VergiL Wang的专栏

个人资料

wangran51

访问： 537149次
积分： 5493
等级： BLOG 6
排名： 第2542名

原创： 32篇
转载： 304篇
译文： 0篇
评论： 95条

文章搜索

文章分类

Machine Learning (43)
Linux (92)
Algorithm (39)
JAVA (15)
Recommend System (2)
Python (43)
Natural Language Process (15)
perl (4)
Matrix Laboratory (1)
Crawler (3)
Larbin (3)
Regular Expression (5)
DataBase (9)
CPlus (55)
Interview (47)
shell (8)
JavaScript (6)
Java Web (7)
Web Knowledge (5)
Cloud (4)
nodejs (0)
Dynamic Programming (2)

Logistic回归

2013-05-06 22:52 15941人阅读 评论(2) 收藏 举报

分类： Machine Learning（42）

转自别处 有很多与此类似的文章 也不知道谁是原创 因原文由少于错误 所以下文对此有修改并且做了适当的重点标记(横线见的内容没大明白 并且有些复杂，后面的运行流程依据前面的得出的算子进行分类)

初步接触

谓LR分类器(Logistic Regression Classifier)，并没有什么神秘的。在分类的情形下，经过学习之后的LR分类器其实就是一组权值w0,w1,...,wm.
当测试样本集中的测试数据来到时，这一组权值按照与测试数据线性加和的方式，求出一个z值：

z = w0+w1*x1+w2*x2+...+wm*xm。 ①（其中x1,x2,...,xm是某样本数据的各个特征，维度为m）

之后按照sigmoid函数的形式求出：

σ(z) = 1 / (1+exp(z)) 。②

由于sigmoid函数的定义域是(-INF, +INF),而值域为(0, 1)。因此最基本的LR分类器适合于对两类目标进行分类。

那么LR分类器的这一组权值w0,w1,...,wm是如何求得的呢？这就需要涉及到极大似然估计MLE和优化算法的概念了。

我们将sigmoid函数看成样本数据的概率密度函数，每一个样本点，都可以通过上述的公式①和②计算出其概率密度

详细描述

1.逻辑回归模型

1.1逻辑回归模型

考虑具有p个独立变量的向量 $x'=(x_1,x_2,\cdots x_p)$,设条件概率 $P(Y=1|x)=p$ 为根据观测量相对于某事件发生的概率。逻辑回归模型可表示为

$$P(Y=1|x)=\pi(x)=\frac{1}{1+e^{-g(x)}} \qquad (1.1)$$

上式右侧形式的函数称为称为逻辑函数。下图给出其函数图象形式。

文章存档

2014年02月 (1)
2014年01月 (1)
2013年12月 (1)
2013年10月 (2)
2013年09月 (5)
展开

阅读排行

SVD奇异值分解 (57683)
C++ Set常用用法 (35182)
Python Dict用法 (18312)
Logistic回归 (15913)
PlayFramework入门教程 (15781)
准确率召回率 (14460)
Latent semantic analysis (11955)
拉格朗日 SVM KKT (7802)
Latent dirichlet allocation (7279)
Linux下获取毫秒级时间 (6511)

评论排行

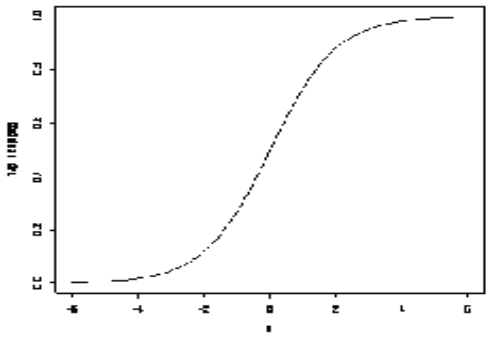
2011阿里巴巴集团实习生 (23)
拉格朗日 SVM KKT (8)
SVD奇异值分解 (8)
准确率召回率 (7)
Latent semantic analysis (5)
Latent dirichlet allocation (5)
JAVA自动补全插件 (4)
学习SVM (3)
隐马尔科夫模型HMM自学 (3)
在 N 条水平线与 M 条竖 (3)

