# 错误类型-特点

模型过度简化问题时，称为 欠拟合

模型过度复杂化时，称为 过拟合

欠拟合 特点是训练集的拟合不够好 ,称这种类型为**偏差引起的误差**。

过拟合 特点是，在训练集中表现良好,但它**趋向于记住**而不是**学习数据特点**,因此在测试集上表现不好。称这种错误为**方差引起的误差。**

# 错误类型-理解

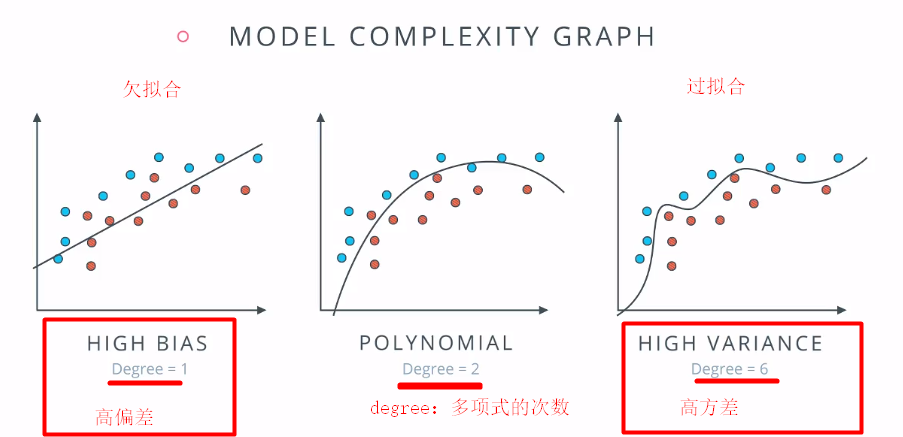
1.欠拟合想象成，在无准备的情况下去参加考试 无论你费多大的力 最终都考不好。

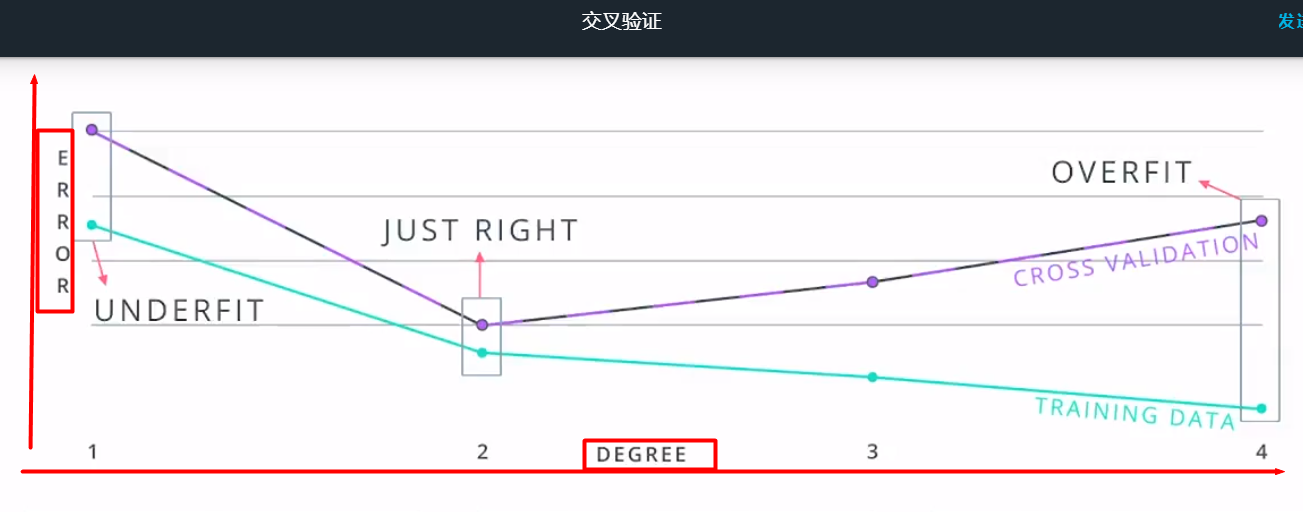
2.过拟合是，做了准备 但你没有去理解那些知识 只是将它一字一句地背下来了

你可以回答得上试卷上所有重复的问题 但遇到关于数据的新问题 就答不出来了

