模型度量指标方法

训练集测试集划分

模型度量指标

- 分类模型度量
  - Accuracy

    预测正确的样本占总体样本的比例，取值范围为[0,1]。精度评价指标对平等对待每个类别，即每一个样本判对 (0) 和判错 (1) 的代价都是一样的。

    缺点是对于有倾向性和数据不均衡的数据不可以只用Accuracy判断，如判断飞行物是否为导弹，宁可错判也不能不管。当数据特别不均衡时候，无脑判为数量较大的那一类别也可以得到很高的accuracy但是模型效果不是很好。
  - confusion matrix

	预测值（正）	预测值（负）
真实值（正）	TP	FN
真实值（负）	FP	TN

    - True positive (TP)

      真实值为Positive，预测正确（预测值为Positive）
    - True negative (TN)

      真实值为Negative，预测正确（预测值为Negative）
    - False positive (FP)

      真实值为Negative，预测错误（预测值为Positive），第一类错误， Type I error。
    - False negative (FN)

      真实值为Positive，预测错误（预测值为 Negative），第二类错误， Type II error。
  - Precision

    Precision指得是**预测的正样本**中正确分类的比例。

$$precision = \frac{TP}{TP + FP}$$

  - Recall

    Recall指的是预测为正且预测正确的样本占所有正样本的概率。

$$recall = \frac{TP}{TP + FN}$$

  - F1-score

为了权衡precision和recall，引入了F1-score，F1-score是precision和recall的调和平均数。相比于算数平均，调和平均数的有点是只要有一个指标低，结果就会很低。如R与P中有一个为0另一个为1，若使用算数平均还是可以得到0.5，但是使用调和平均得到的是0。

$$F1score = \frac{2}{\frac{1}{p} + \frac{1}{R}} = \frac{2PR}{P+R}$$

- ROC (**Receiver Operating Characteristics**)
  - FPR（False positive rate），预测为正但实际为负的样本占所有负例样本的比例
  - TPR（(True positive rate），预测为正且实际为正的样本占所有正例样本的比例

  假设在二分类中，采用逻辑回归分类起，给定一个threshold如0.5，大于0.5的判定为正例，否则为负例，则对应可以算出一组(**FPR和TPR**)，随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着真正的负实例，即TPR和FPR会同时增大。阈值最大时，对应坐标点为(0,0)，阈值最小时，对应坐标点(1,1)。
- AUC(Area under curve)

  AUC其实就是在ROC曲线下的面积，一般用来比较两个模型，一般选取AUC较大的那个模型

- 回归模型度量

  对于回归模型，首先想到的是使用残差的均值来衡量模型的好坏，但是残差是有正有负的，可能会相互抵消，因此引出了MAE
  - MAE(mean absolute error)
    - MAE也称为L1损失范数。

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

    机器学习的本质是优化损失函数，绝对值函数的缺点是在某些点函数不光滑，在这些不光滑的点上不能求导，因此考虑将残差的绝对值改为残差的平方，由此引出均方误差。
  - MSE(mean squared error)

    MSE也称为L2损失范数。

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

    MSE虽然改进了MAE的缺点，但是存在一个量纲不一致的问题。因为将残差平方的同时也就将单位平方了，为改进这个问题，引入了RMSE即将MSE进行了开方操作。
  - RMSE(root MSE)

    RMSE主要的改进就是MSE的量纲问题，即对MSE直接开方

    $$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

    RMSE和其他的MSE、MAE等都这些指标存在的问题是和量纲有关，如对于同一个模型，预测的房产数据，误差是5w元；预测学生的成绩误差是10分，那么这个模型在哪个数据上表现的好就没有办法衡量，因此又引入了R-squared。
  - R squared

    $$R^2 = 1 - \frac{MSE(y, \hat{y})}{Var(y, \hat{y})}$$

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.
- 根据以上两条定义，我们大体可以知道R-squared其实就是因变量的变化可以被线性模型解释的百分比。一个很直接的理解为

  > R-squared = Explained variation / Total variation

  即：

  R-squared is always between 0 and 100%:

  - 0% indicates that the model explains none of the variability of the response data around its mean.
  - 100% indicates that the model explains all the variability of the response data around its mean.

参考

- 1.What does r squared tell us?
- 2.R-squared
- 3.AUC-ROC
- 4.机器学习常见衡量指标