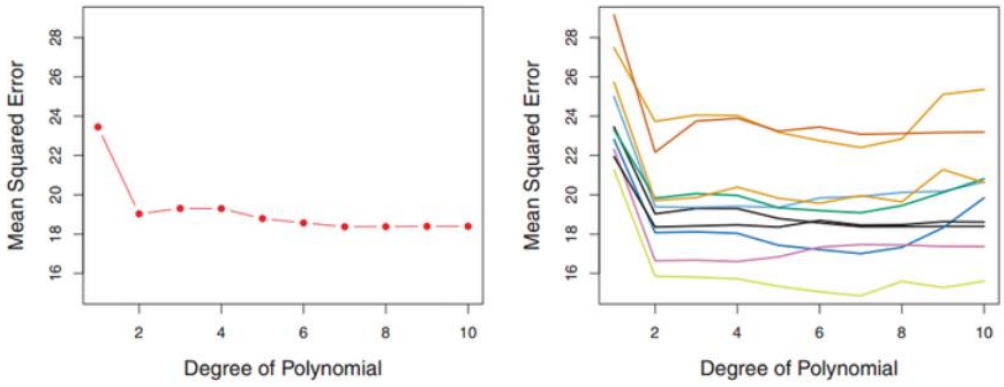
Cross Validation (交叉验证)

在机器学习里，通常不会把所有数据拿去训练，通常会分成训练数据 (Training Data) 和测试数据 (Testing Data)。不使用训练数据来做测试，原因是因为模型已经看过训练数据，再使用训练数据来做测试，会有 Bias 的问题。使用测试数据可以观察模型的 Performance，如果训练的 Loss 很低，但是测试的 Loss 很大，模型有可能已经 Overfit 了。

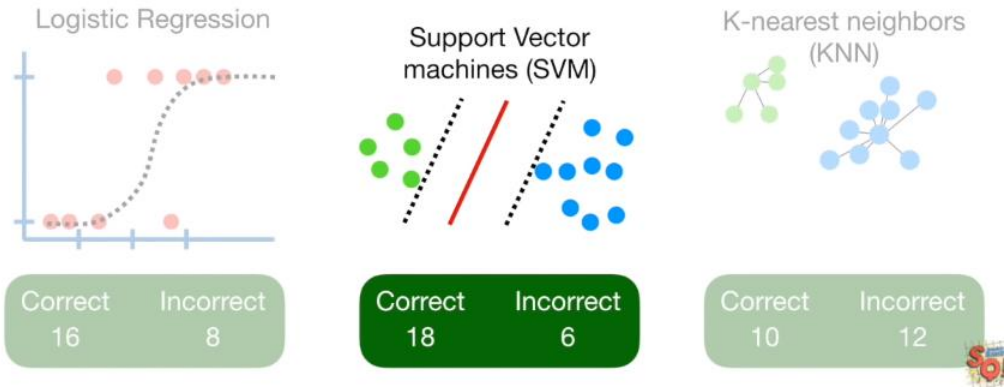
The Validation Set Approach

一般会将**数据随机分成 2 份**，训练数据 (Training Data)，测试数据 (Testing Data)，训练 后进行 MSE 的计算，来算取 Predicted Value 和 Actual Value 的差别。



随机分配数据的方法通常会造成不同的效果，这是因为随机分配的数据可能不够 Balance。上图右侧的 Graph 是 10 种不同的训练集 (Training Set) 和测试集 (Testing Set) Split 得到的 Test MSE 结果，每次的结果都不同。

使用 Cross Validation 可以帮助解决这个问题，也可以帮助选择出更好的模型。

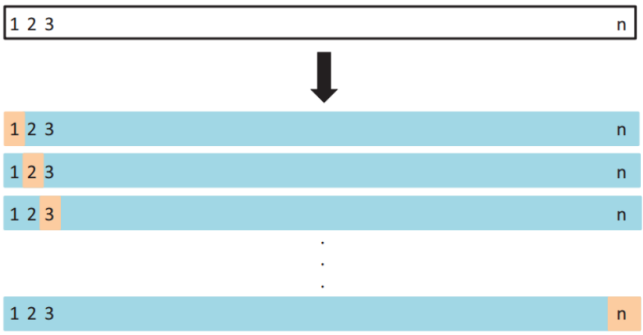


Cross Validation 的目的是用来评估模型在特定数据上训练后的泛化性能 (Generalization Performance) 好坏。

Cross Validation Methods

LOOCV (Leave One Out Cross Validation)