推荐文章

*App竞品技术分析 （6）热修复
*架构设计：系统间通信（17）——服务治理与Dubbo 中篇（分析）
*你的计划为什么执行不下去？怎么破？
*图解堆算法、链表、栈与队列（多图预警）
*【android】仿360手机卫士的简易设计思路及源码
*Android平台Camera实时滤镜实现方法探讨(九)--磨皮算法探讨(一)

最新评论

哈希表等概率情况下查找成功和
whale_yu: 在查找不成功时：地址5和6的计算是否有有错。上述是计算到9的位置，可是备注却是到2的位置？
准确率召回率
另一只蝴蝶: 很有帮助！
在 N 条水平线与 M 条竖直线构成
qq_33425990: #include #include int main(){ int T, N, M, K; ...
SVD奇异值分解
programming2015: 分解的图有点问题



其中 $g(x)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$。如果含有名义变量，则将其变为dummy变量。一个具有k个取值的名义变量，将变为k-1个dummy变量。这样，有

$$g(x)=\beta_0+\beta_1x_1+\cdots+\sum_{i=1}^{k-1}\beta_{ji}D_{ji}+\beta_px_p \quad (1.2)$$

定义不发生事件的条件概率为

$$P(Y=0\mid x)=1-P(Y=1\mid x)=1-\frac{e^{g(x)}}{1+e^{g(x)}}=\frac{1}{1+e^{g(x)}} \quad (1.3)$$

那么，事件发生与事件不发生的概率之比为

$$\frac{P(x=1\mid x)}{P(x=0\mid x)}=\frac{p}{1-p}=e^{g(x)} \quad (1.4)$$

这个比值称为事件的发生比(the odds of experiencing an event),简称为odds。因为0<p<1,故odds>0。对odds取对数，即得到线性函数，

$$\log(\frac{p}{1-p})=\beta_0+\beta_1x_1+\cdots+\sum_{i=1}^{k-1}\beta_{ji}D_{ji}+\beta_px_p \quad (1.5)，$$

1.2极大似然函数

假设有n个观测样本，观测值分别为 $y_1,y_2,\cdots,y_n$，设 $p_i=P(y_i=1\mid x_i)$ 为给定条件下得到yi=1（原文 $p_i=1$ ）的概率。在同样条件下得到yi=0（ $p_i=0$ ）的条件概率为 $P(y_i=0\mid x_i)=1-p_i$。于是，得到一个观测值的概率为

$$P(y_i)=p_i^{y_i}(1-p_i)^{(1-y_i)} \quad (1.6)$$

-----此公式实际上是综合前两个等式得出，并无特别之处

因为各项观测独立，所以它们的联合分布可以表示为各边际分布的乘积。

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

上式称为n个观测的似然函数。我们的目标是能够求出使这一似然函数的值最大的参数估计。于是，最大似然估计的关键就是求出参数 $\beta_0,\beta_1,\cdots,\beta_p$，使上式取得最大值。

对上述函数求对数

$$L(\beta)=\ln[l(\beta)]=\sum_{i=1}^{n}(y_i\ln[\pi(x_i)]+(1-y_i)\ln[1-\pi(x_i)]) \quad (1.8)$$

上式称为对数似然函数。为了估计能使 $L(\beta)$ 取得最大的参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的值。

对此函数求导，得到p+1个似然方程。

$$\sum_{i=1}^{n}[y_i-\pi(x_i)]=\sum_{i=1}^{n}[y_i-\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}]=0 \quad (1.9)$$

$$\sum_{i=1}^{n} x_{ij}[y_i-\pi(x_i)]=\sum_{i=1}^{n}[y_i-\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}]=0$$ ，j=1,2,..,p.-----p为独立向量个数

