[**机器学习评价指标大汇总**](http://www.36dsj.com/archives/42271)

* 雪姬
* 2016-03-03 7:32:08
* [机器学习](http://www.36dsj.com/archives/category/large-data-tools/machine-learning)
* 评论(1)

作者：无影随想

在使用机器学习算法的过程中，针对不同场景需要不同的评价指标，在这里对常用的指标进行一个简单的汇总。

**一、分类**

**1. 精确率与召回率**

精确率与召回率多用于二分类问题。精确率(Precision)指的是模型判为正的所有样本中有多少是真正的正样本;召回率(Recall)指的是所有正样本有多少被模型判为正样本，即召回。设模型输出的正样本集合为$A$，真正的正样本集合为$B$，则有：

$\text{Precision}(A,B)=\frac{|A\bigcap B|}{|A|},\text{Recall}(A,B)=\frac{|A\bigcap B|}{|B|}$。

有时候我们需要在精确率与召回率间进行权衡，一种选择是画出精确率-召回率曲线(Precision-Recall Curve)，曲线下的面积被称为AP分数(Average precision score);另外一种选择是计算$F\_{\beta}$分数：

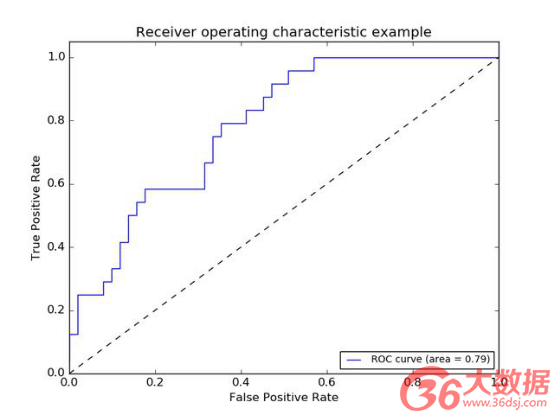
$F\_{\beta}=(1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{\beta^2\cdot\text{precision}+\text{recall}}$。

当$\beta=1$称为$F\_1$分数，是分类与信息检索中最常用的指标之一。

**2. ROC**

设模型输出的正样本集合为$A$，真正的正样本集合为$B$，所有样本集合为$C$，我们称$\frac{|A\bigcap B|}{|B|}$为真正率(True-positive rate)，$\frac{|A- B|}{|C-B|}$为假正率(False-positive rate)。

ROC曲线适用于二分类问题，以假正率为横坐标，真正率为纵坐标的曲线图，如：



AUC分数是曲线下的面积(Area under curve)，越大意味着分类器效果越好。

**3. 对数损失**

对数损失(Log loss)亦被称为逻辑回归损失(Logistic regression loss)或交叉熵损失(Cross-entropy loss)。

对于二分类问题，设$y\in\{0,1\}$且$p={\rm Pr}(y=1)$，则对每个样本的对数损失为：

$L\_{\rm log}(y,p)=-\log{\rm Pr}(y|p)=-(y\log(p)+(1-y)\log(1-p))$。

可以很容易地将其扩展到多分类问题上。设$Y$为指示矩阵，即当样本$i$的分类为$k$时$y\_{i,k}=1$;设$P$为估计的概率矩阵，即$p\_{i,k}={\rm Pr}(t\_{i,k}=1)$，则对每个样本的对数损失为：

$L\_{\log}(Y\_i,P\_i)=-\log{\rm Pr}(Y\_i|P\_i)=\sum\limits\_{k=1}^{K}y\_{i,k}\log p\_{i,k}$。

