交叉验证(Cross-validation)概述及常见交叉验证方法

作者：writer 来源：爱科学

本文原地址：https://www.iikx.com/news/statistics/2592.html

本文主要介绍交叉验证(Cross-validation)的概念、基本思想、目的、常见的交叉验证形式、Holdout 验证、K-fold cross-validation 和留一验证。时亦称循环估计，是一种统计学上将数据样本切割成较小子集的实用方法。主要用于建模应用中，在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。交叉验证的理论是由Seymour Geisser所开始的。 它对于防范testing hypotheses suggested by the data是非常重要的，特别是当后续的样本是危险、成本过高或不可能(uncomfortable science)去搜集。

交叉验证(Cross-validation)：有时亦称循环估计，是一种统计学上将数据样本切割成较小子集的实用方法。主要用于建模应用中，例如PCR 、PLS 回归建模中。**在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和**。

交叉验证概念

这个过程一直进行，直到所有的样本都被预报了一次而且仅被预报一次。把每个样本的预报误差平方加和，称为PRESS(predicted Error Sum of Squares)。

交叉验证基本思想

交叉验证的基本思想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.

交叉验证目的

用交叉验证的目的是为了得到可靠稳定的模型。在建立PCR 或PLS 模型时，一个很重要的因素是取多少个主成分的问题。用cross validation 校验每个主成分下的PRESS值，选择PRESS值小的主成分数。或PRESS值不再变小时的主成分数。

常用的精度测试方法主要是交叉验证，例如10折交叉验证(10-foldcross validation)，将数据集分成十份，轮流将其中9份做训练1份做验证，10次的结果的均值作为对算法精度的估计，一般还需要

进行多次10折交叉验证求均值，例如：10次10折交叉验证，以求更精确一点。

交叉验证有时也称为交叉比对，如：10折交叉比对

交叉验证常见的交叉验证形式

交叉验证 Holdout 验证

常识来说，Holdout 验证并非一种交叉验证，因为数据并没有交叉使用。
随机从最初的样本中选出部分，形成交叉验证数据，而剩余的就当做训练数据。
一般来说，少于原本样本三分之一的数据被选做验证数据。

方法：将原始数据随机分为两组,一组做为训练集,一组做为验证集,利用训练集训练分类器,然后利用验证集验证模型,记录最后的分类准确率为此Hold-Out Method下分类器的性能指标.。Hold-Out Method相对于K-fold Cross Validation 又称Double cross-validation ，或相对K-CV称 2-fold cross-validation(2-CV) 优点：好处的处理简单,只需随机把原始数据分为两组即可
缺点：严格意义来说Hold-Out Method并不能算是CV,因为这种方法没有达到交叉的思想,由于是随机的将原始数据分组,所以最后验证集分类准确率的高低与原始数据的分组有很大的关系,所以这种方法得到的结果其实并不具有说服性.(主要原因是训练集样本数太少，通常不足以代表母体样本的分布，导致 test 阶段辨识率容易出现明显落差。此外，2-CV
中一分为二的分子集方法的变异度大，往往无法达到「实验过程必须可以被复制」的要求。)
交叉验证 K-fold cross-validation

K折交叉验证
，初始采样分割成K个子样本，一个单独的子样本被保留作为验证模型的数据，其他K-1个样本用来训练。交叉验证重复K次，每个子样本验证一次，平均K次的结果或者使用其它结合方式，最终得到一个单一估测。这个方法的优势在于，同时重复运用随机产生的子样本进行训练和验证，每次的结果验证一次，10折交叉验证是最常用的。

优点
：每一个样本数据都即被用作训练数据，也被用作测试数据。避免的过度学习和欠学习状态的发生，得到的结果比较具有说服力。 **缺点** ：K值选取上

交叉验证**留一验证 Leave-One-Out Cross Validation**

正如名称所建议， 留一验证(LOOCV)意指只使用原本样本中的一项来当做验证资料，
而剩余的则留下来当做训练资料。 这个步骤一直持续到每个样本都被当做一次验证资料。
事实上，这等同于和K-fold 交叉验证，其中K为原本样本个数。
在某些情况下是存在有效率的演算法，如使用kernel regression 和Tikhonov regularization。
优点
：每一个分类器或模型都是用几乎所有的样本来训练模型，最接近样本，这样评估所得的结果比较可靠。实验没有随机因素，整个过程是可重复的。 **缺点**
：计算成本高，当N非常大时，计算耗时,因为需要建立的模型数量与原始数据样本数量相同,当原始数据样本数量相当多时,LOO-CV在实作上便有困难几乎就是不显示,除非每次训练分类器得到模型的速度很快,或是可以用并行化计算减少计算所需的时间.