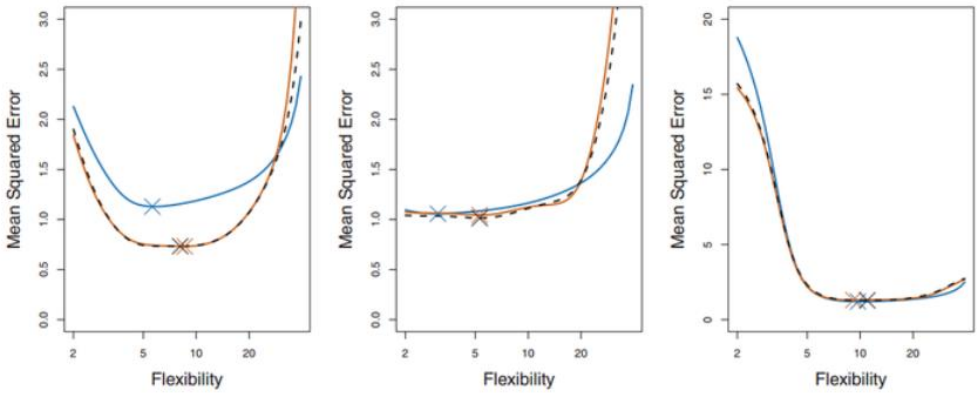
LOOCV 的想法是每一次只使用一个数据当作测试数据，其他数据都当成训练数据。训练完后，将换下一个数据当作测试数据，其他一样当作训练数据，继续训练，直到所有数据都当过 1 次测试数据。



当训练结束后，会得到 n 比模型，然后计算出每个模型对测试数据的 Loss 然后取平均。LOOCV 的优点是不受 Training Set 和 Testing Set 的分配而受到影响，因为每一个数据都单独做过测试。除此之外，以这种方式训练模型，模型比较不容易出现 High Bias 的情况，因为每个模型都使用了 n – 1 个数据来训练。但是 LOOCV 的缺点就是训练量过大，需要训练 n 次模型。

K-Fold Cross Validation (可以看 Bias_Variance 的笔记)

K-Fold 的想法是将数据分成 K 份，然后将分出来的 $\frac{1}{K}$ 比数据当作测试数据，其他都当作训练数据。当训练完后，就换下一份数据当测试数据，其他的数据都当作训练数据继续训练，直到训练完 K 次。最后将这 K 个模型的测试 Loss 取平均。

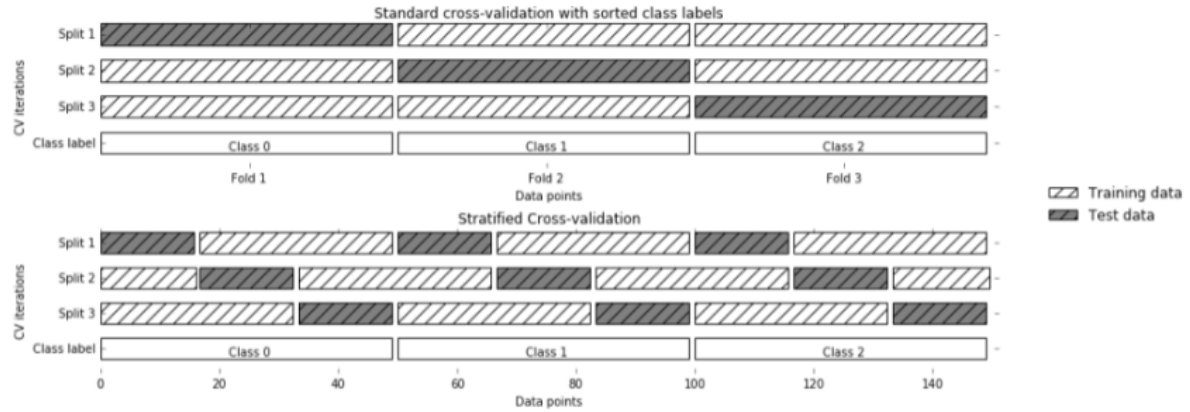


上图是真实的 Test MSE，10-Fold Cross Validation (橙色线) 和 LOOCV (黑色虚线) 的对比。可以看的出来 10-Fold Cross Validation 和 LOOCV 的 Average Loss 是相似的，但是 K-Fold Cross Validation 的方法，它的计算量要比 LOOCV 小很多。

在使用 K-Fold Cross Validation 的时候，需要考虑到 Bias & Variance Tradeoff 的问题。当 K 值越大，也代表训练数据会比较多，这样训练出来的模型的 Bias 会比较小，但是它的问题是训练数据的重复性太高，有可能回导致 Variance 变大 (模型 Overfit)。一般的 K 都设置在 5 或者 10。

Stratified K-Fold Cross Validation

Stratified K-Fold Cross Validation 的想法是将数据分成 K 份，确保每一份数据里面的不同 Class 的比例是一样的。For Example，这笔数据里有 120 个 Class 1，和 120 个 Class 2。假设 K = 3，数据将分配成 3 份，每一份里面都有 40 个 Class 1 和 40 个 Class 2。这个方法可以确保在模型验证的阶段有更好的可靠性。



上图是一个 Stratified K-Fold Cross Validation 和 K-Fold Cross Validation 的对比。