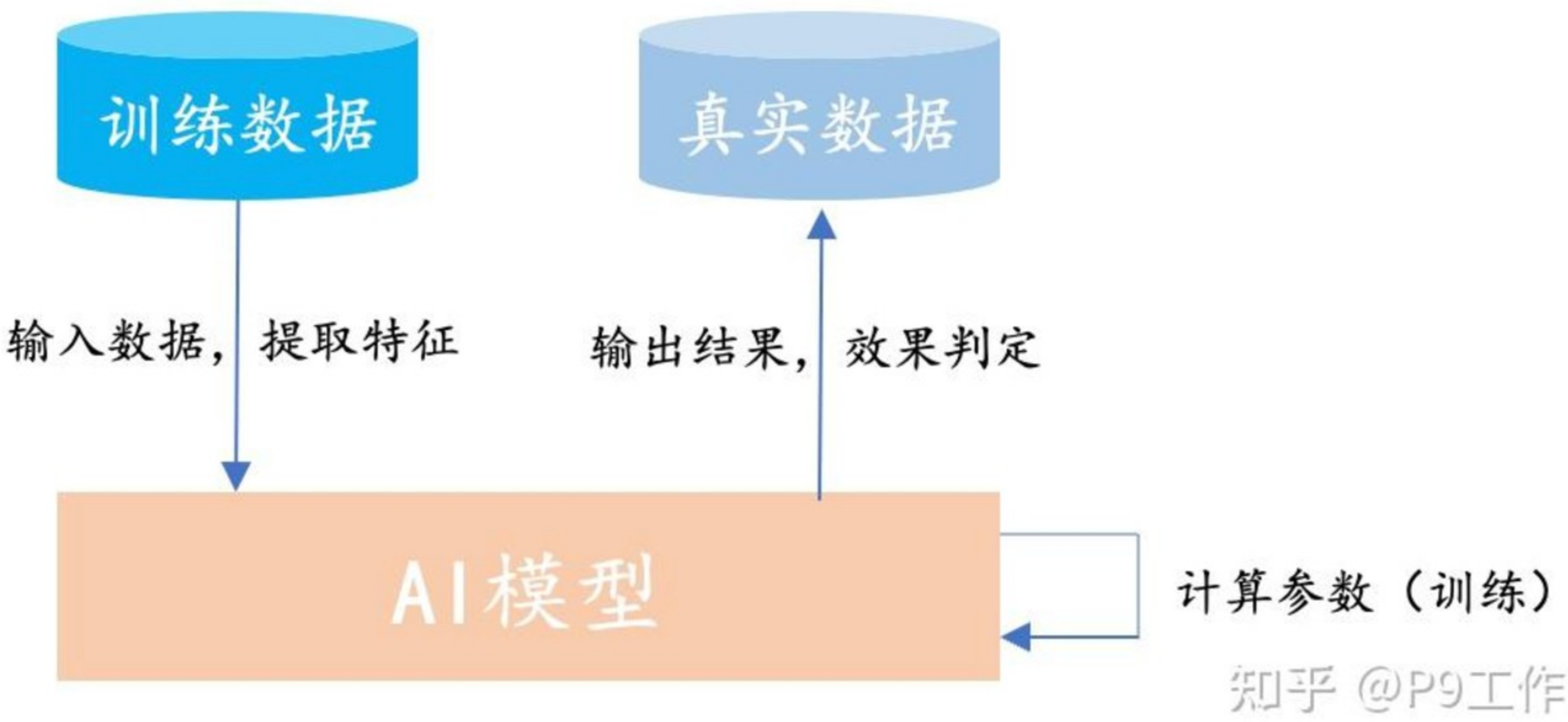
如何构建神经网络

P9工作法
分布式架构、上百人团队管理、全球支付实践与AI技术

已关注

17 人赞同了该文章

既然AI的本质是函数（详见上一篇），那么如何能够得到这个函数就是关键点。回顾上一篇的小结，AI的构建过程是通过数据训练（计算)得到参数，用参数再结合现有数据计算得到结果。如下图所示：