十折交叉验证：10-fold cross validation

英文名叫做10-foldcross-validation，用来测试算法准确性。是常用的测试方法。将数据集分成十分，轮流将其中9份作为训练数据，1份作为测试数据，进行试验。每次试验都会得出相应的正确率(或差错率)。10次的结果的正确率(或差错率)的平均值作为对算法精度的估计，一般还需要进行多次10折交叉验证(例如10次10折交叉验证)，再求其均值，作为对算法准确性的估计。

之所以选择将数据集分为10份，是因为通过利用大量数据集、使用不同学习技术进行的大量试验，表明10折是获得最好误差估计的恰当选择，而且也有一些理论根据可以证明这一点。但这并非最终诊断，争议仍然存在。而且似乎5折或者20折与10折所得出的结果也相差无几。

10折交叉验证是把样本数据分成10份，轮流将其中9份做训练数据，将剩下的1份当测试数据，10次结果的均值作为对算法精度的估计，通常情况下为了提高精度，还需要做多次10折交叉验证。更进一步，还有K折交叉验证，10折交叉验证是它的特殊情况。K折交叉验证就是把样本分为K份，其中K-1份用来做训练建立模型，留剩下的一份来验证，交叉验证重复K次，每个子样本验证一次。

交叉验证插值

交叉验证(cross-validation) 方法是一种评价插值方法质量的方法，通过交叉验证报告文件中的统计量可以确定设置的插值方法相关参数是否合理，从而可以比较出不同插值模型得出结果的不同质量。

交叉验证方法即移去一个已知采样点的数据，用其他采样点的数据来估计该点以检验插值精度的方法。

交叉验证可以使用一些统计指标来进行评价，令 z0为采样值，ze为对应点的估计值，则有

1）误差序列（Error）：

$$E_i = Z_i^e - Z_i^o$$

2)平均误差（Mean Error）：

$$ME = \frac{1}{n}\sum_{i=1}^{n} E_i$$

3)平均绝对误差(Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |E_i|$$

4）均方根误差（Root Mean Square Error）

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n} E_i^2\right]^{\frac{1}{2}}$$

在模式识别与机器学习的相关研究中，经常会将 数据集分为 训练集与测试集
这两个子集，前者用以建立

模式，后者则用来评估该模式对未知样本进行预测时的精确度，正规的说法是 generalization ability(泛化能力)

交叉验证核心原则Cross-validation 是为了有效的估测 generalization error 所设计的实验方法

只有训练集才可以用在 模式的训练过程中，测试集
则必须在模式完成之后才被用来评估模式优劣的依据。

常见的错误运用：许多人在研究都有用到 Evolutionary Algorithms(EA,遗传算法)与 classifiers，所使用的 Fitness Function (适应度函数)中通常都有用到 classifier 的辨识率，然而把Cross-Validation 用错的案例还不少。前面说过，只有 training data 才可以用于 model 的建构，所以只有 training data 的辨识率才可以用在 fitness function 中。而 EA 是训练过程用来调整 model 最佳参数的方法，所以只有在 EA结束演化后，model 参数已经固定了，这时候才可以使用 test data。

EA 与 CV结合研究方法：Cross-Validation 的本质是用来估测某个 classification method 对一组 dataset 的 generalization error，不是用来设计 classifier 的方法，所以 Cross-Validation 不能用在 EA的 fitness function 中，因为与 fitness function 有关的样本都属于 training set，那试问哪些样本才是 test set 呢?如果某个 fitness function 中用了Cross-Validation 的 training 或 test 辨识率，那么这样的实验方法已经不能称为 Cross-Validation.

EA 与 k-CV 正确的搭配方法：是将 dataset 分成 k 等份的 subsets 后，每次取 1份 subset 作为 test set，其余 k-1 份作为 training set，并且将该组 training set 套用到 EA 的 fitness function 计算中(至于该 training set 如何进一步利用则没有限制)。因此，正确的 k-CV 会进行共 k 次的 EA 演化，建立 k 个classifiers。而 k-CV 的 test 辨识率，则是 k 组 test sets 对应到 EA 训练所得的 k 个 classifiers 辨识率之平均值.