上式称为似然方程。为了解上述非线性方程，应用牛顿－拉斐森(Newton-Raphson)方法进行迭代求解。

1.3　牛顿－拉斐森迭代法

对 $L(\beta)$ 求二阶偏导数，即Hessian矩阵为

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^{\ 2}}=-\sum_{i=1}^{n} x_{ij}^{\ 2}\pi_i(1-\pi_i)$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l}=-\sum_{i=1}^{n} x_{ij}x_{il}\pi_i(1-\pi_i) \qquad (1.10)$$

如果写成矩阵形式，以H表示Hessian矩阵， X表示

$$X=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix} \qquad (1.11)$$

令

$$V=\begin{bmatrix}\hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0\\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0\\ & 0 & \ddots & \vdots\\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n)\end{bmatrix} \qquad (1.12)$$

则 $H=X^TVX$ 。再令 $U=\begin{bmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{bmatrix}*\begin{bmatrix}y_1-\pi_1\\ y_2-\pi_2\\ \vdots\\ y_p-\pi_p\end{bmatrix}$ (注：前一个矩阵需转置)，即似然方程的矩阵形式。

得牛顿迭代法的形式为

$$W_{new}=W_{old}-H^{-1}U \qquad (1.13)$$

注意到上式中矩阵H为对称正定的，求解 $H^{-1}U$ 即为求解线性方程H X＝U中的矩阵X。对H进行cholesky分解。

最大似然估计的渐近方差（asymptotic variance）和协方差(covariance)可以由信息矩阵（information matrix）的逆矩阵估计出来。而信息矩阵实际上是 $L(\beta)$ 二阶导数的负值，表示为 $I=\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l}$ 。估计值的方差和协方差表示为 $\mathrm{var}(\beta)=I^{-1}$ ，也就是说，估计值 β_j 的方差为矩阵 I 的逆矩阵的对角线上的值，而估计值 β_j 和 β_l 的协方差(β_j 和 β_l 的协方差等于 $I=\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l}$ ？不解。。。)为除了对角线以外的值。然而在多数情况，我们将使用估计值 β_j 的标准方差，表示为

$SE(\beta_j)=(\mathrm{var}(\beta_j))^{\frac{1}{2}}$ ，for j=0,1,2,…,p　　（1.14）

2.显著性检验

下面讨论在逻辑回归模型中自变量 x_k 是否与反应变量显著相关的显著性检验。零假设 H_0 ： $\beta_k=0$ （表示自变量 x_k 对事件发生可能性无影响作用）。如果零假设被拒绝，说明事件发生可能性依赖于 x_k 的变化。

2.1 Wald test

对回归系数进行显著性检验时，通常使用Wald检验，其公式为

$W=[\hat{\beta}_j/S\hat{E}(\hat{\beta}_j)]^2$ （2.1）

其中，$S\hat{E}(\hat{\beta}_j)$为$\hat{\beta}_j$的标准误差。这个单变量Wald统计量服从自由度等于 1 的λ^2分布。

如果需要检验假设H_0：$\beta_1=\beta_2=\cdots=\beta_p$=0,计算统计量

$$W=[\hat{\beta}'/S\hat{E}(\hat{\beta}')]^2 \quad (2.2)$$

其中，$\hat{\beta}'$为去掉$\hat{\beta}_0$所在的行和列的估计值，相应地，$S\hat{E}(\hat{\beta}')$为去掉$\hat{\beta}_0$所在的行和列的标准误差。这里，Wald统计量服从自由度等于p的λ^2分布。如果将上式写成矩阵形式，有

$$W=(Q\hat{\beta})'[Q\,\mathrm{var}(\hat{\beta})Q']^{-1}(Q\hat{\beta}) \quad (2.3)$$

矩阵 Q 是第一列为零的一常数矩阵。例如，如果检验$\beta_1=\beta_2=0$，则$Q=\begin{bmatrix}0&1&0\\0&0&1\end{bmatrix}$。