**4. 铰链损失**

铰链损失(Hinge loss)一般用来使“边缘最大化”(maximal margin)。

铰链损失最开始出现在二分类问题中，假设正样本被标记为1，负样本被标记为-1，$y$是真实值，$w$是预测值，则铰链损失定义为：

$L\_{\text{Hinge}}(w, y)=\max\{1-wy,0\}=|1-wy|\_+$。

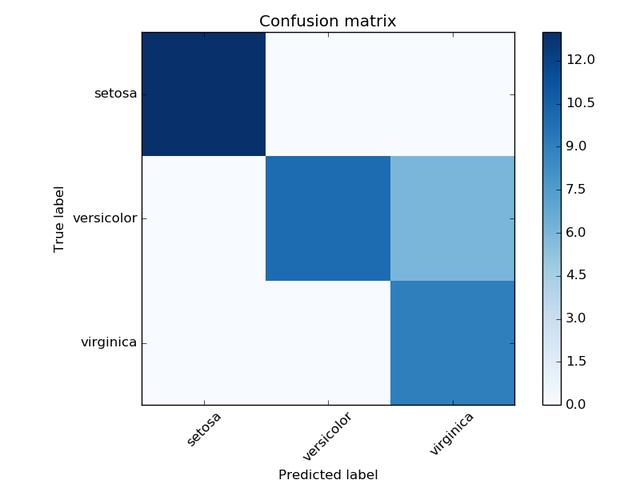
然后被扩展到多分类问题，假设$y\_w$是对真实分类的预测值，$y\_t$是对非真实分类预测中的最大值，则铰链损失定义为：

$L\_{\text{Hinge}}(y\_w, y\_t)=\max\{1+y\_t-y\_w,0\}$。

注意，二分类情况下的定义并不是多分类情况下定义的特例。

**5. 混淆矩阵**

混淆矩阵(Confusion Matrix)又被称为错误矩阵，通过它可以直观地观察到算法的效果。它的每一列是样本的预测分类，每一行是样本的真实分类(反过来也可以)，顾名思义，它反映了分类结果的混淆程度。混淆矩阵$i$行$j$列的原始是原本是类别$i$却被分为类别$j$的样本个数，计算完之后还可以对之进行可视化：



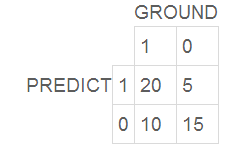
**6. kappa系数**

kappa系数(Cohen’s kappa)用来衡量两种标注结果的吻合程度，标注指的是把N个样本标注为C个互斥类别。计算公式为

$\mathcal{K}=\frac{p\_o-p\_e}{1-p\_e}=1-\frac{1-p\_o}{1-p\_e}$。

其中$p\_o$是观察到的符合比例，$p\_e$是由于随机性产生的符合比例。当两种标注结果完全相符时，$\mathcal{K}=1$，越不相符其值越小，甚至是负的。

是不是云里来雾里去的，现在举个栗子，对于50个测试样本的二分类问题，预测与真实分布情况如下表：



预测与真实值相符共有20+15个，则观察到的符合比例为$p\_o=(20+15)/50=0.7$。计算$p\_e$比较复杂，PREDICT预测为1的比例为0.5，GROUND中1的比例为0.6，从完全随机的角度来看，PREDICT与GROUND均为1的概率为0.5\*0.6=0.3，PREDICT与GROUND均为0的概率为0.5\*0.4=0.2，则PREDICT与GROUND由于随机性产生的符合比例为0.2+0.3=0.5，即$p\_e=0.5$，最后求得$\mathcal{K}=\frac{p\_o-p\_e}{1-p\_e}=\frac{0.7-0.5}{1-0.5}=0.4$。

**7. 准确率**

准确率(Accuracy)衡量的是分类正确的比例。设$\hat{y}\_i$是是第$i$个样本预测类别，$y\_i$是真是类别，在$n\_{\rm sample}$个测试样本上的准确率为

${\rm accuracy}=\frac{1}{n\_{\rm sample}}\sum\limits\_{i=1}^{n\_{\rm sample}}1(\hat{y}\_i=y\_i)$。

其中$1(x)$是indicator function，当预测结果与真实情况完全相符时准确率为1，两者越不相符准确率越低。

虽然准确率适用范围很广，可用于多分类以及多标签等问题上，但在多标签问题上很严格，在有些情况下区分度较差。

**8. 海明距离**

海明距离(Hamming Distance)用于需要对样本多个标签进行分类的场景。对于给定的样本$i$，$\hat{y}\_{ij}$是对第$j$个标签的预测结果，${y}\_{ij}$是第$j$个标签的真实结果，$L$是标签数量，则$\hat{y}\_i$与$y\_i$间的海明距离为

$D\_{Hamming}(\hat{y}\_i,y\_i)=\frac{1}{L}\sum\limits\_{j=1}^L 1(\hat{y}\_{ij}\neq y\_{ij})$。

其中$1(x)$是indicator function。当预测结果与实际情况完全相符时，距离为0;当预测结果与实际情况完全不符时，距离为1;当预测结果是实际情况的真子集或真超集时，距离介于0到1之间。