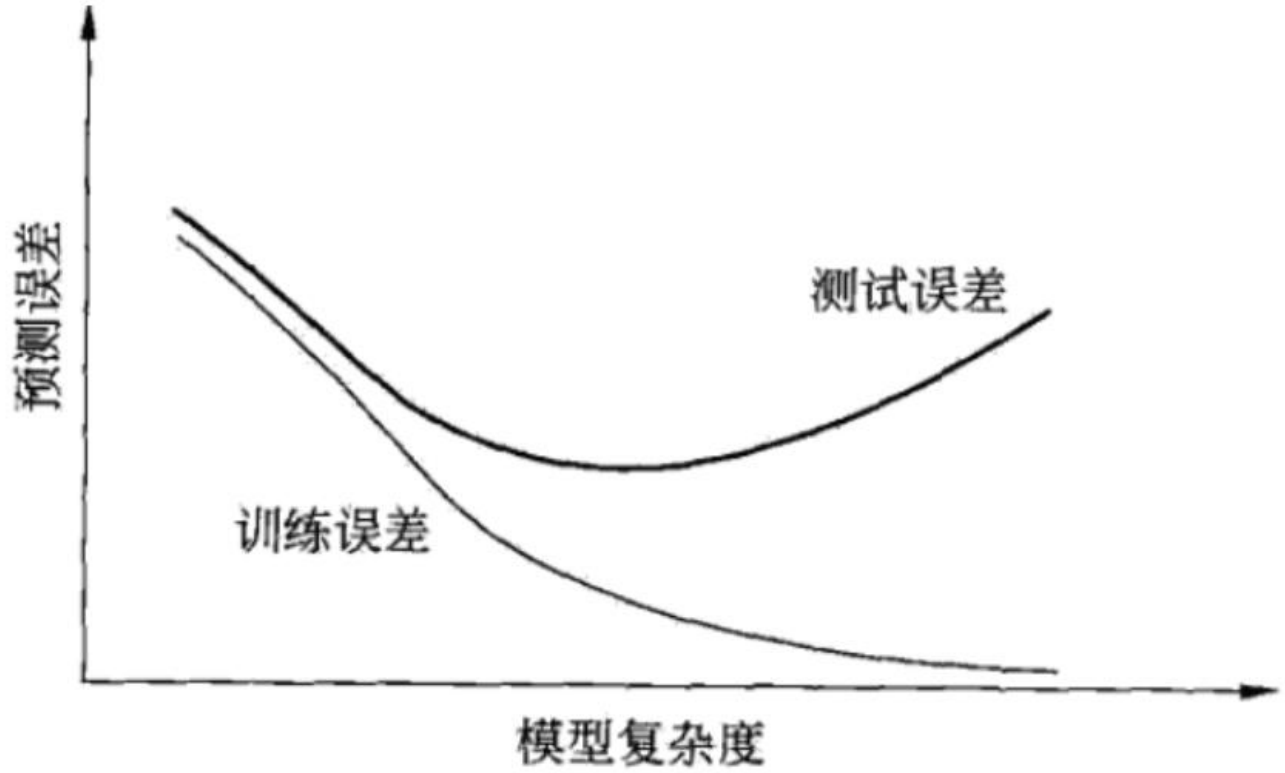
数据集分割原则交叉验证在，原始数据集分割为训练集与测试集，必须遵守两个要点：

训练集中样本数量必须够多，一般至少大于总样本数的 50%。
两组子集必须从完整集合中均匀取样。

其中第2点特别重要，均匀取样的目的是希望减少 训练集/测试集与完整集合之间的偏差(bias)，但却也不易做到。一般的作法是随机取样，当样本数量足够时，便可达到均匀取样的效果。然而随机也正是此作法的盲点，也是经常是可以在数据上做手脚的地方。举例来说，当辨识率不理想时，便重新取样一组训练集与测试集，直到测试集的辨识率满意为止，但严格来说便算是作弊。

拓展：

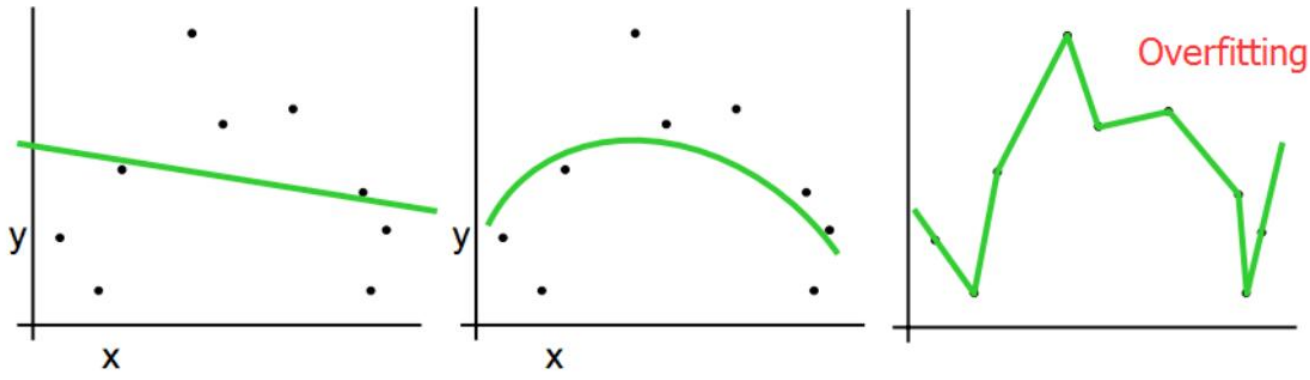
当假设空间含有不同复杂度(如不同参数数量)的模型时，就要进行模型选择。如果过度追求在训练数据集上误差小的模型，那么选出来的模型在测试数据集上的误差就可能很大，此时模型过拟合了训练数据集，图1显示了训练误差和测试误差与模型复杂度之间的关系。



