Error(误差)、Bias(偏差)、Variance(方差)

概念：训练误差（training error）、经验误差（empirical error）、泛化误差（generalization error）、拟合能力、稳定性、波动情况、数据充分性、偏差-方差窘境（bias-variance dilemma）、训练程度、欠拟合（underfitting）、过拟合（overfitting）

【参考】

- [简书 - 总结：Bias(偏差)，Error(误差)，Variance(方差)及CV(交叉验证)]
- [Understanding the Bias-Variance Tradeoff] 有公式
- [斯坦福机器学习笔记 - 偏差与方差]
- [模型评估与选择 （Bias(偏差)，Error(误差)，和Variance(方差)）] 较全面

一般地，通常会把模型输出和真实值之间的差异称为**误差（error）**。在训练集上的误差称为**训练误差（training error）**或者**经验误差（empirical error）**。而在新样本上的误差则称为**泛化误差（generalization error）**。

机器学习中的泛化性能可以拆解为：$Err(x) = Bias^2 + Variance + Noise$

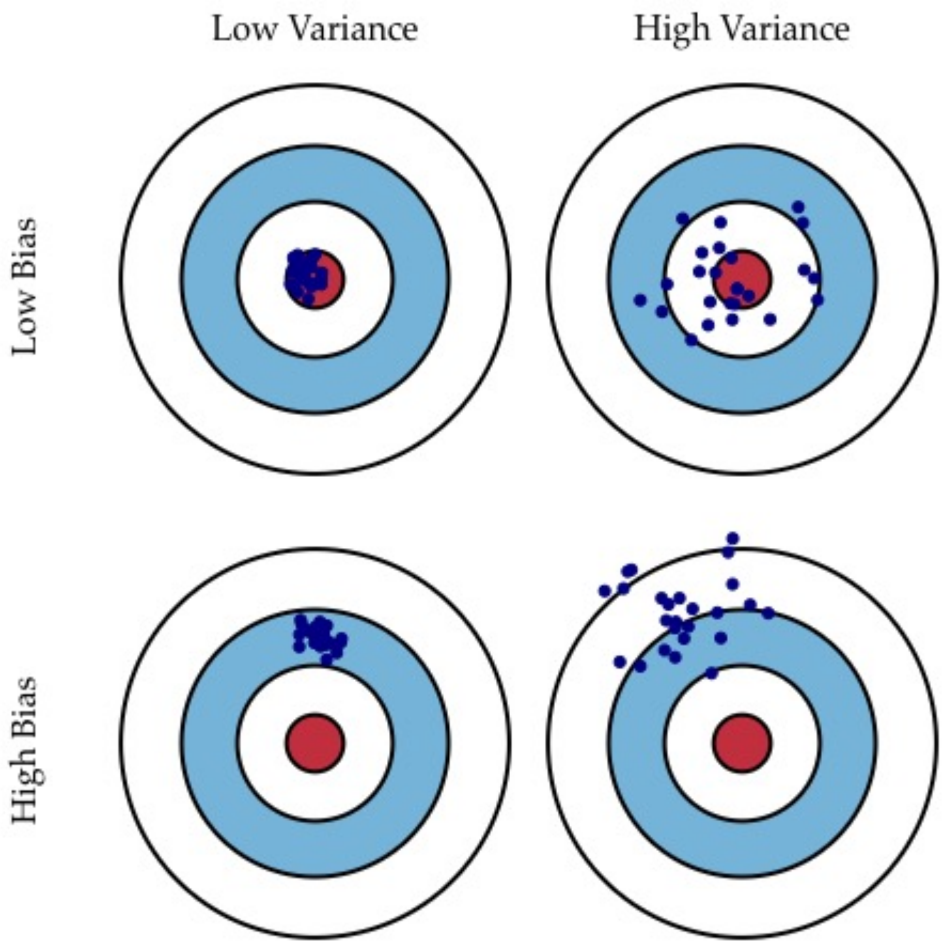
(1) Bias(偏差)
Bias反映的是模型在样本上的**输出与真实值**之间的**误差**，即模型本身的**精准度**，即算法本身的**拟合能力**。偏差越大，越偏离真实数据，如下图的第二行。

