机器学习模型性能评估（一）：错误率与精度

原创 石头 机器学习算法那些事 2018-09-22

机器学习模型性能评估（一）：错误率与精度

机器学习是对给定的原始数据集构建最优学习模型，上篇文章讲到《机器学习模型评估方法》，通过模型评估方法将原始数据集划分训练集和测试集，训练集又可细分为训练集和验证集。机器学习的整个流程包括训练集和验证集构建最优模型，最优模型评估测试集的测试误差，通过测试误差来评价学习方法的泛化能力，泛化能力是评价机器学习模型性能的金标准（如图），即泛化能力强，对应的学习模型好。

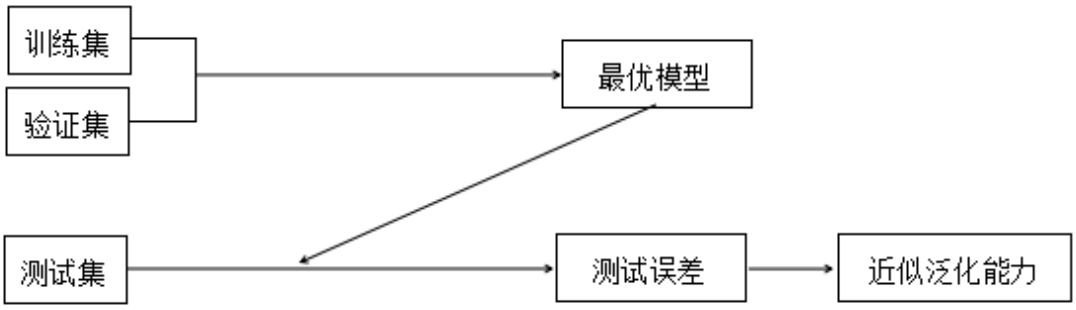

图1 机器学习流程

测试误差评价学习模型的泛化能力具有一定的局限性，抽样的测试数据集具有随机性，不同的测试集表现出不同测试误差，即评价模型的泛化能力也有所不同。

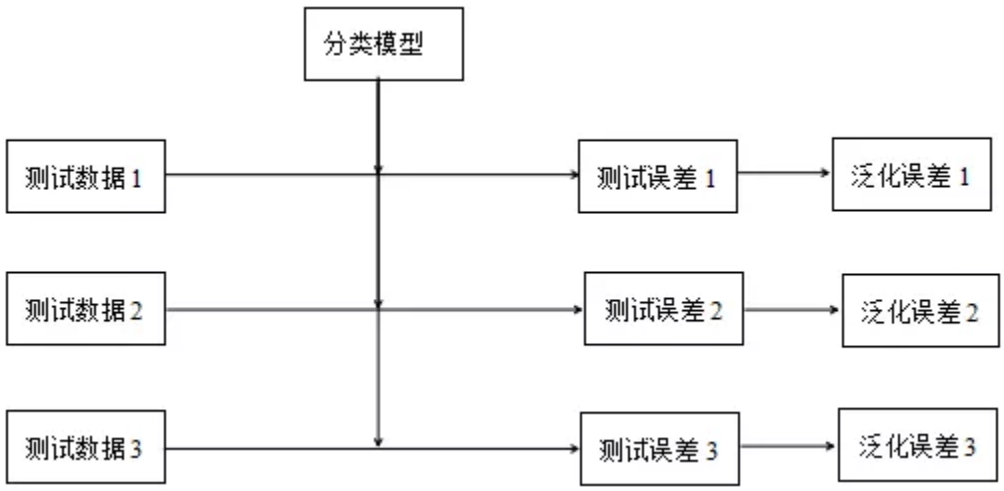

图2 测试数据集对泛化能力的影响

本文介绍了评价机器学习模型性能的四种方法：

（1）错误率与精度，错误率与精度是评价学习模型泛化能力的最常用的方法；

（2）从查准率和查全率的角度来评价学习模型泛化能力的优劣，并引用了P-R曲线和度量参数F1；

（3）ROC（Receiver Opreating Characteristic，受试者工作特征）曲线则评价“一般情况下”学习模型的泛化能力，并引用了度量参数AUC（Area Under Curve，曲线下的面积）；