所以模型选择时应特别注意防止过拟合，本文首先回顾了过拟合，之后介绍防止过拟合常用的方法之一——交叉验证。

过拟合

若训练得到的模型的复杂度超过真实模型的复杂度，就称发生了过拟合，反之为欠拟合。过拟合发生的原因是训练数据集中存在随机噪声和确定性噪声。



(注：图片来自 Tutorial Slides by Andrew Moore)

交叉验证

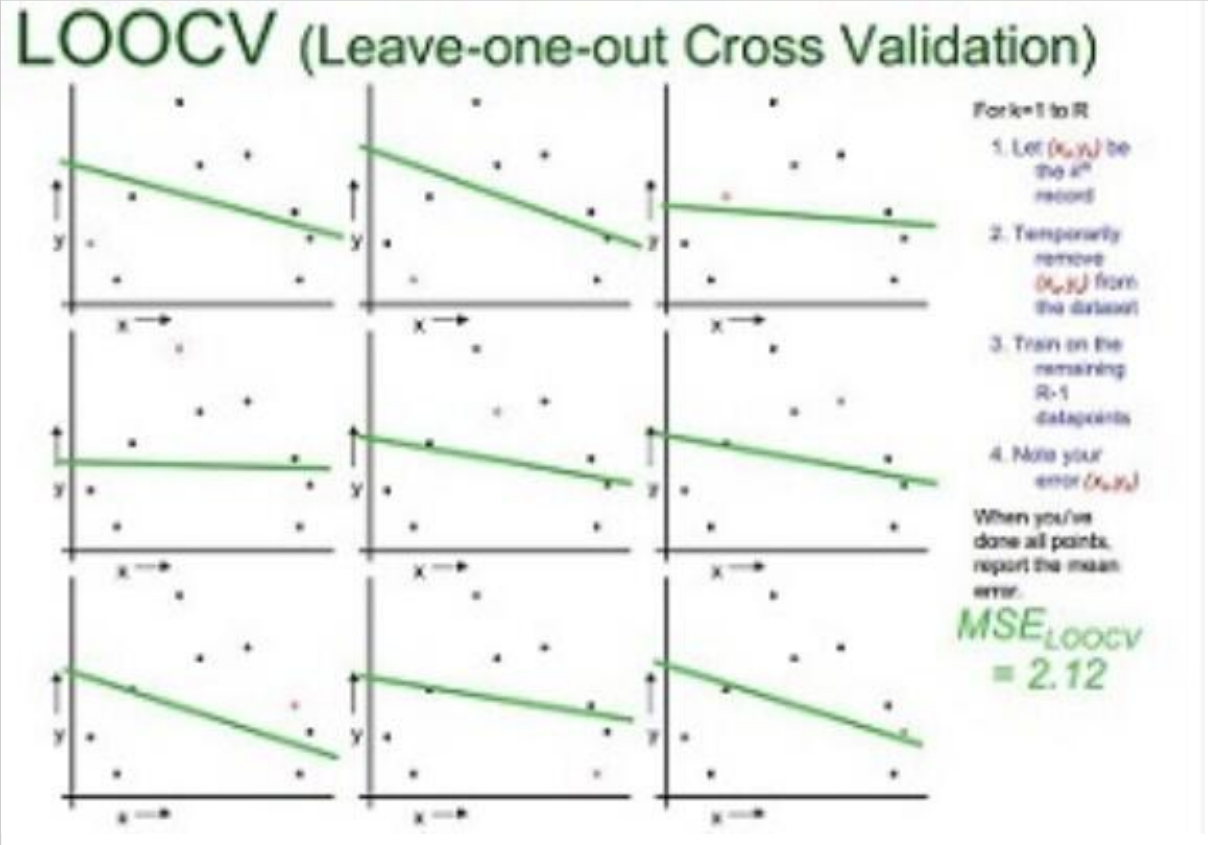
交叉验证(Cross-validation，CV)目的：检测和预防过拟合

交叉验证方法	优点	缺点
Test-set	计算开销小	无法评估模型泛化能力
Leave-one-out cross validation（LOOCV）	不浪费数据	计算开销大
k-fold cross validation	计算开销相对LOOCV小	浪费1/k的数据