然而当回归系数的绝对值很大时，这一系数的估计标准误就会膨胀，于是会导致Wald统计值变得很小，以致第二类错误的概率增加。也就是说，在实际上会导致应该拒绝零假设时却未能拒绝。所以当发现回归系数的绝对值很大时，就不再用Wald统计值来检验零假设，而应该使用似然比检验来代替。

2.2　似然比（Likelihood ratio test）检验

在一个模型里面，含有变量x_i与不含变量x_i的对数似然值乘以-2的结果之差，服从χ^2分布。这一检验统计量称为似然比(likelihood ratio)，用式子表示为

$$G=-2\ln\left(\frac{\text{不含}x_i\text{似然}}{\text{含有}x_i\text{似然}}\right) \quad (2.4)$$

计算似然值采用公式（1.8）。

倘若需要检验假设H_0：$\beta_1=\beta_2=\cdots=\beta_p$=0,计算统计量

$$G=2\left[\sum_{i=1}^{n}y_i\ln(\hat{\pi}_i)+(1-y_i)\ln(1-\hat{\pi}_i)\right]-[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)] \quad (2.5)$$

式中，n_0表示y_i=0的观测值的个数，而n_1表示y_i=1的观测值的个数，那么n就表示所有观测值的个数了。实际上，上式的右端的右半部分$[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)]$表示只含有$\beta_0$的似然值。统计量G服从自由度为p的$\chi^2$分布

2.3 Score检验

在零假设H_0：β_k=0下，设参数的估计值为$\beta_{(0)}$，即对应的β_k=0。计算Score统计量的公式为

$$U(\beta_{(0)})^T I^{-1}(\beta_{(0)})U(\beta_{(0)}) \quad (2.6)$$

上式中，$U(\beta_{(0)})$表示在β_k=0下的对数似然函数（1.9）的一价偏导数值，而$I(\beta_{(0)})$表示在β_k=0下的对数似然函数（1.9）的二价偏导数值。Score统计量服从自由度等于 1 的λ^2分布。

2.4　模型拟合信息

模型建立后，考虑和比较模型的拟合程度。有三个度量值可作为拟合的判断根据。

(1)-2LogLikelihood

$$-2LogL=-2\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\hat{\pi}_i}{y_i}\right)+(1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \quad (2.7)$$

(2) Akaike信息准则（Akaike Information Criterion,简写为AIC）

$$AIC=-2LogL+2(K+S) \quad (2.8)$$

其中 K 为模型中自变量的数目，S 为反应变量类别总数减 1，对于逻辑回归有S=2-1=1。-2LogL的值域为0至∞，其值越小说明拟合越好。当模型中的参数数量越大时，似然值也就越大，-2LogL就变小。因此，将 2 (K+S)加到AIC公式中以抵销参数数量产生的影响。在其它条件不变的情况下，较小的AIC值表示拟合模型较好。

(3)Schwarz准则

这一指标根据自变量数目和观测数量对-2LogL值进行另外一种调整。SC指标的定义为

$$SC = -2LogL + 2(K+S)*\ln(n)$$ (2.9)

其中ln(n)是观测数量的自然对数。这一指标只能用于比较对同一数据所设的不同模型。在其它条件相同时，一个模型的AIC或SC值越小说明模型拟合越好。

3.回归系数解释

3.1发生比

odds=[p/(1-p)]= $e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k+\cdots\beta_p x_p}$，即事件发生的概率与不发生的概率之比。而发生比率(odds ration),即

$$OR = \frac{odds_i}{odds_j}$$

(1)连续自变量。对于自变量 x_k，每增加一个单位，odds ration为

$$OR = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k(x_k+1)+\cdots\beta_p x_p}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k+\cdots\beta_p x_p}}$$

$$= e^{\beta_k}$$ (3.1)

(2)二分类自变量的发生比率。变量的取值只能为0或1，称为dummy variable。当 x_k 取值为1，对于取值为0的发生比率为

$$OR = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k *1+\cdots\beta_p x_p}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k *0+\cdots\beta_p x_p}}$$

$$= e^{\beta_k}$$ (3.2)