(2) Variance(方差)
Variance反映的是模型每一次**输出结果与模型输出期望**之间的误差，即模型的**稳定性**，反应预测的**波动情况**（即使用同规模的不同训练集进行训练时带来的性能变化，刻画数据扰动带来的影响）。他描述了预测值的变化范围，离散程度，也就是离其期望值的距离。方差越大，数据的分布越分散，如下图的第二列。

(3) Noise(噪音)
当前任务上任何算法所能达到的期望泛化误差的下界（即不可能有算法取得更小的误差），刻画**问题本身的难度**。即使数据集本身上的问题，不管你算法在怎么努力都不可能取得更小的误差，无论如何都不可能逾越过去，除非更换数据集。
就是不是你想要的真正数据，你可以想象为来破坏你实验的元凶和造成你可能过拟合的原因之一，至于为什么是过拟合的原因，因为模型过度追求Low Bias会导致训练过度，对测试集判断表现优秀，导致噪声点也被拟合进去了。

也即是说，泛化性能包含了学习算法的拟合能力（偏差）、数据的充分性（方差）以及问题本身的难度（噪音）共同决定的。给定一个任务，噪声是固定的，我们需要做得就是尽量降低偏差和方差。

偏差与方差的关系可以参考如下图：



但是这两者其实是有冲突的，这称为**偏差-方差窘境（bias-variance dilemma）**。给定一个任务，我们可以控制算法的训练程度（如决策树的层数）。

- 在训练程度较低时，拟合能力较差（欠拟合（underfitting）），因此训练数据的扰动不会让性能有显著变化，此时偏差主导泛化错误率；
- 在训练程度较高时，拟合能力很强（过拟合（overfitting）），以至于训练数据自身的一些特性都会被拟合，从而产生过拟合问题，训练数据的轻微扰动都会令模型产生很大的变化，此时方差主导泛化错误率。

需要注意的是，将泛化性能完美地分解为方差、偏差、噪声这三项仅在**基于均方误差的回归任务中**得以推导出，分类任务由于损失函数的跳变性导致难以从理论上推导出分解形式，但已经有很多方法可以通过实验进行估计了。

通过诊断判断出现了高偏差或者高方差，我们就可以才去相应的处理方法：

手段	使用场景	备注
采集更多的样本	高方差	适应数据集的变化
降低特征维度	高方差	去除冗余的特征，防止学习到与数据集无关的特征
采集更多的特征	高偏差	学习更多的特征，增加拟合能力
进行高次多项式回归	高偏差	增加拟合能力
降低参数 λ（正则项）	高方差	降低模型的复杂度，适应数据集的变化
增大参数 λ（正则项）	高偏差	增加模型复杂度，增加拟合能力

数据集：训练集、验证集、测试集

- sohu - 业界 | 似乎没区别，但你混淆过验证集和测试集吗？
- 个站 - What is the Difference Between Test and Validation Datasets? 上文的原文
- 知乎 - 训练集(train)验证集(validation)测试集(test)与交叉验证法(Cross Validation)
- 交叉验证（简单交叉验证、k折交叉验证、留一法）
- 知乎 - 留一法交叉验证和普通交叉验证有什么区别？
- wikipedia - Training, validation, and test sets

通常情况下，「验证数据集」指模型训练过程中留出的样本集，可与「测试数据集」这个术语互换。

- 训练集：用来学习的样本集，用于分类器参数的拟合。
- 验证集：用来调整分类器超参数的样本集，如在神经网络中选择隐藏层神经元的数量。
- 测试集：仅用于对已经训练好的分类器进行性能评估的样本集。