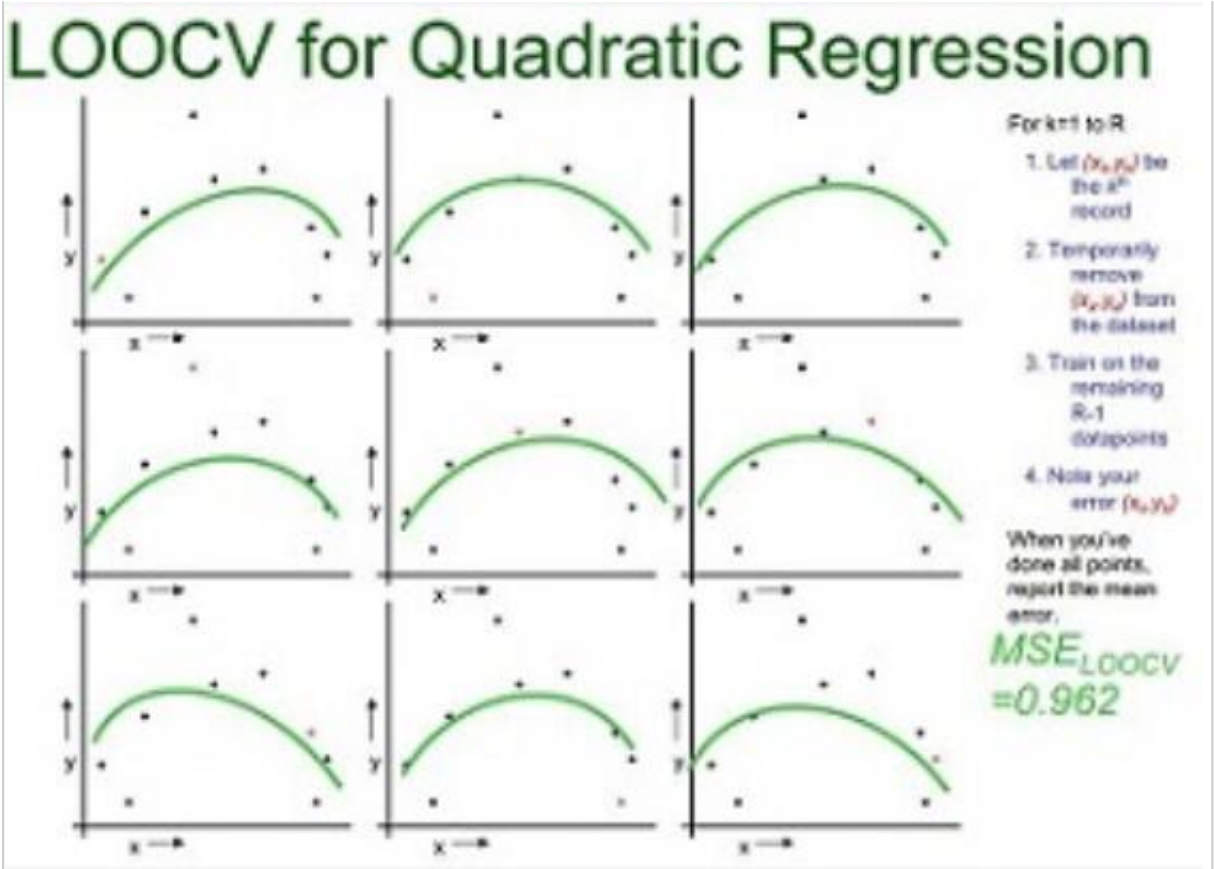
Test-set将数据集中的全部数据用于模型训练，不考虑模型验证，选择训练集上误差最小的模型为最优模型，易产生过拟合。

LOOCV (Leave-one-out Cross Validation)

