

# 机器学习

- 机器学习常规套路：收集数据并给定标签；训练一个分类器；测试，评估
- 机器学习流程：数据获取,特征工程（特征工程的作用：数据特征决定了模型的上限；预处理和特征提取是最核心的；算法与参数选择决定了如何逼近这个上限）,建立模型,评估与应用

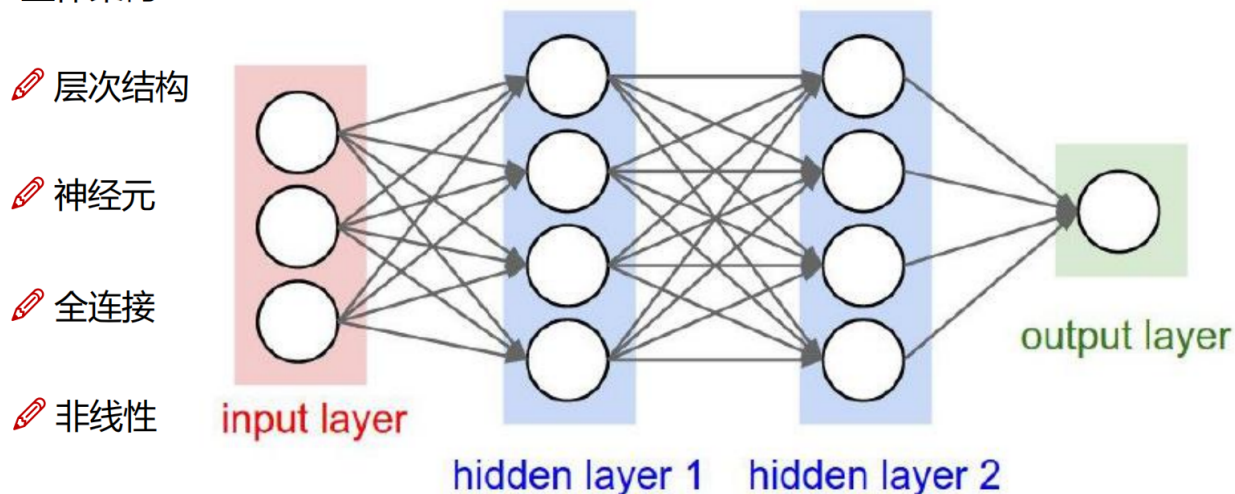
## k近邻

- 近邻计算流程:
  - 1.计算已知类别数据集中的点与当前点的距离
  - 2.按照距离依次排序
  - 3.选取与当前点距离最小的K个点
  - 4.确定前K个点所在类别的出现概率
  - 5.返回前K个点出现频率最高的类别作为当前点预测分类

## 神经网络基本概念与应用

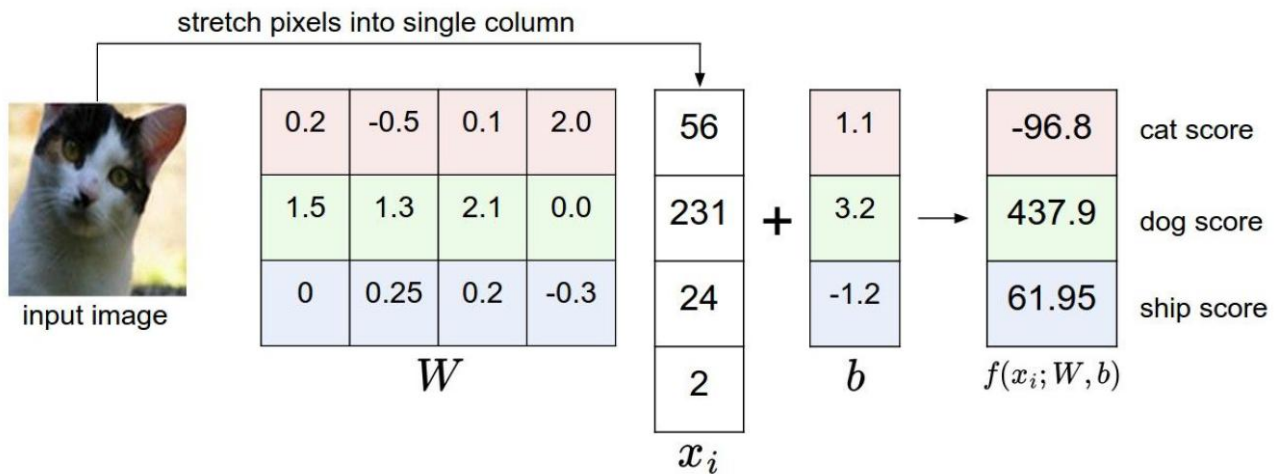
1. 神经网络是深度学习中的核心模块，用于处理分类、回归等任务。
2. 神经网络与深度学习的区别在于端到端(end2end)的处理方式，简化了中间过程。