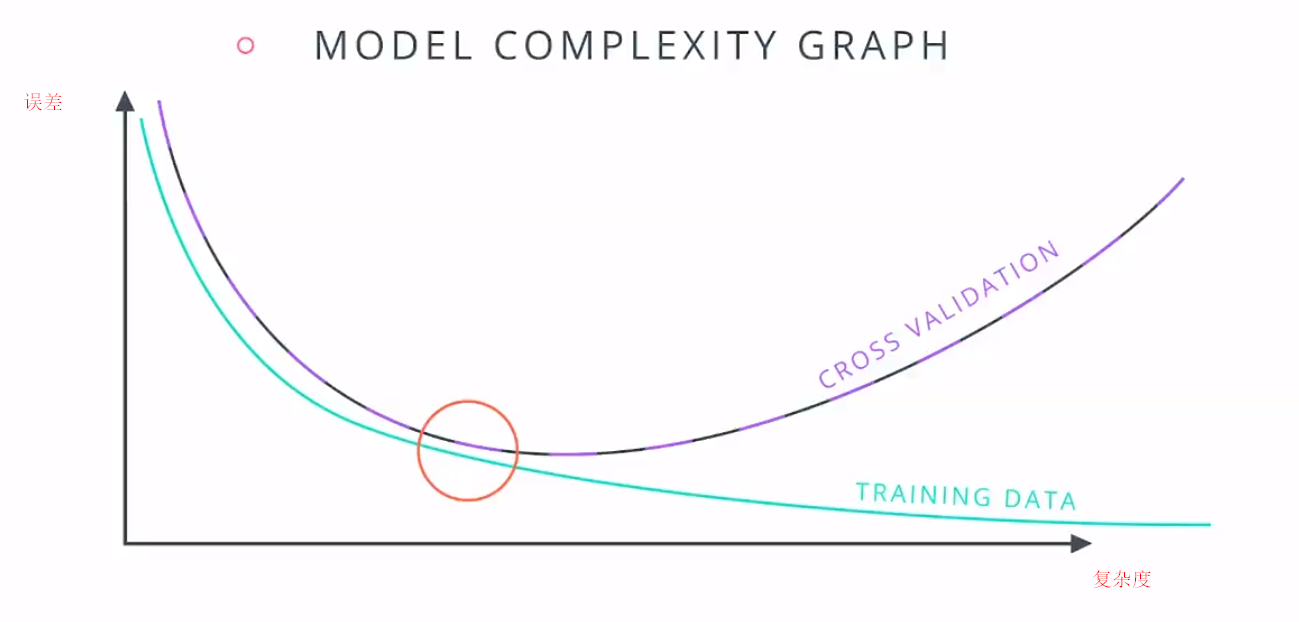
3.完美的模型就像是你学习了知识 并准备好回答试卷中遇到的任何新问题

# 模型复杂度图表





# 模型复复杂度一般规律



模型复杂度一般图形表现行为：

1.模型越复杂 训练误差越来越小。

2.测试误差开始很大 然后减小 然后再变大。

# k 折交叉验证

产生问题原因：

黄金法则：永远不能将测试数据用于训练。

也就是说，除非到最后一步，否则我们不应该使用测试数据，更不能利用它 决定 我们选择的模型

解决方法：

1、添加 交叉验证集

训练集将用于 训练参数

**交叉验证集(cross validation)** 用于对模型做出决定 例如多项式的次数

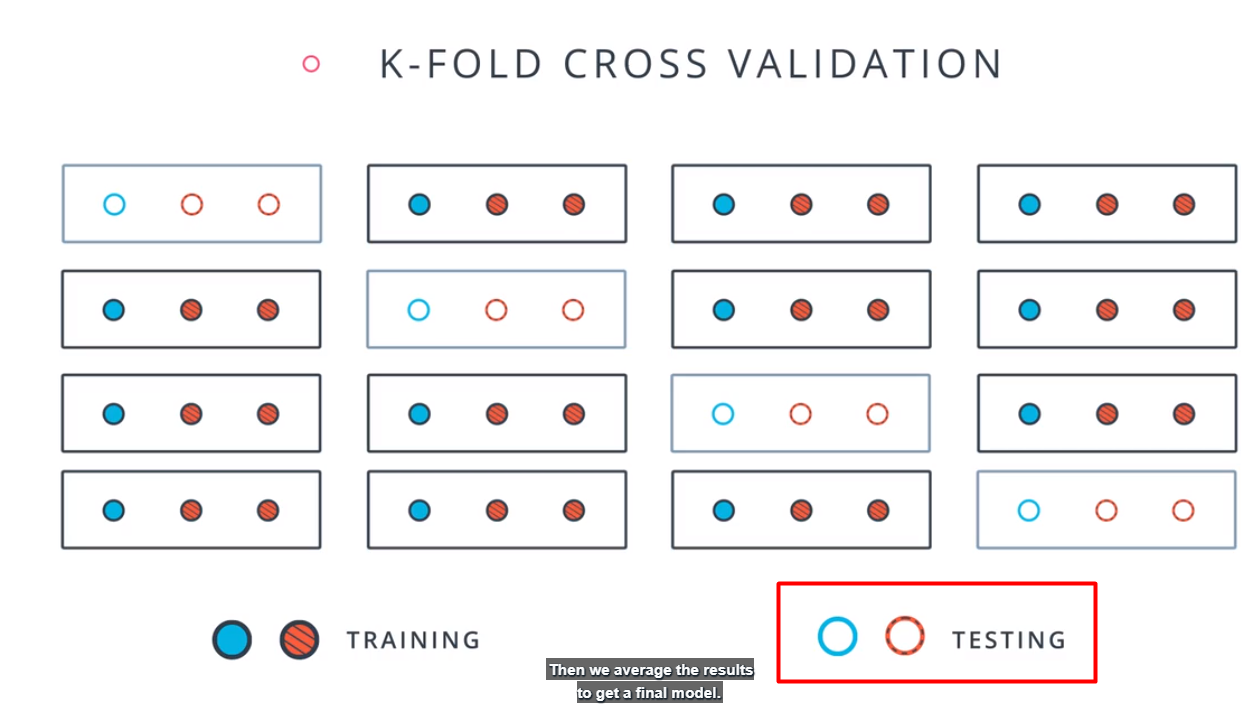