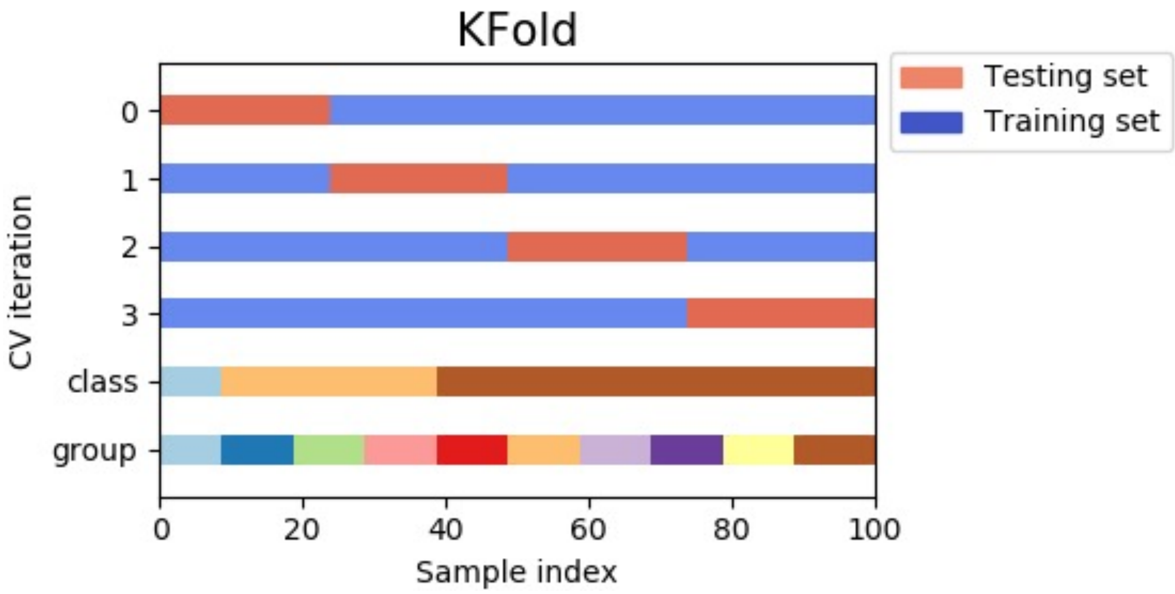
三者的区别
形象上来说训练集就像是学生的课本，学生 根据课本里的内容来掌握知识，验证集就像是作业，通过作业可以知道 不同学生学习情况、进步的速度快慢，而最终的测试集就像是考试，考的题是平常都没有见过，考察学生举一反三的能力。

为什么要测试集
训练集直接参与了模型调惨的过程，显然不能用来反映模型真实的能力，这样一些 对课本死记硬背的学生(过拟合)将会拥有最好的成绩，显然不对。同理，由于验证集参与了人工调参(超参数)的过程，也不能用来最终评判一个模型，就像刷题库的学生也不能算是学习好的学生是吧。所以要通过最终的考试(测试集)来考察一个学(模)生(型)真正的能力。

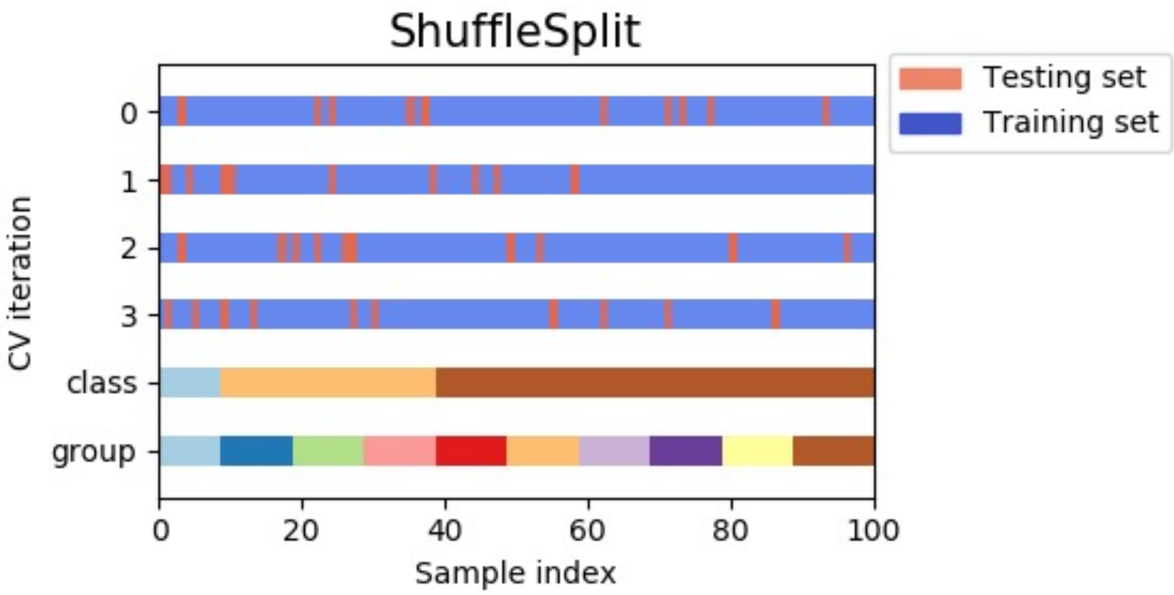
但是仅凭一次考试就对模型的好坏进行评判显然是不合理的，所以就需要使用`交叉验证法`，交叉验证法的作用就是尝试利用不同的训练集/测试集划分来对模型做多组不同的训练/测试，来应对单次测试结果过于片面以及训练数据不足的问题。