✓ 整体架构



## 神经网络基础与工作原理

1. 神经网络通过矩阵乘法处理输入数据，输出分类概率。



2. 权重参数的初始化对模型训练有影响，可以通过预训练模型进行初始化。

$$f(x_i, W, b) = Wx_i + b$$

3. 损失函数用于指导模型学习，通过反向传播算法调整权重参数。

- 损失函数 = 数据损失 + 正则化惩罚项

$$L = \frac{1}{N} \sum_{i=1}^N \sum_{j \neq y_i} \max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1) + \lambda R(W)$$

## 神经网络优化与正则化

1. 正则化通过惩罚权重参数过大来防止过拟合，提高模型泛化能力。
2. 正则化包括L1正则化和L2正则化，分别对权重参数的大小进行不同的惩罚。
3. 数据预处理和特征选择对模型性能有重要影响，包括标准化、归一化等操作。

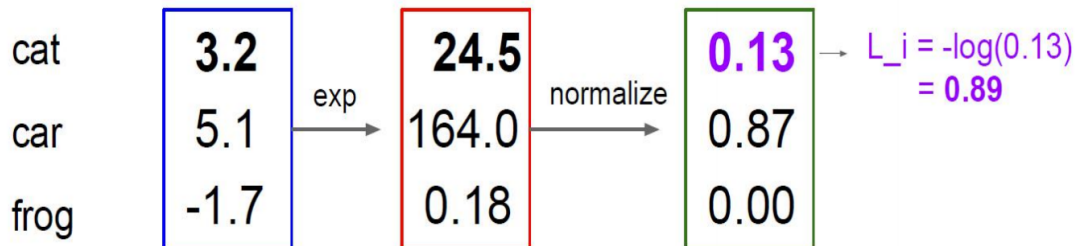
## 神经网络激活函数与损失函数

1. 激活函数如ReLU、Sigmoid等用于引入非线性，增强模型表达能力。

✓ Softmax分类器

✎ 归一化:  $P(Y = k|X = x_i) = \frac{e^{s_k}}{\sum_j e^{s_j}}$  where  $s = f(x_i; W)$