测试集用于 模型的最终测试

2、使用k 折交叉验证 方法

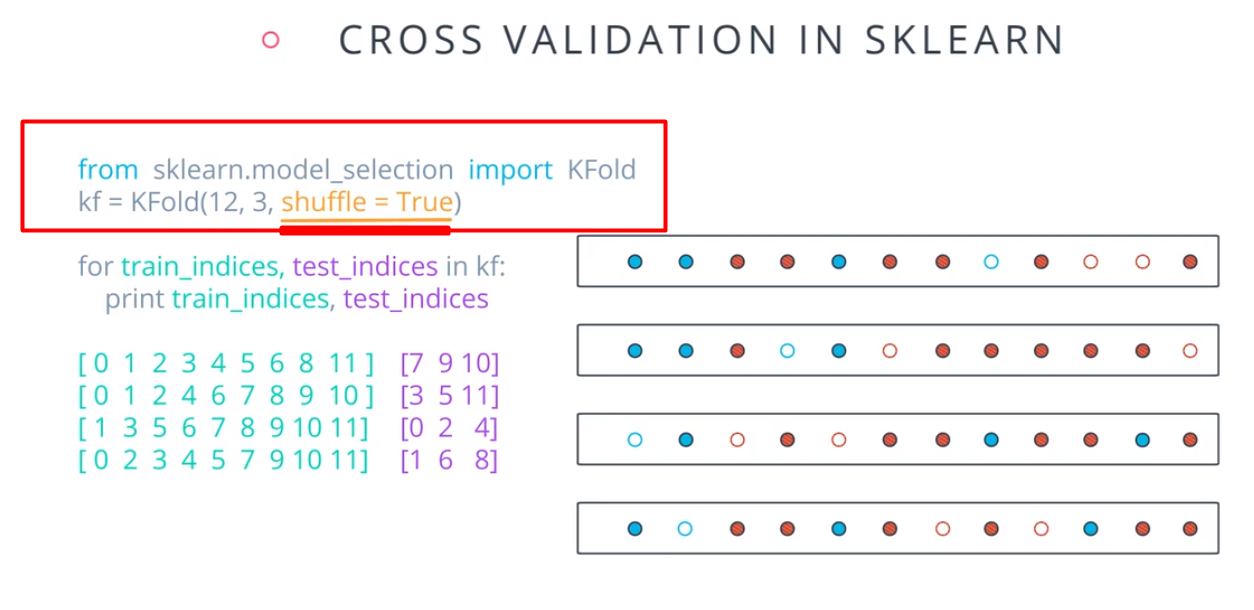
k 折交叉验证是非常有用的循环利用数据的方法。

在 K 折交叉验证中 我们将训练数据分为 K 个包（例如k=5），然后将模型训练 K 次 ,每次将不同的包用作测试集，而剩下的包作为训练集，最后我们求结果的平均值 来得到最终模型。

# 数据拆分图



Sklearn 实现



始终建议随机化我们的数据 以消除任何可能的偏差，这里我们仍然将数据拆分为包 但是是随机而非顺序选择

sklearn 中进行随机化也很容易 只需，在初始化 KFold 对象时将 shuffle 参数设为 true 即可。