验证

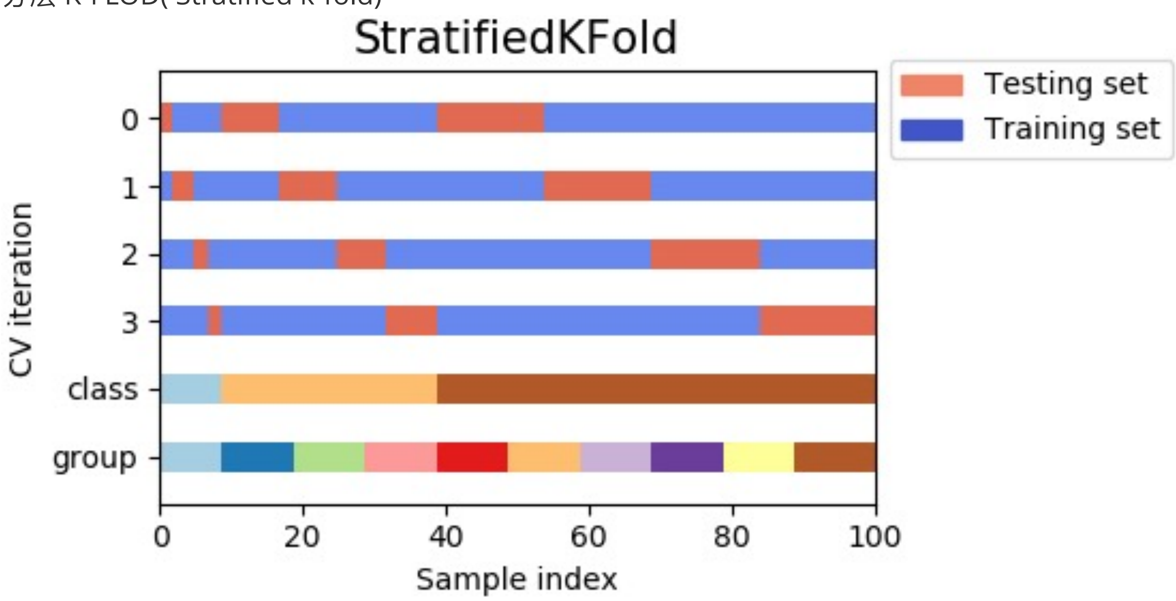
- sklearn - Cross-validation: evaluating estimator performance
- KFold