✎ 计算损失值:  $L_i = -\log P(Y = y_i|X = x_i)$



2. 损失函数如交叉熵损失用于分类任务，均方误差损失用于回归任务。

3. 损失函数的设计根据具体任务和数据特点进行选择和优化。

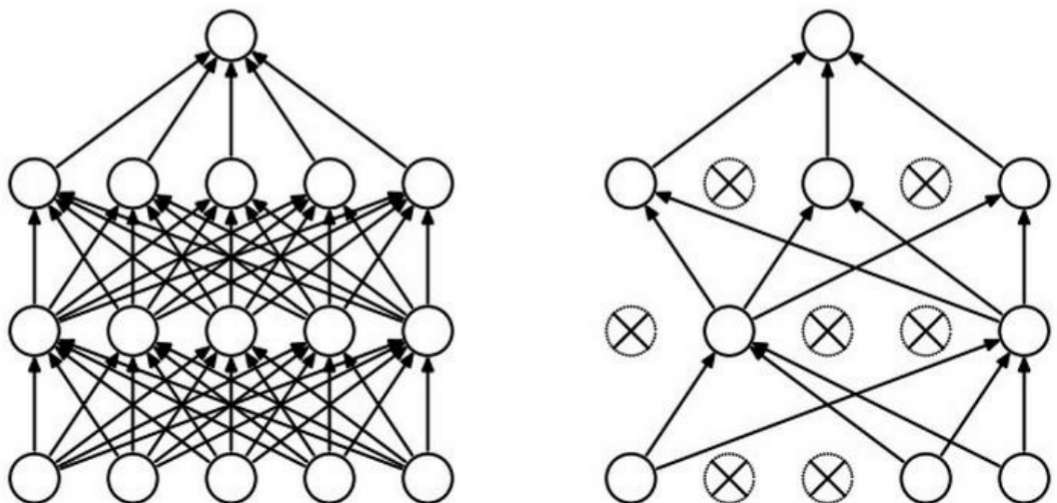
## 神经网络训练技巧与优化器

1. 优化器如Adam用于调整权重更新步长，影响模型收敛速度和稳定性。

2. 训练技巧包括学习率调整、梯度累积等，用于提高训练效率和模型性能。

✓ DROP-OUT (传说中的七伤拳)

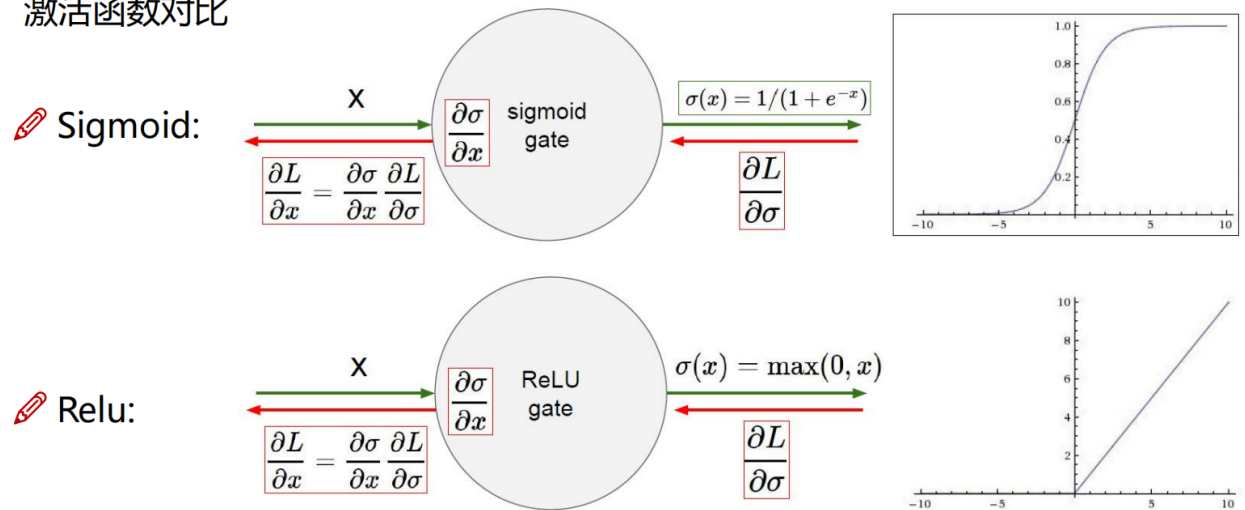
✎ 过拟合是神经网络非常头疼的一个大问题!



## 神经网络结构设计与特征提取

1. 神经网络结构设计包括选择合适的网络层数、激活函数和损失函数。

### ✓ 激活函数对比



2. 特征提取是神经网络的核心任务，通过学习输入数据的特征表示来解决问题。

3. 不同的网络结构适用于不同的数据和任务，通过尝试和调整找到最优结构。