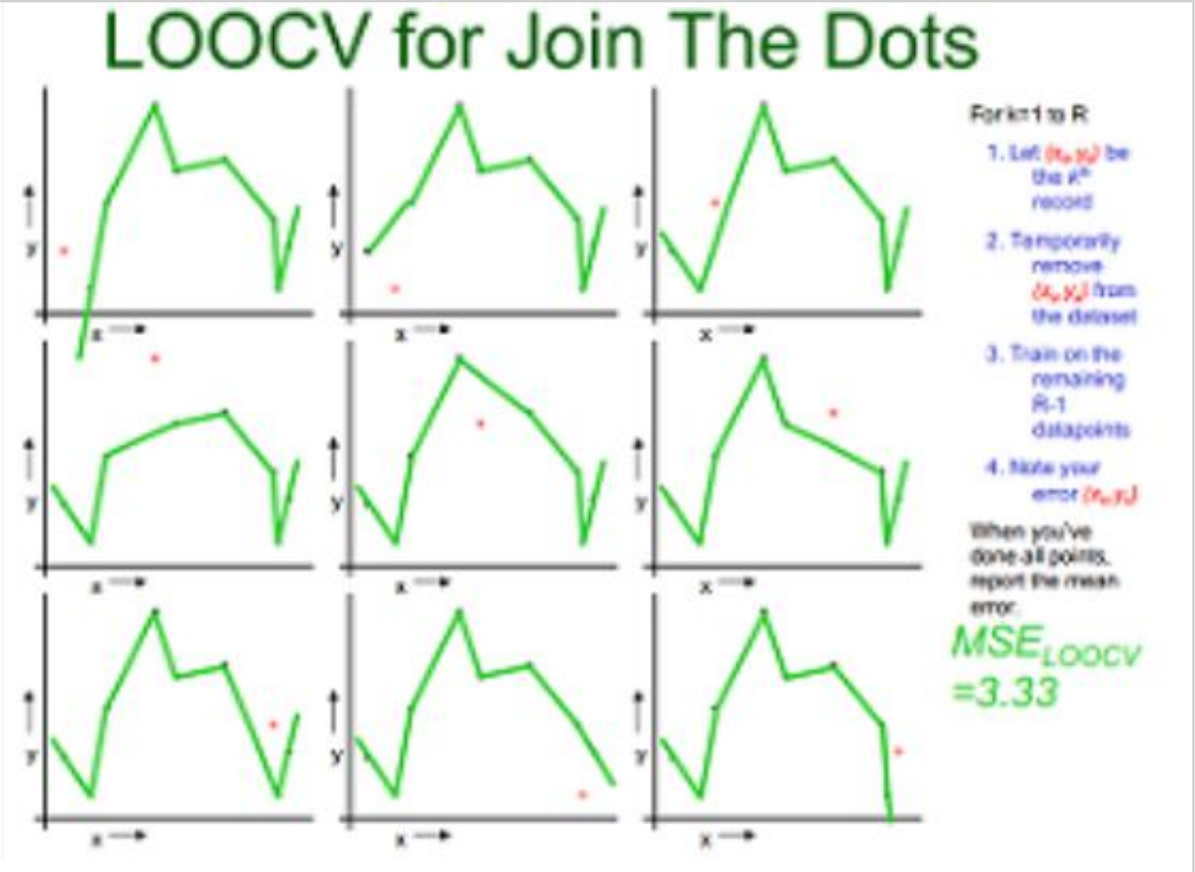
下图示例了使用LOOCV方法对线性回归、二次回归、直接点连接模型进行选择的过程.从大小为n的数据集中抽出一个作为模型验证样本，其他的(n-1)个样本用于模型训练，这样对于线性拟合、二次拟合、点连接三种模型分别有n个模型和对应得3个的均方误差(MSE)，选择均方差最小的，即二次拟合为最优模型。



线性拟合


LOOCV for Quadratic Regression
For k=1 to R
1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)
When you've done all points, report the mean error.
MSE_{LOOCV} =0.962