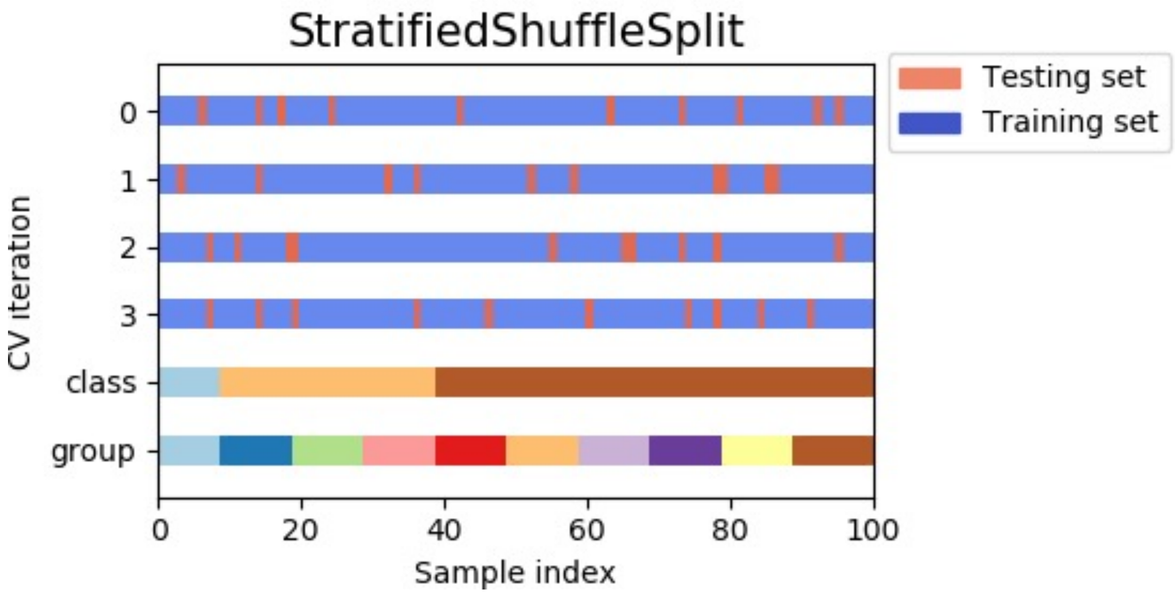
- 留一法(Leave One Out, LOO)
- Leave P Out(LPO)
- 随机排列交叉验证(Random permutations cross-validation a.k.a. Shuffle & Split)



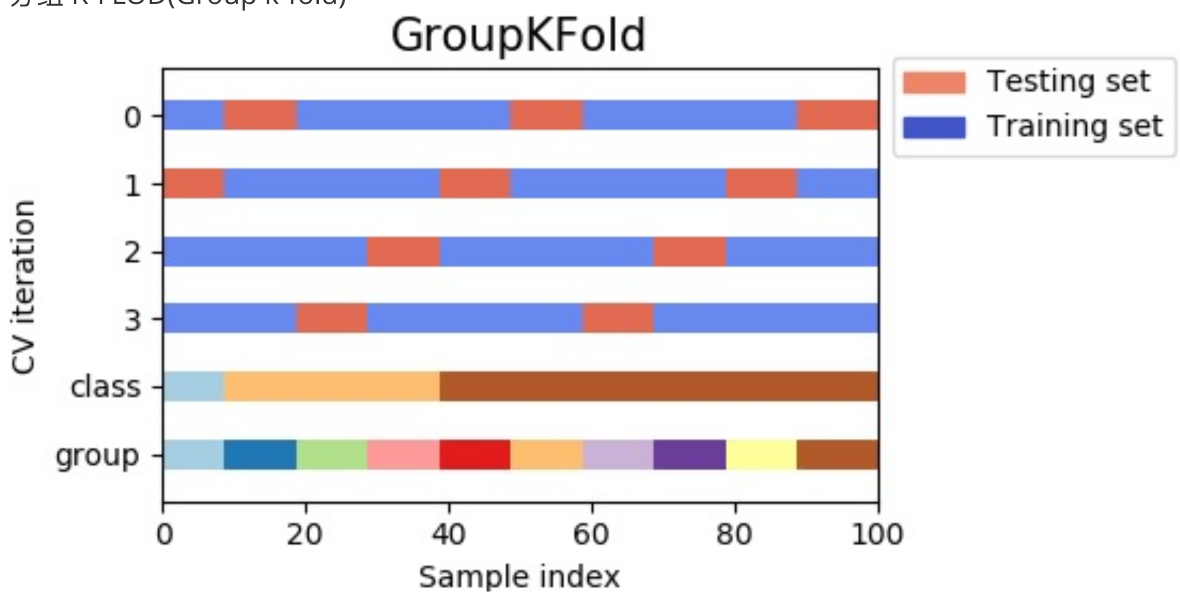
- 分层 K-FLOD(Stratified k-fold)



