

回归模型的评价

评价回归模型,一般是基于预测值和实际值的函数.

设定

对于回归模型,我们不妨设定n个样本,每个样本为 (x_i, y_i) ,预测为 $\hat{y}_i, i \in 1, 2, \dots, n$. \bar{y} 为 $\{y_i\}_{i=1}^n$ 的平均值.

平均绝对误差(MAE, Mean Absolute Error)

这个指标是对绝对误差损失的预期值.

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

平均绝对百分比误差(MAPE, Mean Absolute Percentage Error)

这个指标是对相对误差损失的预期值.所谓相对误差,就是绝对误差和真值的百分比.

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

均方误差(MSE, Mean Squared Error)

该指标对应于平方(二次)误差的期望.

$$\text{MSE}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n \|y_i - \hat{y}_i\|_2^2$$

均方误差根(RMSE, Root Mean Squared Error)

该指标对应于平方(二次)误差的期望.

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|y_i - \hat{y}_i\|_2^2}$$

均方误差对数(MSLE, Mean Squared Log Error)

该指标对应平方对数(二次)差的预期.

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

其中 $\log(x)$ 表示 x 的自然对数.当 y 具有指数增长的趋势时,该指标最适合使用,例如人口数量,跨年度商品的平均销售额等.MSLE实际上包含了对 y 的对数化处理.

中位绝对误差(MedAE, Median Absolute Error)

通过取目标和预测之间的所有绝对差值的中值来计算损失.相比MAE,对极端值不敏感.

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

R Squared(r2 score)

R Squared又叫可决系数(coefficient of determination)也叫拟合优度,反映的是自变量 x 对因变量 y 的变动的解释的程度.越接近于1,说明模型拟合得越好.

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

其中,

$$TSS(\text{Total Sum of Squares}) = \sum_{i=0}^n (y_i - \bar{y})^2$$

表示的是 y 的变动的程度,正比于方差.

$$RSS(\text{Residual Sum of Squares}) = \sum_{i=0}^n (y_i - \hat{y}_i)^2$$

表示的是模型和真实值的残差.

$$ESS(\text{Explained Sum of Squares}) = \sum_{i=0}^n (\hat{y}_i - \bar{y})^2$$

表示的是模型对 y 的变动的预测.

各个度量的对比

以上提到的度量,除了R Squared,都是越小越好. 使用哪种取决于目标,多种度量综合考虑则允许我们更深入得分析模型.

单个度量使用的情况

- MAE和MedAE基于绝对误差,如果看重真实值和预测值的绝对误差,则选用MAE或MedAE,两者分别是误差的均值和中位数.MAE对极端值比较敏感.
- 如果看重真实值和预测值的差的平方,则选用MSE或RMSE.
- 如果存在不同样本的真实值有量级差或者更加关注预测和真实值的百分比差异的情况,最好选用MAPE.
- 如果y具有随着x进行指数变动的趋势时,适合用MSLE.recall that $\log(1 + x)$ 是x的等阶无穷小和对x的对数化处理.
- 如果模型希望的是找到能够解释目标y变动的因变量,则选用R Squared更合适.

多个度量使用的情况

- MAE和RMSE一起使用,可以看出样本误差的离散程度.比如RMSE远大于MAE时,可以得知不同样例的误差差别很大.
- MAE和MAPE一起使用,再结合 \bar{y} ,可以估算模型对不同数量级样例的拟合程度.比如MAE远大于 $MAPE * \bar{y}$ 则可能是模型对真实值小的样本预测更准.可以考虑为不同数量级的样本建立不同的模型.

交叉检验中使用各种评价指标

1. 不论是回归模型还是分类模型,机器学习要达到的都是好的泛化能力.而泛化能力是由测试误差估计的泛化误差来评判的.
其中测试误差是测试集的预测值和真实值的统计量,也就是以上谈到的各种评价指标.
2. 我们不妨设 μ 为测试误差的均值, σ 为测试误差的标准差, α 为显著性水平,一般是5%.
假设测试误差和泛化误差服从正态分布,置信度95%的区间为 $(\mu - Z_{0.025} * \sigma, \mu + Z_{0.025} * \sigma)$
3. 我们认为A模型比B模型好,是A模型的误差区间上界小于B模型的误差区间下界.参看[樱花书中机器学习基础部分 \(http://www.deeplearningbook.org/contents/ml.html\)](http://www.deeplearningbook.org/contents/ml.html).
4. 所以我们不仅要关注各个评级指标的均值,也要关注方差,因为方差反映了模型泛化能力随着训练集改变的程度.

使用sklearn做模型评估

对于上面的这些指标,sklearn提供了部分接口来做分类模型的评估

	接口	说明
	<code>sklearn.metrics.mean_absolute_error(y_true, y_pred)</code>	平均绝对误差
	<code>sklearn.metrics.mean_squared_error(y_true, y_pred[, ...])</code>	均方误差
	<code>sklearn.metrics.mean_squared_log_error(y_true, y_pred)</code>	均方误差对数
	<code>sklearn.metrics.median_absolute_error(y_true, y_pred)</code>	中位绝对误差
	<code>sklearn.metrics.r2_score(y_true, y_pred[, ...])</code>	R ² score可决系数

MAPE我在metrics.py中有实现.

In []: