模型选择

1. 模型误差分类

训练误差：模型在训练数据上的误差
泛化误差：模型在新数据集上的误差
训练误差小不代表泛化误差好，可能就是仅仅针对该数据集效果比较好

2. 训练数据集 (Training Dataset)：

训练数据集用于训练模型。通过这个数据集，模型通过学习数据中的规律，调整其参数（如权重）以提高其对数据的拟合度。训练数据集包含了输入特征和对应的标签，模型根据这些数据来“学习”如何做出预测。**train loss, train acc主要是用来计算训练误差和准确率，检验模型对数据的拟合情况**

3. 验证数据集 (Validation Dataset)：

验证数据集用于评估模型在训练过程中的表现。它不参与模型的训练过程，但用于在每次训练后对模型进行评估，帮助选择最优的超参数（如学习率、正则化系数等）。通常使用交叉验证方法，或者在训练过程中分出一个子集作为验证集。验证集可以帮助检测模型是否出现过拟合问题，并提供对模型泛化能力的初步判断。**验证集主要是用来检测模型的泛化性能，根据此来调整模型的超参数，防止过拟合**

验证数据集一定不能和训练数据集混在一起

4. 测试数据集 (Test Dataset)：

测试数据集用于在模型训练完成后评估模型的最终性能。测试集的数据没有在训练中使用，它用于模拟模型在未知数据上的表现。测试集的目的是衡量模型的泛化能力，并为模型的实际应用提供性能

指标。测试数据集应当在训练和验证过程完成后使用，并且只能使用一次。**测试集是评估模型最终性能的客观指标**

5. K-折交叉验证 (K-fold Cross Validation)：

交叉验证是用于评估模型性能的一种方法，通过将训练数据集分成K个相同的子集（折），**每次使用K-1个子集进行训练，剩余的一个子集用于验证模型**。重复K次，确保每个子集都能作为验证集进行一次。这种方法的优点是能够充分利用每个数据样本，提高模型的稳定性，并减少因数据划分导致的偏差。

- **K=5或K=10**：常见的K值设置，通常选取5折或10折交叉验证。K值过小可能导致评估不稳定，过大则计算成本较高。
  在K折交叉验证中，最终的评估结果通常是K次验证结果的平均值，用以更准确地评估模型的泛化能力。

6. 模型容量与拟合效果

模型容量/数据	简单	复杂
低	正常	欠拟合
高	过拟合	正常

欠拟合：欠拟合是指模型无法捕捉数据中的规律，导致在训练数据上的表现不佳，也无法在测试数据上取得良好的结果。通常发生在模型过于简单时，参数过少，模型的假设空间不足以拟合训练数据的复杂性。
过拟合：过拟合是指模型在训练数据上表现极好，但在测试数据上表现较差。模型在训练集上学得过于精细，以至于记住了训练数据中的噪声或偶然模式，而不是学习到数据的潜在规律。过拟合通常发生在模型过于复杂时，参数过多，能够记住训练数据的每一个细节，但这些细节并不具有普适性。

7. 模型容量与正则化

模型容量：模型拟合任意函数的能力

- **低容量模型**：如线性回归，参数少，拟合能力弱
- **高容量模型**：如深度神经网络，参数多，拟合能力强
  控制模型容量的因素：
- 模型参数个数
- 参数取值范围

数据复杂度的因素

- 样本数量
- 特征维度
- 样本结构与多样性

模型容量限制方法

- **使用均方范数来作为硬性限制**
  通过限制参数的值的选择范围来控制模型容量
  $min\, e(w,b)\ subject\, to\, ||w||^2 \leq \theta$
  通常不限制偏移b
- **使用均方函数作为柔性限制**
  $min\, e(w,b) + \frac{\lambda}{2}||w||^2$
  可以使用拉格朗日乘子来证明
  超参数λ控制了正则项的重要程度，$\lambda = 0$无作用，
  $\lambda \to \infty\, w^* \to 0$
  参数更新法则
  计算梯度

$$\frac{\partial(e(w,b) + \frac{\lambda}{2}||w||^2)}{\partial w} = \frac{\partial(e(w,b))}{\partial w} - \lambda w$$

时间t更新参数：

$$w_{t+1} = (1 - \eta\lambda)w_t - \eta\frac{\partial e(w_t, b_t)}{\partial w_t}$$

通常$\eta\lambda < 1$，每一次更新的时候先将w_t减小一次，这在深度学习中叫做权重衰退
权重衰退通过L2正则使得模型参数不会过大，从而控制模型复杂度，正则项权重λ是控制模型复杂程度的超参数

如何解决过拟合
调整λ参数，调整数据量大小，调整训练轮次等等

Dropout 扰动方法

Dropout只在训练中使用：影响参数的更新
一个好的模型需要对输入数据由扰动鲁棒
丢弃法·：在层之间加入噪音
对`x`增加噪音得到`x'`，我们希望$E[x'] = x$
丢弃法对每个元素进行如下扰动：

$$x_i' = \begin{cases} 0 & \text{with probablity p} \\ \frac{x_i}{1-p} & \text{otherwise} \end{cases}$$

dropout可以用来将隐藏**全连接层**的输出上
$h = \sigma(W_1 x + b_1)$
$h' = dropout(h)$
$o = W_2 h' + b_2$
$y = F(0)$
dropout丢弃法将一些输出项随机置为0来控制模型复杂度，通常作用在多层感知机的隐藏层输出上，丢弃概率是控制模型复杂度的超参数，控制模型自由度，减少**各个神经元的依赖性，强制学到更加普遍的特征减少过拟合**
每一次采样相当于在原神经网络中随机采样一个子网络，最终的模型是就相当于是多个子模型的集成平均
在推理的时候直接返回输入$h' = dropout(h)$