（4）P-R曲线和ROC曲线认为学习器对不同类的分类错误产生的代价损失相同，则实际情况可能是不同类的分类错误产生的代价损失不相同，即非均衡代价，因此。从非均衡代价的角度去分析模型性能的优劣，并引用了代价曲线和期望总体代价。由于周老师编写的《机器学习》对代价曲线和期望总体代价描述的比较简练，因此，本文会详细去解释这两者的含义。

在介绍这四种方法前，需要理解期望和均值的概念，本文会首先引出期望和均值的概念，然后介绍评价机器学习模型性能的四种方法，最后对本文进行总结。

1、期望和均值

期望和均值这两个概念，相信在看我这篇文章的童鞋都不陌生，最近在自学贝叶斯概率的时候发现自己并没有充分理解期望和均值的概念，且这篇文章很多知识点要涉及到期望和均值的思想，因此，本节简单介绍了期望和均值的定义。

假设某一离散变量X的取值范围来自于集合A，A={X1,X2,X3,X4,...XN}，对集合A进行可放回抽样M次（参考上节），产生容量为M的抽样数据集S，数据集S的离散变量X的取值为：{X(1),X(2),X(3),...X(M)}，X(K)表示第K次可放回抽样的值。

变量X期望E(X)：

$$E(X) = \sum_{n=1}^{N} Xn * p(Xn)$$

由上式可知，期望是累加变量所有可能的取值与对应该值出现概率的乘积。

抽样数据集S的变量X的均值 $\overline{X}$：

$$\overline{X}=\frac{1}{M}\sum_{k=1}^{M}X(k)$$

均值是所有抽样数据变量的求和平均。

实际工作项目中，我们获取的数据都是抽样数据，只能够通过抽样数据来计算均值，期望是计算不出来的，若知道某一个变量的真实分布，马上就能知道该变量的期望，那么训练数据也没有什么存在的意义了。机器学习的一个重要内容就是通过已知的抽样数据集去估计总体的分布。

辛钦大数定理：

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}=1$$

辛钦大数定理：若抽样数据的样本量足够大，那么变量均值等于期望。

2、模型性能评估方法

模型性能评估即评估模型的泛化能力，本节介绍四种性能评估方法，分别为错误率与精度评估方法、P-R曲线评估法，ROC曲线评估法以及代价曲线评估法，根据实际应用的侧重点选择性能评估方法。

2.1 错误率与精度

错误率与精度常用于分类任务，错误率是测试样本中分类错误的样本数占总样本数的比例，精度是测试样本中分类正确的样本数占总样本数的比例。
假设数据集D，模型f：
分类错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$$

精度：

$$\begin{aligned} \mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)=y_i) \\ &= 1-E(f;D)\ . \end{aligned}$$

更一般地，若对于数据分布D和概率密度函数p(·)，错误率与精度可分别描述为：

$$E(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}(f(\boldsymbol{x})\neq y)\,p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$$

$$\begin{aligned} \mathrm{acc}(f;\mathcal{D}) &= \int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}(f(\boldsymbol{x})=y)\,p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= 1-E(f;\mathcal{D})\ . \end{aligned}$$

若测试数据集的精度高或错误率小，则模型的泛化能力强；反之，则泛化能力弱。

用测试数据集的精度来表示模型的泛化能力在正负样本比例相差较大的时候就不适用了。

如：

真实情况	预测结果	
	正例	反例
正例	100	0
反例	1	0

上表的混淆矩阵可知:

测试数据集正样本和负样本的比例是100：1，把所有样本都检测为正样本，准确率99%，但不能说明模型的泛化能力强，因为不能反映模型对负样本的检测能力。运用精度来表示测试数据集的泛化能力，测试数据集的正负样本比例应该均衡（1:1）。

未完，待续。。。