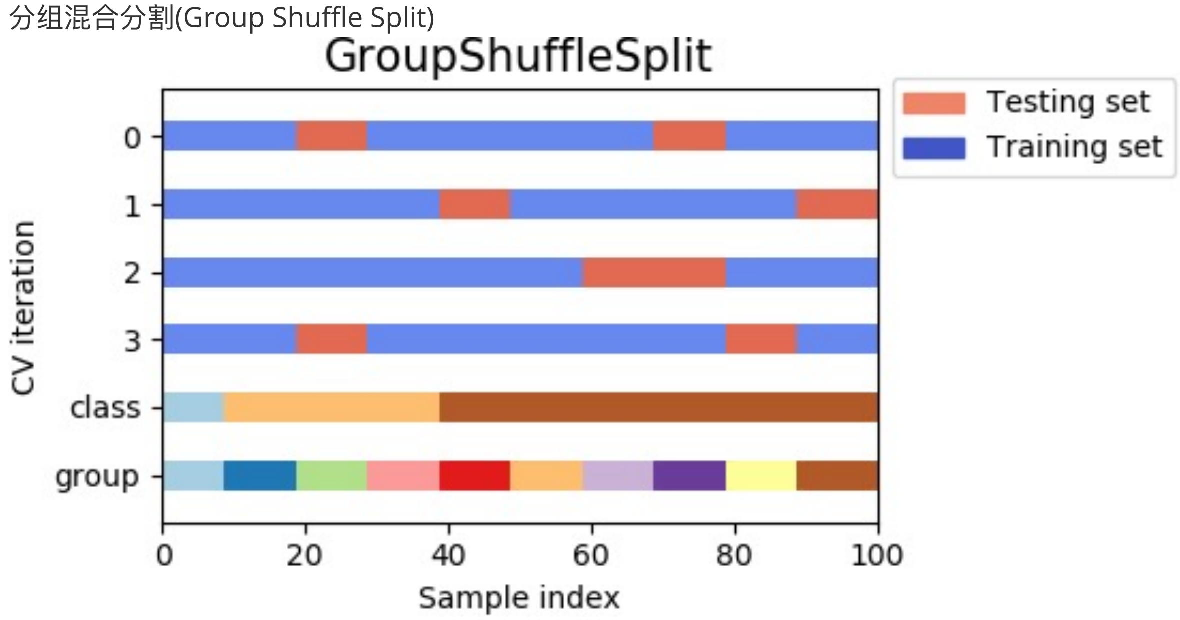
- 分层混合分割(Stratified Shuffle Split)