AI模型的构建过程

把这个过程进一步拆解：

1、首先是训练数据的处理，这是AI最重要的一部分，有数据才可能得到AI模型，有了高质量的数据才可能有好的AI模型，才能够更节省计算资源去获得这样一个AI模型。

2、其次是AI模型的选取，即用什么样函数来拟合这些数据是最有效的，有些简单问题用直线就好了，有些复杂问题要用曲线甚至是空间。 虽然神经网络具备通用近似定理的能力，但如何搭建这个模型仍然值得考究，比如有多少个神经元，多少个隐藏层等等。

3、最后是如何计算参数，也就是模型的训练。如果模型用函数表达为

$$y = W_1x_1 + W_2x_2 + W_3x_3 + W_4x_4 + W_5x_5 + \ldots W_nx_n + b$$

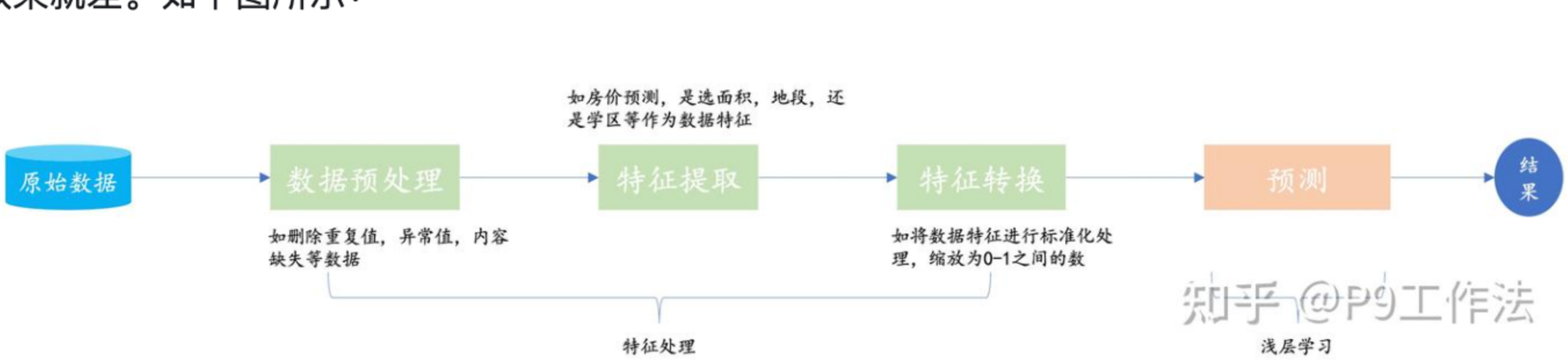
即如何求得这些 $W_1, W_2 \ldots W_n$ 。

接下来将逐个分析上面三个步骤

数据处理

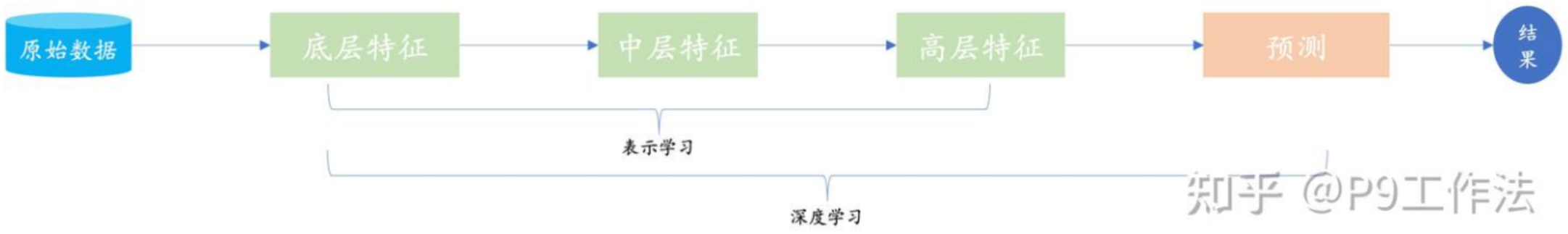
俗话说garbage in,garbage out，如果低质的数据给到神经网络肯定什么都得不到。但神经网络需要的数据并非是传统机器学习模型的特征数据，而是原始数据。

比如在逻辑回归模型中，是需要对原始数据进行特征提取的，也就是要按照专家经验去提取可能有用的数据，然后交给模型去学习找规律，特征工程做得越好效果也就越好，如果特征工程做得差那效果就差。如下图所示：



