样本方差：

错误：

欠拟合：模型没有很好地捕捉到数据特征，不能够很好地拟合数据。（模型太简单，偏差高）

过拟合：为了得到一致假设而使假设变得过度严格称为过拟合。（为了很好的拟合训练数据，而是模型太复杂，方差高，泛化能力差）

欠拟合：平时不好好学习，考试没考好。

好的模型：平时好好学习，考试考的不错

过拟合：平时一个字一个字记住了教科书，可以回答书上的任何问题，但没有泛化能力，无法通过考试。

交叉验证（Cross Validation, CV）：验证分类器的性能的统计分析方法。将原始数据进行分组，一部分作为训练集（train set），另一部分作为验证集（validation set）。用训练集对分类器进行训练，利用验证集来测试训练的模型。

测试集数据不参与交叉验证。

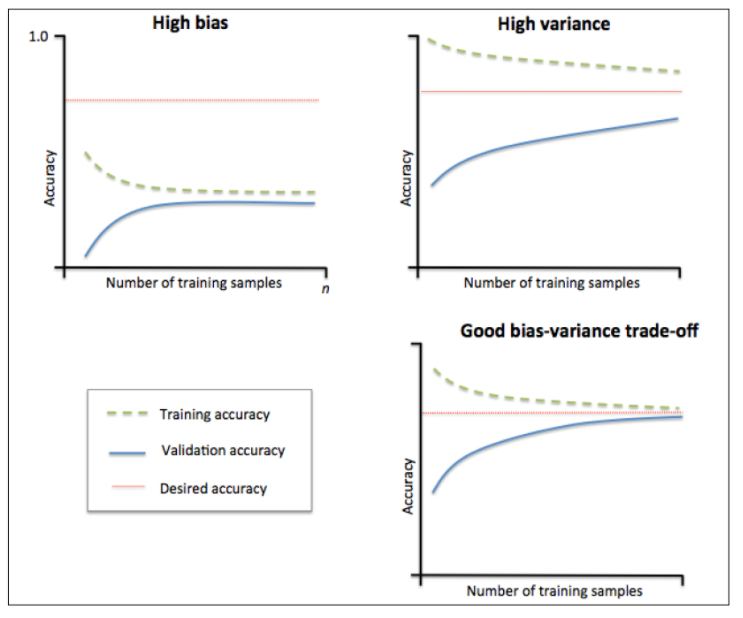
K折交叉验证：将原始数据分为K组，将每个子集分别做1次验证集，其余K-1个子集作为训练集，这样会得到K个模型，用这K个模型最终的验证集的分类准确率的平均数作为此K-CV下分类器的性能指标，常用的是10折交叉验证。

留一法交叉验证：假设原始数据有N个样本，每个样本单独作为验证集，其余N-1个样本作为训练集，这样就会得到N个模型。

优点：每个模型几乎都是所有样本都参与训练，得到的模型最接近原始样本分布；实验没有随机因素，可重复。

缺点：如果N非常大，则计算量非常大。

学习曲线：横轴为训练样本数目，纵轴为准确率（优达教程中为误差），有两条曲线（有时会有3条），分别为：不同样本数目下训练准确率曲线和验证准确率曲线。



左上角是高偏差模型，训练准确率和验证准确率会逐渐收敛到一起，但准确率都比较低，代表着欠拟合。对于欠拟合，可以增加模型参数，构建更多的特征，不能通过增加数据量来解决。

右上角为高方差模型，训练准确率很高，但验证准确率很低，二者不收敛，代表着过拟合。对于过拟合，可以增大训练集，降低模型复杂度，减少特征来解决。

网格搜索：利用穷举法尝试模型中参数的各种组合，从而得到最优的模型参数。

可以对多个参数（大于2）进行参数调优。

参考：<https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html#sphx-glr-auto-examples-model-selection-plot-grid-search-digits-py>

该例子对SVC的kernel，gamma和C同时进行搜索。