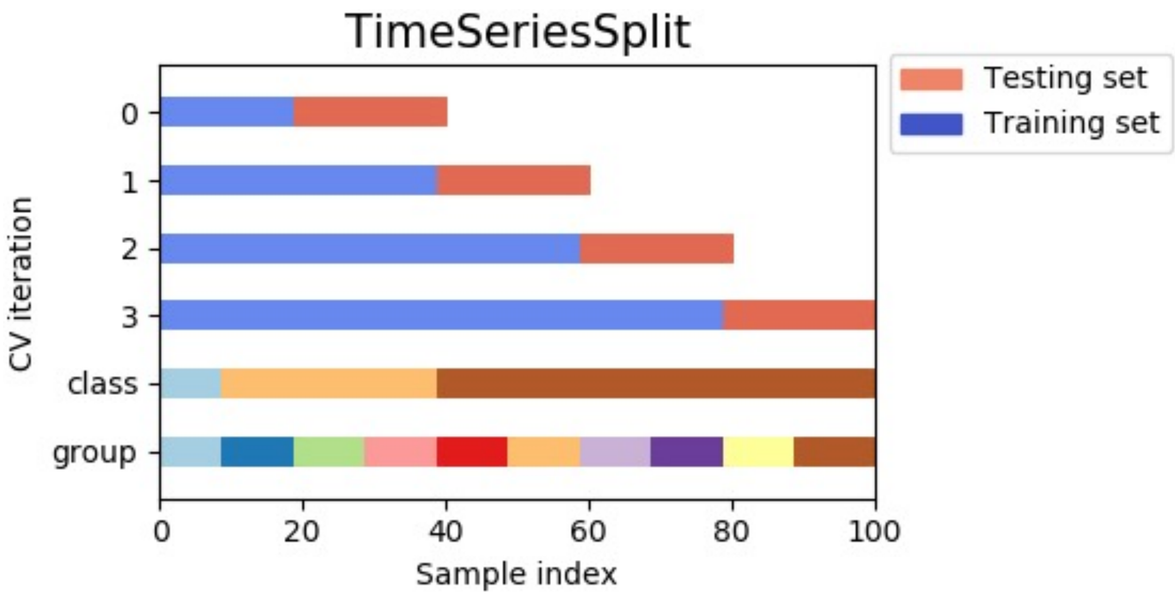
- 分组 K-FLOD(Group k-fold)



- 留一组法(Leave One Group Out, LOGO)
- 留 P 组(Leave P Groups Out, LPGO)
- 分组混合分割(Group Shuffle Split)



- 时序分割(Time series Split)



评估

分类

精确度(accuracy)

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量
- $\mathbf{1}(x)$ 指示函数，条件为真是值为1，否则为 0.

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}(\hat{y}_i = y_i)$$

平衡精确度(balanced accuracy score)

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- $\mathbf{1}(x)$ 指示函数，条件为真是值为1，否则为 0.

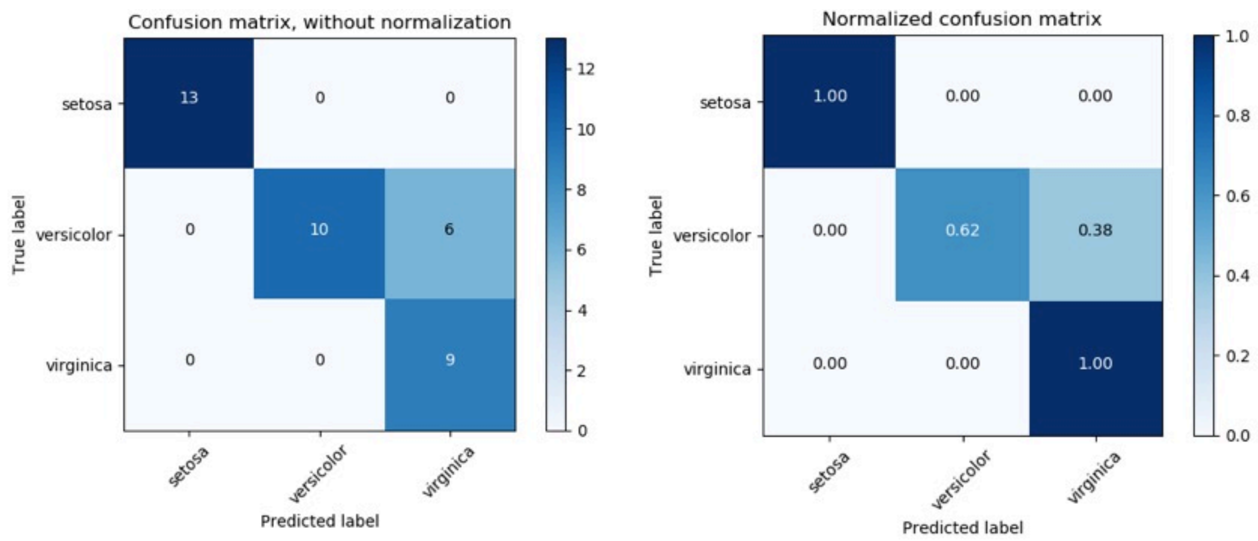
$$\texttt{balanced-accuracy}(y, \hat{y}, \omega) = \frac{1}{\sum \hat{\omega}_i} \sum_i \mathbf{1}(\hat{y}_i = y_i) \hat{\omega}_i$$

