机器学习的分类

1. 监督学习，训练数据中，每个样本都带有正确答案。
2. 无监督学习，训练数据中，每个样本都没有正确答案。
3. 半监督学习，训练数据中，少部分样本有正确答案。
4. 强化学习。

维数灾难及避免

当数据的特征值维度增大时，对应的特征值空间的样本数据将会呈现指数级增长。

通过特征抽取与特征选择进行避免。

机器学习算法

决策树

我曾经在暑假期间学过一个图论问题，求解有向图中从某点出发到所有点的必经点。

学习到了支配树这一算法，与决策树很类似，必经点就像是决策树上的决策点。

朴素贝叶斯分类

一类简单的概率分类器，基于贝叶斯定理和特征间的强大的独立假设。

应用识别垃圾邮件，新闻分类。

逻辑回归

一种统计学方法，通过使用逻辑函数来估计概率，从而衡量类别依赖变量和一个或多个独立变量之间的关系。

应用于信用评分。

K均值

K临近算法，曾经实现过一个二维的最近点对算法，平面最近点对问题。

随机森林

为了解决决策树的过拟合问题而诞生，本质上是若干决策树构成的森林。

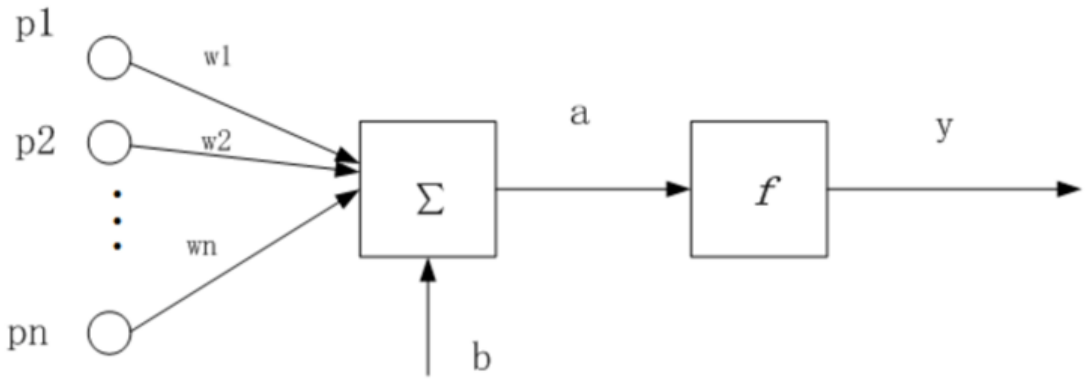
欠拟合与过拟合

“欠拟合”常常在模型学习能力较弱，而数据复杂度较高的情况出现，此时模型由于学习能力不足，无法学习到数据集中的“一般规律”，因而导致泛化能力弱。

“过拟合”常常在模型学习能力过强的情况中出现，此时的模型学习能力太强，以至于将训练集单个样本自身的特点都能捕捉到，并将其认为是“一般规律”，同样这种情况也会导致模型泛化能力下降。

区别：欠拟合在训练集和测试集上的性能都较差，而过拟合往往能较好地学习训练集数据的性质，而在测试集上的性能较差。在神经网络训练的过程中，欠拟合主要表现为输出结果的高偏差，而过拟合主要表现为输出结果的高方差。

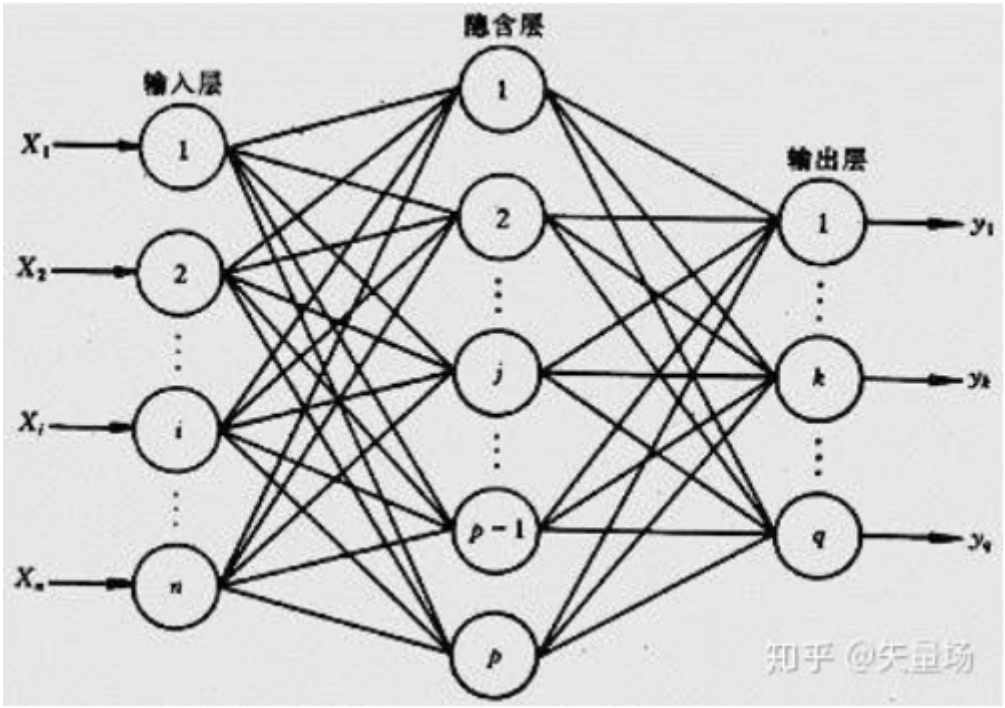
感知器模型



p_i为第i个输入，w_i为第i个输入的连接权值，$\sum$表示求和，f为一种函数，y为输出值。

BP神经网络

图源知乎水印：



输入层神经元读取输入数据x_i，同层神经元不互相连接，非同层神经元之间才有连接，感觉这里可以类比二分图。将数据通过输入层送到隐含层后，经过矩阵乘法运算，将结果送到输出层予以输出。

参考自知乎用户：退乎。