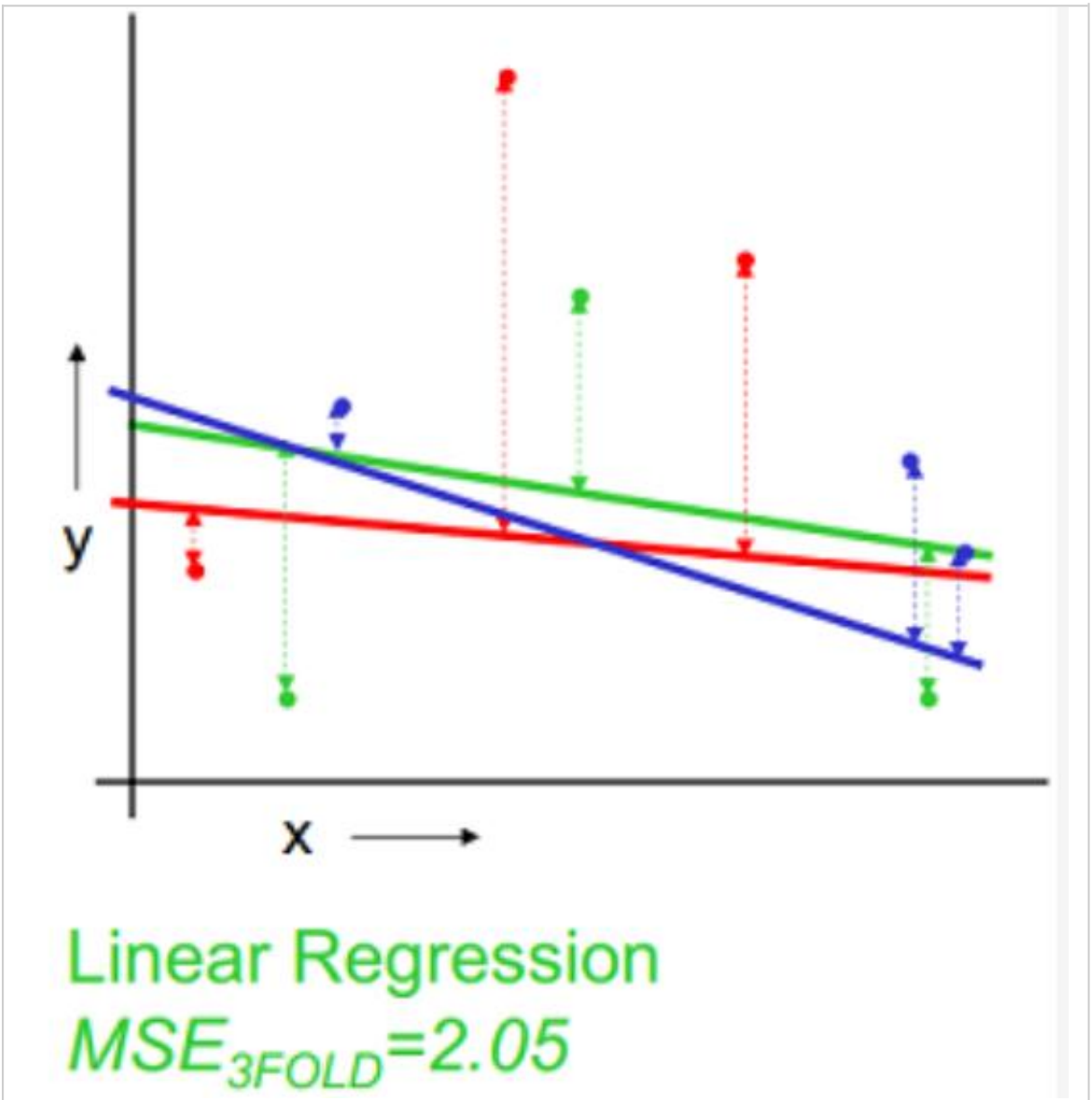
二次拟合


LOOCV for Join The Dots
For k=1 to R
1. Let (x_k, y_k) be the k^{th} record
2. Temporarily remove (x_k, y_k) from the dataset
3. Train on the remaining R-1 datapoints
4. Note your error (x_k, y_k)
When you've done all points, report the mean error.
MSE_{LOOCV} =3.33