亦即对应系数的幂。

(3)分类自变量的发生比率。

如果一个分类变量包括m个类别，需要建立的dummy variable的个数为m-1,所省略的那个类别称作参照类(reference category)。设dummy variable为 x_k，其系数为 β_k，对于参照类，其发生比率为 e^{β_k}。

3.2 逻辑回归系数的置信区间

对于置信度 1 - α，参数 β_k 的100%（ 1 - α ）的置信区间为

$$\hat{\beta}_k \pm Z_{\frac{\alpha}{2}} \times SE_{\hat{\beta}_k}$$ （3.3）

上式中，$Z_{\frac{\alpha}{2}}$ 为与正态曲线下的临界 Z 值（critical value），$SE_{\hat{\beta}_k}$ 为系数估计 $\hat{\beta}_k$ 的标准误差，$\hat{\beta}_k - Z_{\frac{\alpha}{2}} \times SE_{\hat{\beta}_k}$ 和 $\hat{\beta}_k + Z_{\frac{\alpha}{2}} \times SE_{\hat{\beta}_k}$ 两值便分别是置信区间的下限和上限。当样本较大时，α =0.05水平的系数 $\hat{\beta}_k$ 的95%置信区间为

$$\hat{\beta}_k \pm 1.96 \times SE_{\hat{\beta}_k}$$ （3.4）

4.变量选择

4.1前向选择（forward selection）：在截距模型的基础上，将符合所定显著水平的自变量一次一个地加入模型。

具体选择程序如下

（1） 常数（即截距）进入模型。

（2） 根据公式（2.6）计算待进入模型变量的Score检验值，并得到相应的P值。

（3） 找出最小的p值，如果此p值小于显著性水平 α_{in},则此变量进入模型。如果此变量是某个名义变量的单面化(dummy)变量，则此名义变量的其它单面化变理同时也进入模型。不然，表明没有变量可被选入模型。选择过程终止。

（4） 回到(2)继续下一次选择。

4.2 后向选择（backward selection）：在模型包括所有候选变量的基础上，将不符合保留要求显著水平的自变量一次一个地删除。

具体选择程序如下

(1) 所有变量进入模型。

(2) 根据公式（2.1）计算所有变量的Wald检验值，并得到相应的p值。

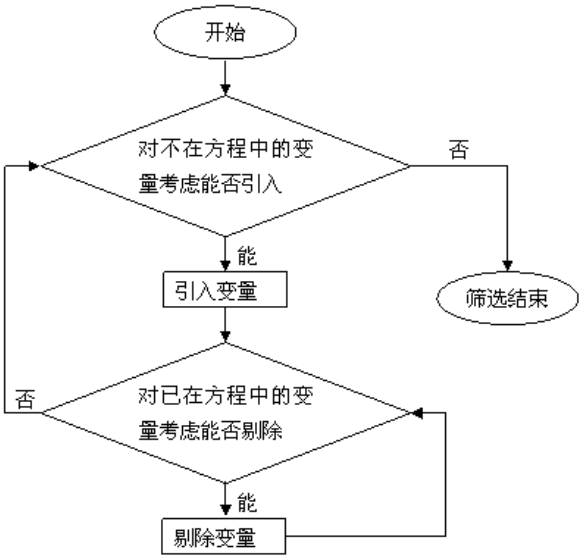
(3) 找出其中最大的p值，如果此P值大于显著性水平 α_{out}，则此变量被剔除。对于某个名义变量的单面化变量，其最小p值大于显著性水平 α_{out}，则此名义变量的其它单面化变量也被删除。不然，表明没有变量可被剔除，选择过程终止。

(4) 回到(2)进行下一轮剔除。

4.3逐步回归(stepwise selection)

(1)基本思想：逐个引入自变量。每次引入对Y影响最显著的自变量，并对方程中的老变量逐个进行检验，把变为不显著的变量逐个从方程中剔除掉，最终得到的方程中既不漏掉对Y影响显著的变量，又不包含对Y影响不显著的变量。

(2)筛选的步骤：首先给出引入变量的显著性水平 α_{in} 和剔除变量的显著性水平 α_{out}，然后按下图筛选变量。



(3)逐步筛选法的基本步骤

逐步筛选变量的过程主要包括两个基本步骤：一是从不在方程中的变量考虑引入新变量的步骤；二是从回归方程中考虑剔除不显著变量的步骤。

假设有p个需要考虑引入回归方程的自变量.

