1.机器学习的一些概念：

(1) 监督学习：就是指训练数据的标签是已知的，在模型训练的时候，要将训练数据和标签数据相匹配，实现模型的训练。也就是用训练数据不断拟合标签，实现模型的监督学习。
(2) 无监督学习：无监督学习是指训练的模型的数据的标签是未知的。
泛化能力：训练好的模型在测试集上的适应能力，也就是对新样本的识别或预测能力，如果识别率高或者预测结果的误差小，就说明泛化能力高。
(3) 过拟合：训练好的模型对于训练过的数据的适应能力特别强，但是对于测试样本（也就是没有训练过的样本）的适应能力弱，此时就叫过拟合。过拟合带了的结果是低偏差，高方差。

解决方法：增加训练数据，降低模型复杂度（如减少神经网络的层数），加入正则化约束，神经网络的 dropout 策略，批归一化策略（具有轻微正则化效果），early-stopping 方法
(4) 欠拟合：如果模型在训练的时候，在训练集上的性能表现的不好的话，就叫欠拟合，此时成为高偏差

解决方法：新模型，新方法，新的网络，增大训练数据，归一化输入数据，更好的梯度优化算法
(5) **交叉验证：**一种统计学上将数据样本切割成较小子集的实用方法，因数据集的样本数有限，于是把在某种意义下将原始数据(dataset)进行分组,一部分做为训练集(train set),另一部分做为验证集(validation set or test set),首先用训练集对分类器进行训练,再利用验证集来测试训练得到的模型(model),以此来做为评价分类器的性能指标。常用的交叉验证是 K 折交叉验证，也就是把数据集平均分为 K 个部分，K-1 个部分作为训练集，另外一个部分作为验证，将 K 次训练得到的结果进行平均来评价模型的能力。

2.线性回归的原理：

我们经常说的 $y=ax+b$，就是一个简单的线性回归的思想，我们当时在学习这个函数的时候，如果给定方程上的两个不同的点，就可以联系方程组求得参数 a 和 b。
但是现在我们进一步对其扩展延伸到多参数的例子，也就是我们的自变量有很多个，因此我们的系数也是多维度的。通过这种线性模型，加上监督学习的训练数据集，通过训练迭代更新就可以得到相应的线性回归模型。

3. 线性回归的损失函数

损失函数：常常分为 0-1 损失函数，感知损失函数，平方和损失函数，绝对值损失函数，对数损失函数

线性回归的代价函数：在我的理解来函，损失函数就是一个定义，而代价函数才是应用，也就是说损失函数我们只是一个直观上的认知；但是代价函数面对的是大量的样本，也就是我们需要将所有训练数据的损失函数做一个权衡来优化最终的模型参数。比如，我们会将多个样本数据的每个损失函数最终做一个平均之后，作为代价函数。

线性回归的目标函数：均方差损失，$MSE=(\mathrm{y}_i-\hat{\mathrm{y}}_i)^2$

4. 优化方法：https://blog.csdn.net/v_JULY_v/article/details/81350035

5. 线性回归的评估指标：

（1）R 方评估指标，它是用 1 减去样本总体平方和与残差平方和的比值

（2）残差估计

总体思想是计算实际值与预测值间的差值简称残差。从而实现对回归模型的评估，一般可以画出残差图，进行分析评估、估计模型的异常值、同时还可以检查模型是否是线性的、以及误差是否随机分布。

（3）均方误差(Mean Squared Error, MSE)

均方误差是线性模型拟合过程中，最小化误差平方和(SSE)代价函数的平均值。MSE 可以用于不同模型的比较，或是通过网格搜索进行参数调优，以及交叉验证等。

（4）决定系数

可以看做是 MSE 的标准化版本，用于更好地解释模型的性能。换句话说，决定系数是模型捕获相应反差的分数。

6. Sklearn 详解

调用的包是：sklearn.linear_model.LinearRegression()

参数详解：

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression

https://blog.csdn.net/weixin_42451864/article/details/81352878

1、fit_intercept：bool 量，选择是否需要计算截距 b，默认为 True，如果中心化了的数据可以选择 false。

2、normalize：bool 量，选择是否需要标准化（中心化），默认为 false，和参数 fit_intercept 有关。

3、copy_x:bool 量，选择是否复制 X 数据，默认 True,如果否，可能会因为中心化把 X 数据覆盖。

4、n_job:int 量，选择几核用于计算，默认 1，-1 表示全速运行，也就是电脑里的线程数，越多的线程数处理能力更强。