我们可以通过对所有样本的预测情况求平均得到算法在测试集上的总体表现情况，当标签数量$L$为1时，它等于1-Accuracy，当标签数$L>1$时也有较好的区分度，不像准确率那么严格。

**9.杰卡德相似系数**

杰卡德相似系数(Jaccard similarity coefficients)也是用于需要对样本多个标签进行分类的场景。对于给定的样本$i$，$\hat{y}\_i$是预测结果，${y}\_i$是真实结果，$L$是标签数量，则第$i$个样本的杰卡德相似系数为

$J(\hat{y}\_i,y\_i)=\frac{|\hat{y}\_i\bigcap y\_i|}{|\hat{y\_i}\bigcup y\_i|}$。

它与海明距离的不同之处在于分母。当预测结果与实际情况完全相符时，系数为1;当预测结果与实际情况完全不符时，系数为0;当预测结果是实际情况的真子集或真超集时，距离介于0到1之间。

我们可以通过对所有样本的预测情况求平均得到算法在测试集上的总体表现情况，当标签数量$L$为1时，它等于Accuracy。

**10. 多标签排序**

在这节我们介绍一些更精细化的多标签分类效果衡量工具。设真实标签分类情况为$y\in\{0, 1\}^{n\_\text{samples} \times n\_\text{labels}}$，分类器预测情况为$\hat{f}\in\mathbb{R}^{n\_\text{samples} \times n\_\text{labels}}$。

**10.1 涵盖误差**

涵盖误差(Coverage error)计算的是预测结果中平均包含多少真实标签，适用于二分类问题。涵盖误差定义为：

$coverage(y,\hat{f})=\frac{1}{n\_\text{samples}}\sum\limits\_{i=1}^{n\_\text{samples}}\max\limits\_{j:y\_{ij}=1}\text{rank}\_{ij}$，

其中$\text{rank}\_{ij}=\left|\left\{k:\hat{f}\_{ik}\ge \hat{f}\_{ij} \right\}\right|$。可以看到它实际衡量的是真实标签中有多少排在预测结果的前面。

**10.2 标签排序平均精度**

标签排序平均精度(Label ranking average precision)简称LRAP，它比涵盖误差更精细：

$LRAP(y,\hat{f})=\frac{1}{n\_\text{samples}}\sum\limits\_{i=1}^{n\_\text{samples}}\frac{1}{|y\_i|}\sum\limits\_{j:y\_{ij}=1}\frac{|\mathcal{L}\_{ij}|}{\text{rank}\_{ij}}$，

其中$\mathcal{L}\_{ij}=\left\{k:y\_{ik}=1,\hat{f}\_{ik}\ge\hat{f}\_{ij}\right\}$，$\text{rank}\_{ij}=\left|\left\{k:\hat{f}\_{ik}\ge \hat{f}\_{ij} \right\}\right|$。

**10.3 排序误差**

排序误差(Ranking loss)进一步精细考虑排序情况：

$ranking(y,\hat{f})=\frac{1}{n\_\text{samples}}\sum\limits\_{i=1}^{n\_\text{samples}}\frac{1}{|y\_i|(n\_\text{labels}-|y\_i|))}\left|\mathcal{L}\_{ij} \right|$，

其中$\mathcal{L}\_{ij}=\left\{(k,l):\hat{f}\_{ik}<\hat{f}\_{ij}, y\_{ik}=1, y\_{il}=0\right\}$。

**二、拟合**

拟合问题比较简单，所用到的衡量指标也相对直观。假设$y\_i$是第$i$个样本的真实值，$\hat{y}\_i$是对第$i$个样本的预测值。

**1. 平均绝对误差**

平均绝对误差MAE(Mean Absolute Error)又被称为$l1$范数损失($l1$-norm loss)：

${\rm MAE}(y, \hat{y})=\frac{1}{n\_{\rm samples}}\sum\limits\_{i=1}^{n\_{\rm samples}}|y\_i-\hat{y}\_i|$。

**2. 平均平方误差**

平均平方误差MSE(Mean Squared Error)又被称为$l2$范数损失($l2$-norm loss)：

${\rm MSE}(y, \hat{y})=\frac{1}{n\_{\rm samples}}\sum\limits\_{i=1}^{n\_{\rm samples}}(y\_i-\hat{y}\_i)^2$。