传统机器学习

但神经网络就不再需要进行特征抽取，而是端到端的，如下图所示：



深度学习过程

至于从底层特征到中层特征，再到高层特征可以这样理解：加入神经网络看到一张数字图片，先把学习到的横、竖，小圆弧看成是底层特征，中层特征可以认为是小圆弧组成的半圆，横折（横与竖的组合），高层特征可以认为是半圆组成的0，那这样这些特征就可以复用了，比如8就是两个0的组合，7就是横折再加一个竖等等。

而这些数据的特征都被神经网络模型自动学习到（可以简单理解为底层、中层、高层的特征提取就是通过多个隐藏层来实现），不再需要依赖人工，这简直是超级的进步。为什么说AI时代，对应用架构师来说是有巨大的机会，这是其中一个原因。应用架构师不需要有深厚的特征工程的经验，就能够用上AI。

这也解释了为什么神经网络是属于深度学习的一个原因。就是因为通过多层网络结构不断自动提取特征，最后给到预测函数去做预测。

AI模型

模型从本质上来说就是函数，要找到一个很好的函数去拟合这些数据。在神经网络中，其实就是要去考虑这个模型怎么搭建，输入层多少个参数，多少层隐藏层，输出层用什么等。这是对模型的一

个方面理解，很多公司都在比谁的模型做得大，就是在拼这方面的内容。但要把模型做大肯定不是调整这几个超参数的事，还有一系列的工程配套支持。

另外一方面的理解是，为了解决不同类型的问题，还需要不同的模型来应对。比如需要用卷积神经网络来处理图片，用RNN来处理语言。这可以理解为神经网络是最基础组件，需要根据场景再做一些应用层研发。如卷积神经网络就是在神经网络前面针对信息输入做一个卷积操作。

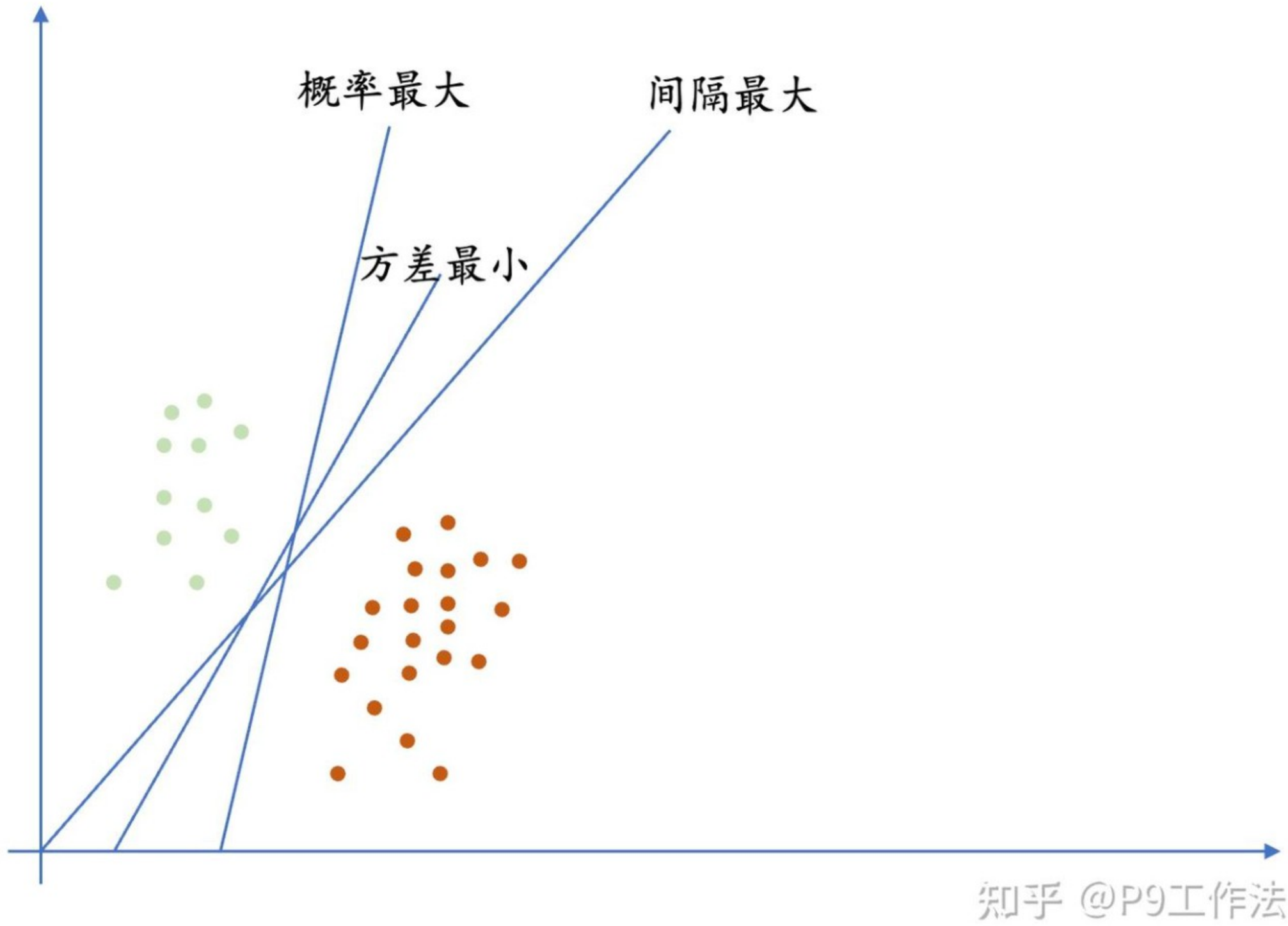
在AI模型中还有非常多优秀的模型，如transformer,bert,gpt等，这些都是基于神经网络这个基础组件扩展而来的场景化模型。