① 设仅有截距项的最大似然估计值为 L_0。对p个自变量每个分别计算Score检验值，

设有最小p值的变量为 x_{e1}，且有 $p_{e1}=\min(p_j)$，对于单面化(dummy)变量，也如此。若 $p_{e1}<\alpha_{in}$，则此变量进入模型，不然停止。如果此变量是名义变量单面化(dummy)的变量，则此名义变量的其它单面化变量也进入模型。其中 α_{in} 为引入变量的显著性水平。

② 为了确定当变量 x_{e1} 在模型中时其它p-1个变量也是否重要，将 $x_j, j=1,2,\cdots p, j\neq e$ 分别与 x_{e1} 进行拟合。对p-1个变量分别计算Score检验值，其p值设为 p_j。设有最小p值的变量为 x_{e2}，且有 $p_{e2}=\min(p_j)$.若 $p_{e2}<\alpha_{in}$，则进入下一步，不然停止。对于单面化变量，其方式如同上步。

③ 此步开始于模型中已含有变量 x_{e1} 与 x_{e2}。注意到有可能在变量 x_{e2} 被引入后，变量 x_{e1} 不再重要。本步包括向后删除。根据(2.1)计算变量 x_{e1} 与 x_{e2} 的Wald检验值，和相应的p值。设 x_{e3} 为具有最大p值的变量，即 p_{e3} =max(p_j), $j=e_1,e_2$.如果此p值大于 α_{out}，则此变量从模型中被删除，不然停止。对于名义变量，如果某个单面化变量的最小p值大于 α_{out}，则此名义变量从模型中被删除。

④ 如此进行下去，每当向前选择一个变量进入后，都进行向后删除的检查。循环终止的条件是：所有的p个变量都进入模型中或者模型中的变量的p值小于 α_{out}，不包含在模型中的变量的p值大于 α_{in}。或者某个变量进入模型后，在下一步又被删除，形成循环。

本文适合有少许文本分类实践经验的同学。

1.什么是文本分类？

简单点说，给定类别，将文本分到某个或某几个类别中。比如，一篇网页，判断它是体育类还是政治类还是娱乐类。当然网页比文本稍微复杂一些，需要先做一些页面解析等预处理工作。文本分类可看作网页分类的一个子问题。

想继续了解文本分类，推荐看计算所王斌老师的PPT ，点击这里。

2.什么是逻辑回归（**LR, logistic regression**）？

英文，参考wikipedia的定义，点击这里。

中文，可参考这篇，点击这里。

目前有不少机器学习方面的开源实现，本人采用了liblinear开源库，实现高效，使用简单，它支持LR和SVM，点击这里了解。

3.什么是模型调优？

对于文本分类问题，收集若干类别样本，确定好文本特征后，采用一些成熟的分类算法（朴素贝叶斯、SVM、决策树、LR等），即可得到一个分类器，采用交叉验证(cross validation)可得到这个分类器的大致效果。要想达到比较理想的分类效果（准确率/召回率），则需要进行模型调优。以下列举本人在利用LR的实践过程中觉得比较重要的调优点。

4. 训练样本调优

理想情况下，对于任何分类算法来讲，只要训练样本足够好（什么算好？），分类效果的差别并不是特别大。训练样本的好坏直接决定了分类效果。矛盾的是，理想中的训练样本几乎无法得到。主要原因有二：1）训练样本无法正确映射出现实世界中的各类别比例。比如现实世界里A类/B类=40，如果按照这个比例来确定训练样本，则显然不行。2）对于有监督学习来说，训练样本往往需要人工标注，这使得训练样本数量无法得到保证。另外人工标注不可避免会产生错误，也会对分类造成影响。