其中 $\hat{\omega}_i$ 为：

$$\hat{\omega}_i = \frac{\omega_i}{\sum_j \mathbf{1}(y_j = y_i) \omega_j}$$

混淆矩阵(confusion matrix)

对角线上的值越大越好，表明被预测正确的值越多。



查准率、查全率、F-score

- tp: true positive, Correct result
- fp: false positive, Unexpected result
- fn: false negative, Missing result
- tn: true negative, Correct absence of result

查准率(precision)：预测为正例的样本中有多少真正例

$$\text{precision} = \frac{tp}{tp + fp}$$

查全率(recall)：正例样本中预测为正例的样本比例

$$\text{recall} = \frac{tp}{tp + fn}$$

F score：precision与 recall 的折中

$$F_\beta = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

多类别和多标签分类

- y： 预测 `(sample, label)` 对集合
- $\hat{y}$： 真实 `(sample, label)` 对集合
- L：标签集合
- S：样本集合
- y_s： 样本为 s 的 y 的子集，如 $y_s := \{(s', l) \in y \mid s' = s\}$
- y_l： 标签为 l 的 y 的子集
- $\hat{y}_s, \hat{y}_l$与上面类似，是$\hat{y}$ 的子集
- $P(A, B) := \frac{|A \cap B|}{|A|}$
- $R(A, B) := \frac{|A \cap B|}{|B|}$

- $F_\beta(A, B) = (1 + \beta^2) \frac{P(A,B) \times R(A,B)}{\beta^2 \cdot P(A,B) + R(A,B)}$

定义如下的度量标准：

average	Precision	Recall	F_beta
"micro"	$P(y, \hat{y})$	$R(y, \hat{y})$	$F_\beta(y, \hat{y})$
"samples"	$\frac{1}{\|S\|} \sum_{s \in S} P(y_s, \hat{y}_s)$	$\frac{1}{\|S\|} \sum_{s \in S} R(y_s, \hat{y}_s)$	$\frac{1}{\|S\|} \sum_{s \in S} F_\beta(y_s, \hat{y}_s)$
"macro"	$\frac{1}{\|L\|} \sum_{l \in L} P(y_l, \hat{y}_l)$	$\frac{1}{\|L\|} \sum_{l \in L} R(y_l, \hat{y}_l)$	$\frac{1}{\|L\|} \sum_{l \in L} F_\beta(y_l, \hat{y}_l)$
"weighted"	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| P(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| R(y_l, \hat{y}_l)$	$\frac{1}{\sum_{l \in L} \|\hat{y}_l\|} \sum_{l \in L} \|\hat{y}_l\| F_\beta(y_l, \hat{y}_l)$
None	$\langle P(y_l, \hat{y}_l) \| l \in L \rangle$	$\langle R(y_l, \hat{y}_l) \| l \in L \rangle$	$\langle F_\beta(y_l, \hat{y}_l) \| l \in L \rangle$

回归

平均绝对值误差（L1）

Mean absolute error（MAE）

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量

$$\text{MAE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} |y_i - \hat{y}_i|$$

均方误差

Mean squared error（MSE）

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2$$

均方对数误差

Mean squared logarithmic error（MSLE）

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量

$$\text{MSLE}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} (\log_e(1 + y_i) - \log_e(1 + \hat{y}_i))^2$$

中位数绝对值误差

Median absolute error（MedAE），不支持多输出

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量

$$\text{MedAE}(y, \hat{y}) = \text{median}(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid)$$

R2分数

- y_i 第 i 个样本的真是值
- $\hat{y}_i$ 第 i 个样本的预测值
- n 样本的数量

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n-1}(y_i - \bar{y})^2}, \ \ \bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i$$

其他

- 数据 Independent and Identically Distributed (i.i.d.) 假设