**3. 解释变异**

解释变异(Explained variance)是根据误差的方差计算得到的：

${\rm explained variance}(y,\hat{y})=1-\frac{{\rm Var}\{y-\hat{y}\}}{{\rm Var}{y}}$。

**4. 决定系数**

决定系数(Coefficient of determination)又被称为$R^2$分数：

$R^2(y,\hat{y})=1-\frac{\sum\_{i=1}^{n\_{\rm samples}}(y\_i-\hat{y}\_i)^2}{\sum\_{i=1}^{n\_{\rm samples}}(y\_i-\bar{y})^2}$，

其中$\bar{y}=\frac{1}{n\_{\rm samples}}\sum\_{i=1}^{n\_{\rm samples}}y\_i$。

**三、聚类**

**1 . 兰德指数**

兰德指数(Rand index)需要给定实际类别信息$C$，假设$K$是聚类结果，$a$表示在$C$与$K$中都是同类别的元素对数，$b$表示在$C$与$K$中都是不同类别的元素对数，则兰德指数为：

${\rm RI}=\frac{a+b}{C\_2^{n\_{\rm samples}}}$，

其中$C\_2^{n\_{\rm samples}}$数据集中可以组成的总元素对数，RI取值范围为$[0,1]$，值越大意味着聚类结果与真实情况越吻合。

对于随机结果，RI并不能保证分数接近零。为了实现“在聚类结果随机产生的情况下，指标应该接近零”，调整兰德系数(Adjusted rand index)被提出，它具有更高的区分度：

${\rm ARI}=\frac{{\rm RI}-E[{\rm RI}]}{\max({\rm RI})-E[{\rm RI}]}$，

具体计算方式参见Adjusted Rand index。

ARI取值范围为$[-1,1]$，值越大意味着聚类结果与真实情况越吻合。从广义的角度来讲，ARI衡量的是两个数据分布的吻合程度。

**2. 互信息**

互信息(Mutual Information)也是用来衡量两个数据分布的吻合程度。假设$U$与$V$是对$N$个样本标签的分配情况，则两种分布的熵(熵表示的是不确定程度)分别为：

$H(U)=\sum\limits\_{i=1}^{|U|}P(i)\log (P(i)), H(V)=\sum\limits\_{j=1}^{|V|}P'(j)\log (P'(j))$，

其中$P(i)=|U\_i|/N,P'(j)=|V\_j|/N$。$U$与$V$之间的互信息(MI)定义为：

${\rm MI}(U,V)=\sum\limits\_{i=1}^{|U|}\sum\limits\_{j=1}^{|V|}P(i,j)\log\left ( \frac{P(i,j)}{P(i)P'(j)}\right )$，

其中$P(i,j)=|U\_i\bigcap V\_j|/N$。标准化后的互信息(Normalized mutual information)为：

${\rm NMI}(U,V)=\frac{{\rm MI}(U,V)}{\sqrt{H(U)H(V)}}$。

与ARI类似，调整互信息(Adjusted mutual information)定义为：

${\rm AMI}=\frac{{\rm MI}-E[{\rm MI}]}{\max(H(U), H(V))-E[{\rm MI}]}$。

利用基于互信息的方法来衡量聚类效果需要实际类别信息，MI与NMI取值范围为$[0,1]$，AMI取值范围为$[-1,1]$，它们都是值越大意味着聚类结果与真实情况越吻合。

**3. 轮廓系数**

轮廓系数(Silhouette coefficient)适用于实际类别信息未知的情况。对于单个样本，设$a$是与它同类别中其他样本的平均距离，$b$是与它距离最近不同类别中样本的平均距离，轮廓系数为：

$s=\frac{b-a}{\max(a,b)}$。

对于一个样本集合，它的轮廓系数是所有样本轮廓系数的平均值。

轮廓系数取值范围是$[-1,1]$，同类别样本越距离相近且不同类别样本距离越远，分数越高。

**四、信息检索**

信息检索评价是对信息检索系统性能(主要满足用户信息需求的能力)进行评估，与机器学习也有较大的相关性，感兴趣的可以参考这篇不错的博文。

**五、总结**

上面介绍了非常多的指标，实际应用中需要根据具体问题选择合适的衡量指标。那么具体工作中如何快速使用它们呢?优秀的Python机器学习开源项目Scikit-learn实现了上述绝指标的大多数，使用起来非常方便。