简答题：

①监督学习（数据集有输入和输出数据）：通过已有的一部分输入数据与输出数据之间的相应关系。生成一个函数，将输入映射到合适的输出，比如分类。

无监督学习（数据集中只有输入）：直接对输入数据集进行建模，比如聚类。

半监督学习：综合利用有类标的数据和没有类标的数据，来生成合适的分类函数。

②训练集:顾名思义指的是用于训练的样本集合,主要用来训练神经网络中的参数.

验证集:从字面意思理解即为用于验证模型性能的样本集合.不同神经网络在训练集上训练结束后,通过验证集来比较判断各个模型的性能.这里的不同模型主要是指对应不同超参数的神经网络,也可以指完全不同结构的神经网络.

测试集:对于训练完成的神经网络,测试集用于客观的评价神经网络的性能.

数据集分为训练数据和测试数据。测试数据即为测试集，是需要应用模型进行预测的那部分数据，是机器学习所有工作的最终服务对象。为了防止训练出来的模型只对训练数据有效，一般将训练数据又分为训练集和验证集，训练集用来训练模型，而验证集一般只用来验证模型的有效性，不参与模型训练。

在有监督的分类模型中，训练集和验证集都是事先标记好的有标签数据，测试集是无标记的数据。在无监督模型中，训练集、验证集和测试集都是无标记的数据。

③归一化的目的就是使得预处理的数据被限定在一定的范围内（比如[0,1]或者[-1,1]），从而消除奇异样本数据导致的不良影响。

1. 提升模型的收敛速度

2.提升模型的精度。从经验上说，归一化是让不同维度之间的特征在数值上有一定比较性，可以大大提高分类器的准确性。

3. 深度学习中数据归一化可以防止模型梯度爆炸。

④泛化能力（generalization ability）是指机器学习算法对新鲜样本的适应能力，简而言之是在原有的数据集上添加新的数据集，通过训练输出一个合理的结果。

欠拟合（under-fitting），是指模型拟合程度不高，数据距离拟合曲线较远，或指模型没有很好地捕捉到数据特征，不能够很好地拟合数据。

过拟合（over-fitting），是指模型在训练集上表现很好，在测试集上效果差。

选择填空：

# ①减少过拟合的方法：

# 增加样本数据量 2. 数据增强，人为扩展数据量 3. 加入正则化项 4. Dropout 5. early stopping 6. batch normalization

# 决策树剪枝

# L1L2正则、Dropout、early stopping

# ②激活函数

# 什么是激活函数

# 激活函数（Activation functions）对于人工神经网络模型去学习、理解非常复杂和非线性的函数来说具有十分重要的作用。它们将非线性特性引入到我们的网络中。如图1，在神经元中，输入的 inputs 通过加权，求和后，还被作用了一个函数，这个函数就是激活函数。引入激活函数是为了增加[神经网络模型](https://baike.baidu.com/item/%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C%E6%A8%A1%E5%9E%8B/10658952" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)的非线性。没有激活函数的每层都相当于矩阵相乘。就算你叠加了若干层之后，无非还是个矩阵相乘罢了。

# 为什么要用激活函数

# 如果不用激活函数，每一层输出都是上层输入的线性函数，无论神经网络有多少层，输出都是输入的线性组合，这种情况就是最原始的[感知机](https://baike.baidu.com/item/%E6%84%9F%E7%9F%A5%E6%9C%BA/12723581" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)（Perceptron）。

# 如果使用的话，激活函数给神经元引入了非线性因素，使得神经网络可以任意逼近任何[非线性函数](https://baike.baidu.com/item/%E9%9D%9E%E7%BA%BF%E6%80%A7%E5%87%BD%E6%95%B0/16029251" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)，这样神经网络就可以应用到众多的[非线性模型](https://baike.baidu.com/item/%E9%9D%9E%E7%BA%BF%E6%80%A7%E6%A8%A1%E5%9E%8B/10463547" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)中。

1. 常见激活函数

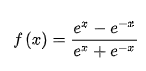
# 1）Sigmoid函数

# [Sigmoid函数](https://baike.baidu.com/item/Sigmoid%E5%87%BD%E6%95%B0/7981407" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)是一个在生物学中常见的S型函数，也称为[S型生长曲线](https://baike.baidu.com/item/S%E5%9E%8B%E7%94%9F%E9%95%BF%E6%9B%B2%E7%BA%BF/5581189" \t "https://baike.baidu.com/item/%E6%BF%80%E6%B4%BB%E5%87%BD%E6%95%B0/_blank)。在信息科学中，由于其单增以及反函数单增等性质，Sigmoid函数常被用作神经网络的阈值函数，将变量映射到0,1之间 [2]  。公式如下：

捕获

# 2）Tanh函数

# Tanh是双曲函数中的一个，Tanh()为双曲正切。在数学中，双曲正切“Tanh”是由基本双曲函数双曲正弦和双曲余弦推导而来。公式如下：



# 3）ReLU函数

# Relu激活函数（The Rectified Linear Unit），用于隐层神经元输出。公式如下：

捕获

# ③决策树的分叉依据

④评估方法：

关键：怎么获得“测试集”(test set) ？测试集应该与训练集“互斥”。

常见方法：

1. 留出法 (hold-out)
2. 交叉验证法 (cross validation)
3. 自助法 (bootstrap)

交叉验证法：

⑤机器学习周期：

机器学习经过了三个阶段,分别是推理期,知识期和学习期。

⑥CNN计算：

⑦batch\_size最小为什么是2x：