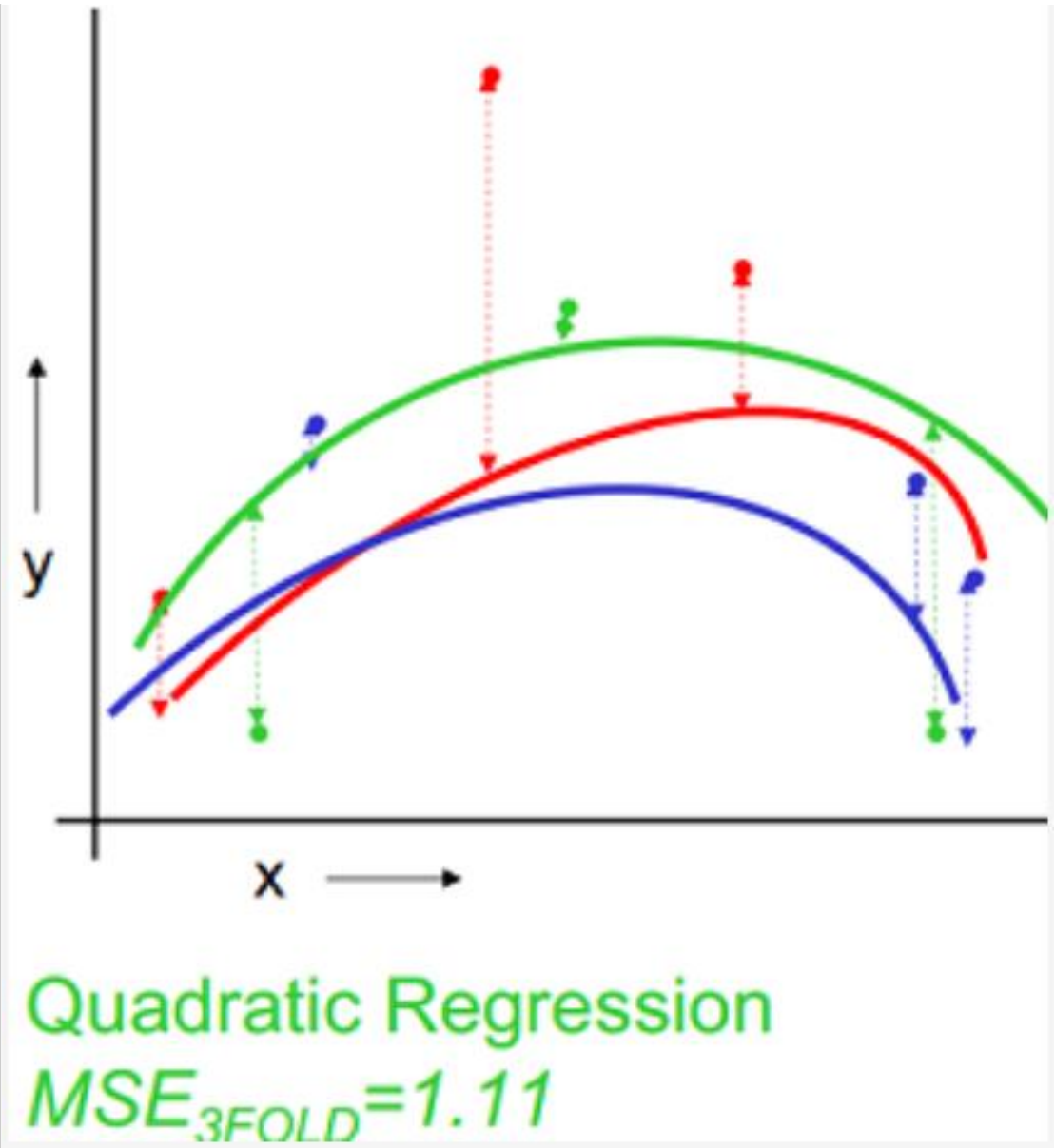
点连接

k-fold cross validation

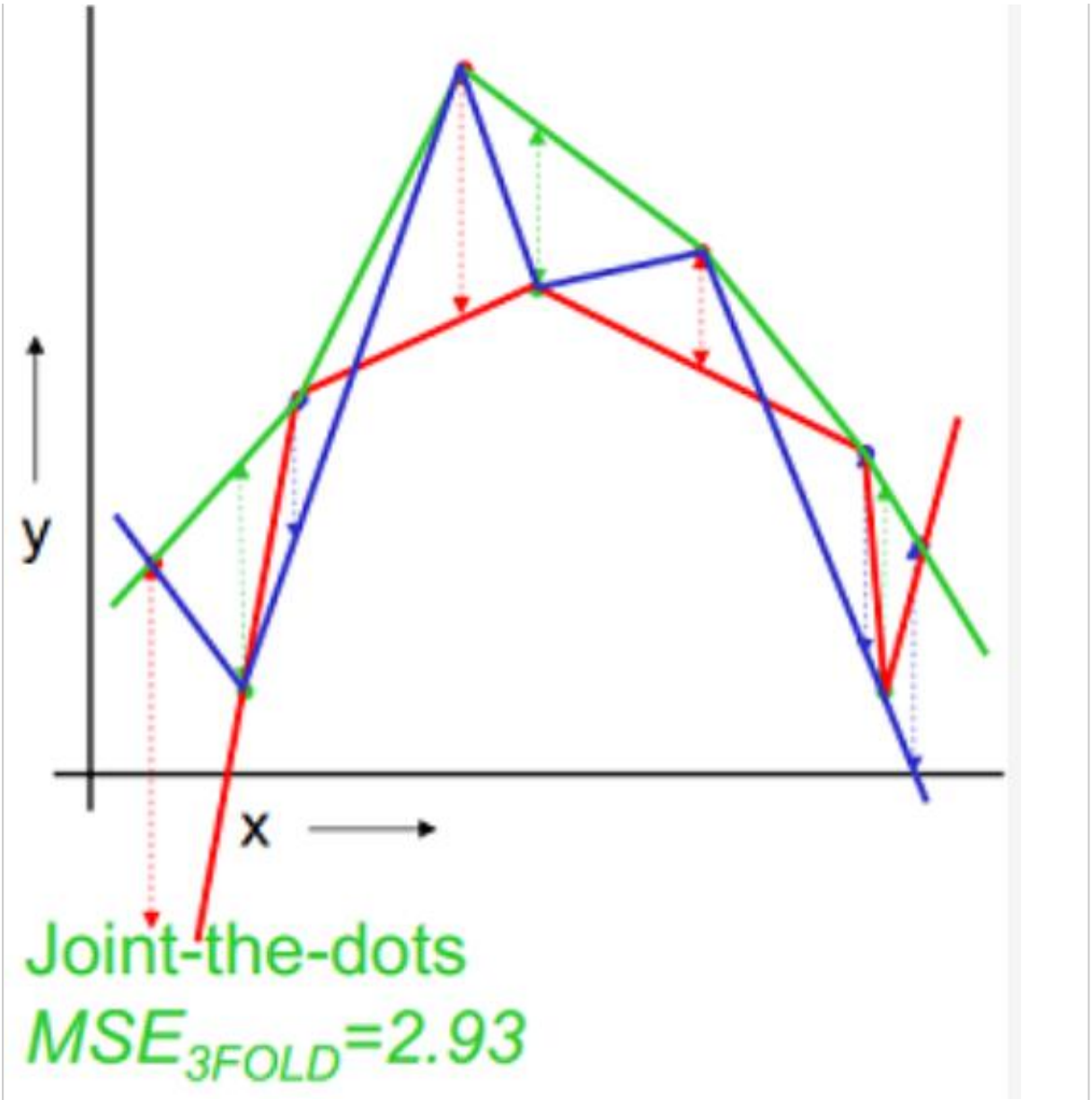
以k=3为例，下图示例了使用3-fold交叉验证的方法对线性回归、二次回归、直接点连接模型进行选择的过程，数据集被随机划分为3份，其中2份用来训练模型，1份用来验证，这样针对线性、二次拟合、点连接模型分别有3个训练好的模型和均方误差(MSE)，选择均方差最小的，即二次拟合为最优模型。



线性拟合



二次拟合



点连接

更多 统计方法 请访问 https://www.iikx.com/news/statistics/