参数计算

如何评判

在AI领域，说一个模型好不好常说这个模型的性能怎么样，其实性能表达了两个意思，一个是准不准，一个是推理（也就是计算）得快不快。关于准不准怎么评判呢，其实也好理解，就是把模型输出结果与真实的结果（在模型训练时，不仅仅是输入训练数据，还会有结果数据）进行比对，把误差表达为一个函数，这就是损失函数。

以之前的预测房价的为例，真实值与预测值的误差有以下图三种：



- 一个是真实值与函数值的方差最小，精确度较高。
- 二是预测值出现在函数范围的概率最大，代表预测值归属于真实值的概率情况。比如求解房价大于500万的概率。
- 三是真实值与函数值的间隔越大越好，代表预测的包容性越强，但未必精确。

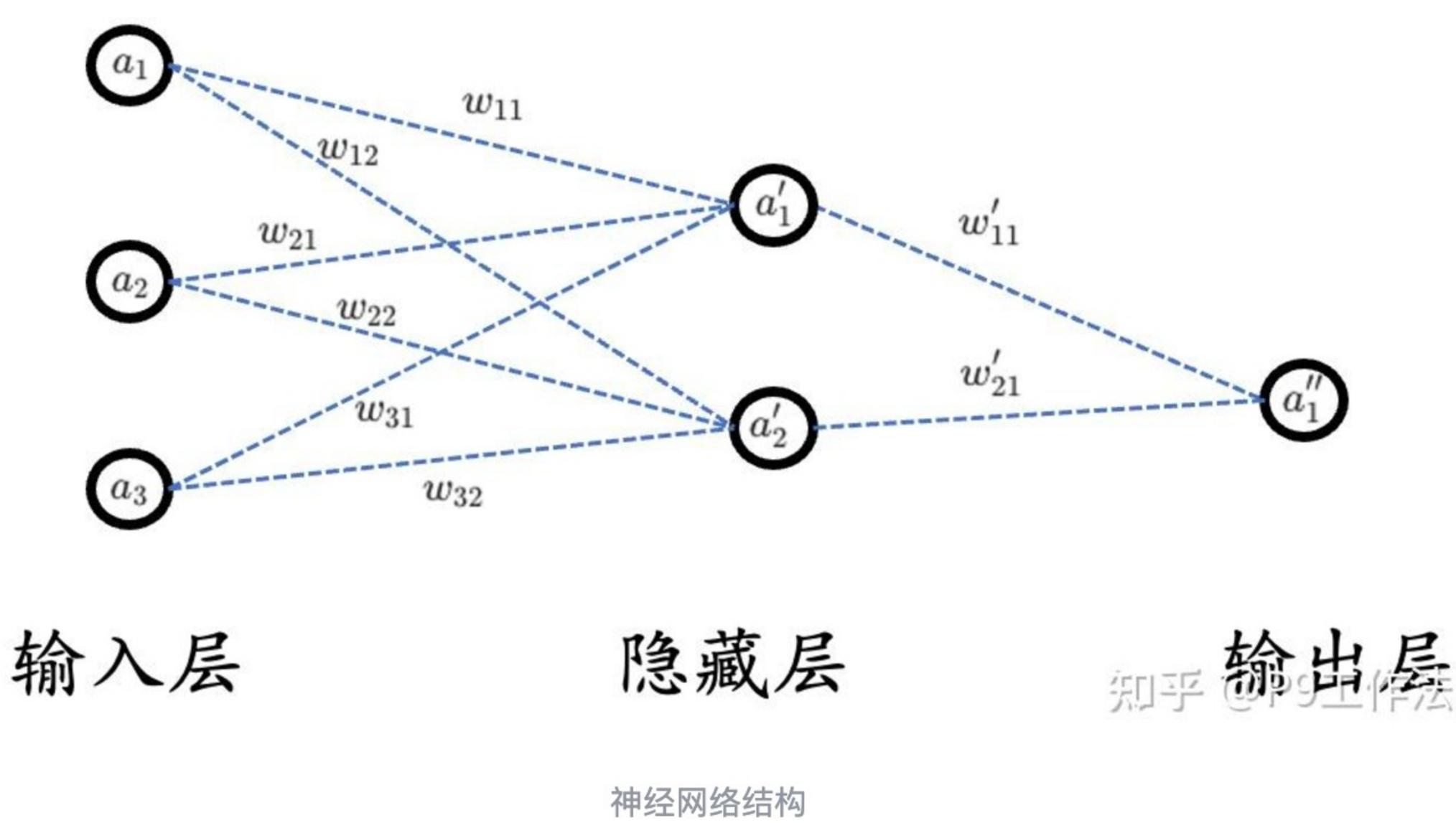
这三种判断方法，也叫做最终效果的策略，未必哪一个一定好，不同的问题用到的策略就不一样。比如你希望最终的准确率高那么选择间隔大的就行，但这有可能失去判断价值。

损失函数是AI中最重要的概念，只有写出来了损失函数才能够指导模型去调整参数，也是模型终止计算的条件（比如我的误差小于某个极小值就可以终止训练）。

如何计算

得到损失函数不是目的，目的是通过损失函数去求得这个函数的最优解，那么就能够得到这些参数的值。所以优化理论也是神经网络中非常重要的一个分支（这里包括非凸函数的优化理论，是一个蛮深的学科）。