在实践过程当中，要保持对数据的敏感性，对于模型的错误/有偏输出结果，要不断分析和猜测并加以验证。比如某个非政治类词与政治类的关联度特别大，则可断定是训练样本的有偏性造成的（比如训练样本大部分来自新浪政治类网页，则新浪这个词肯定与政治类关联度特别大，要想办法消除这种有偏性）。

5. 特征调优

如何表示一个文本？向量空间模型（VSM）是比较常用的。对于文本分类问题，VSM的每一维可以表示一个word，而tfidf是比较常见的权重计算方法，但是tfidf的具体计算方法又有很多种（log形式, normalized形式、tf=1形式等）。任何一种都没有绝对的优劣性。需要在实践中根据具体数据来选择对应形式。

另外，特征的维数及各维定义也需要商榷。维数过大会带来训练时间过长和数据稀疏性问题。维数过小无法完整表示文本显然也不行。一般通过特征选择（feature selection）方法来确定特征维数和组成方式。实际使用过程中CHI和IG是效果比较好的两种。各维数含义则可简单可复杂，简单的，各维可表示一个word，直观明了；复杂的可使用LSI等方法来对其进行重构。

对于特征选择的计算结果（每维特征与各类别的关联度排序），可稍加分析，看是否存在训练样本的有偏问题。

6. 保持对数据的敏感性

模型调优是一个不断迭代的过程，在实践过程中，要善于根据分类器的输出（打分分布、区间样本抽查、误判分析）来发现问题所在。走一步，看一步。不要盲目地去调整，要根据模型目前的状态，分析其可能的问题所在，然后有针对性地去优化。另外还要确保测试集合的开放性，防止over-fitting.

7. 保持耐心、细致

模型调优又是一个繁琐的工作，需要不断的迭代优化，需要不断的抽查样本，需要不断的分析和对比数据。往往有时模型的输出结果与预测不符，会令人沮丧。但最重要的是要保持耐心和细心。如果确定目前的方法可以解决这类问题，则要坚定不移地走下去，同时细致地发现可能存在的问题并加以改进。相信总会得到一个令人满意的结果。

上一篇 抛硬币 直到连续出现两次字为止
下一篇 google面试

我的同类文章

Machine Learning（42）

- EM Alogrithm
- Adaboost from Baidu
- Adaboost
- Boosting for PRML
- 隐马尔科夫模型HMM自学（2）
- 隐马尔科夫模型HMM自学（1）
- 隐马尔科夫模型HMM自学（3）
- Normalized Cut
- Entropy
- 准确率召回率

更多

主题推荐 color rgb

猜你在找

有趣的算法（数据结构）
数据结构和算法
《C语言/C++学习指南》加密解密篇（安全相关算法）
前端性能测试分析与精要【小强测试出品】
Jmeter性能测试全程实战
逻辑斯蒂回归logistic regression学习笔记
logistic 回归模型
机器学习实战笔记5logistic回归
Logistic regression 逻辑回归 概述
logistic回归简介

查看评论

2楼 xugguo 2015-05-13 07:03发表

mark

1楼 拾壹女 2015-03-26 18:43发表

mark

您还没有登录,请[登录]或[注册]

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

核心技术类目

全部主题 Hadoop AWS 移动游戏 Java Android iOS Swift 智能硬件 Docker OpenStack
VPN Spark ERP IE10 Eclipse CRM JavaScript 数据库 Ubuntu NFC WAP jQuery
BI HTML5 Spring Apache .NET API HTML SDK IIS Fedora XML LBS Unity
Splashtop UML components Windows Mobile Rails QEMU KDE Cassandra CloudStack
FTC coremail OPhone CouchBase 云计算 iOS6 Rackspace Web App SpringSide Maemo
Compuware 大数据 aptech Perl Tornado Ruby Hibernate ThinkPHP HBase Pure Solr
Angular Cloud Foundry Redis Scala Django Bootstrap