神经网络的参数其实是神经元之间的连接（这里还省略了偏执参数b），如下图所示：



神经网络结构

数据是从左边输入，然后经过复杂的计算到隐藏层（隐藏层可能是一个或者多个），然后隐藏层再计算得到输出结果。

加入输入的信息为猫的图片信息，为了简化理解假设猫的图片信息是数字1,2,3，分别对应为输入层的三个神经元 a_1, a_2, a_3 接收到的数据。

那么 a'_1 的值应该为 $a'_1 = a_1 * w_{11} + a2 * w_{21} + a_3 * w_{31}$ ，此处省略偏执参数b。

a_2' 的值应该为 $a_2' = a_1 * w_{12} + a2 * w_{22} + a_3 * w_{32}$

最后一个神经元 a_1'' 的值为 $a_1'' = a_1' * w_{11}' + a'2 * w_{21}'$

假设模型在最开始就把参数初始化为0到1之间的小数，那最终应该会计算得到 a_1'' 的值，假设为0.2。我们简化地理解这是概率为猫的概率为0.2，便判定此图片不是猫。这便得到了预测数据与真实数据的误差。需要根据某些数学方法来更新这些参数，至于如何更新这些w系列的参数，这就是优化算法需要涵盖的范畴。这会放到下一章的反向传播来继续讲解。