CSDN新首页上线啦,邀请你来立即体验!(http://blog.csdn.net/)

CSDN

മ

49

原创

[...]

逻辑回归

什么是逻辑回归?

博客 (http://blog.csdn.net/?ref=toolbar)

学院 (http://edu.csdn.net?ref=toolbar)

下载 (http://db

2014年07月16日 15:42:14

• 如果是连续的,就是多重线性回归;

• 如果是二项分布,就是Logistic回归;

• 如果是Poisson分布,就是Poisson回归;

• 寻找危险因素:寻找某一疾病的危险因素等;

3. 想办法使得]函数最小并求得回归参数 (θ)

个人有多大的可能性是属于某病。

如果是负二项分布,就是负二项回归。

用的就是二分类的Logistic回归。

Regression问题的常规步骤为: 1. 寻找h函数 (即hypothesis); 2. 构造J函数 (损失函数);

Logistic回归的主要用途:

这一家族中的模型形式基本上都差不多,不同的就是因变量不同。

更多 ▼

因为如此,这两种回归可以归于同一个家族,即广义线性模型(generalizedlinear model)。

• 预测:根据模型,预测在不同的自变量情况下,发生某病或某种情况的概率有多大;

Logistic回归与多重线性回归实际上有很多相同之处,最大的区别就在于它们的因变量不同,其他的基本都差不多。正是

Logistic回归的因变量可以是二分类的,也可以是多分类的,但是二分类的更为常用,也更加容易解释。所以实际中最常

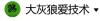
● 判别:实际上跟预测有些类似,也是根据模型,判断某人属于某病或属于某种情况的概率有多大,也就是看一下这

Logistic回归主要在流行病学中应用较多,比较常用的情形是探索某疾病的危险因素,根据危险因素预测某疾病发生的概 率,等等。例如,想探讨胃癌发生的危险因素,可以选择两组人群,一组是胃癌组,一组是非胃癌组,两组人群肯定有不 同的体征和生活方式等。这里的因变量就是是否胃癌,即"是"或"否",自变量就可以包括很多了,例如年龄、性别、饮食

Q

标签:逻辑回归(http://so.csdn.net/so/search/s.do?q=逻辑回归&t=blog)





(http://w/nittep.bblogigbcooknometr/prostreit/activ

ref=toolltar) source=csdnblog

203845



(http://blog.csdn.net/pakko)

码云

未开通 粉丝 原创 查欢 (https://gi 18 121 utm_sourc

他的最新文章

更多文章 (http://blog.csdn.net/pakko)

CentOS6.5 安装apache/mysql/php (http://blog.csdn.net/pakko/article/

Openswan在CentOS6.4上的编译安装

Kerberos for linux配置 (http://blog.c sdn.net/pakko/article/details/48245



逻辑分析仪



思维逻辑训练

移民澳洲的条件

details/53045900)

与配置 (http://blog.csdn.net/pakko/ article/details/50242361)

gmat逻辑网课

构造预测函数h

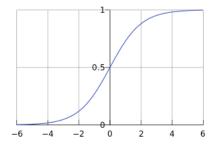
常规步骤

Logistic回归虽然名字里带"回归",但是它实际上是一种分类方法,主要用于两分类问题(即输出只有两种,分别代表两 个类别),所以利用了Logistic函数(或称为Sigmoid函数),函数形式为:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid 函数在有个很漂亮的"S"形,如下图所示(引自维基百科):

习惯、幽门螺杆菌感染等。自变量既可以是连续的,也可以是分类的。



下面左图是一个线性的决策边界,右图是非线性的决策边界。

在线课程



腾讯云容器服务架构实现 介绍 () 讲师: 董晓杰



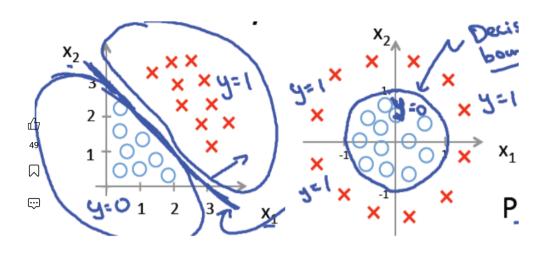
⚠ 内容举报 TOP

(http://edu.csdn.net/hu utm_source=blog9) 妍师owse/series_detail/ 73?utm_source=blog9)

返回顶部

他的热门文章

逻辑回归 (http://blog.csdn.net/pakko/ article/details/37878837)



203570

磁盘IO:缓存I p://blog.csdn s/8779110) **3** 7505



4 6526 线性回归-理论

90452)

pakko/article

Socket中的TIME_WAIT状态 (http://blo g.csdn.net/pakko/article/details/8780 540)

4403

对于线性边界的情况,边界形式如下:

$$\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n = \sum_{i=1}^n \theta_i x_i = \theta^T x$$

构造预测函数为:

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

函数 $h_{ heta}(x)$ 的值有特殊的含义,它表示结果取1的概率,因此对于输入x分类结果为类别1和类别0的概率分别为:

$$P(y=1 | x; \theta) = h_{\theta}(x)$$

$$P(y=0 | x; \theta) = 1 - h_{\theta}(x)$$
(1)

构造损失函数J

Cost函数和J函数如下,它们是基于最大似然估计推导得到的。

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1\\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{n} Cost(h_{\theta}(x_i), y_i) = -\frac{1}{m} \left[\sum_{i=1}^{n} y_i \log h_{\theta}(x_i) + (1 - y_i) \log(1 - h_{\theta}(x_i)) \right]$$

下面详细说明推导的过程:

(1)式综合起来可以写成:

$$P(y | x; \theta) = (h_{\theta}(x))^{y} (1 - h_{\theta}(x))^{1-y}$$

取似然函数为:

$$L(\theta) = \prod_{i=1}^{m} P(y_i \mid x_i; \theta) = \prod_{i=1}^{m} (h_{\theta}(x_i))^{y_i} (1 - h_{\theta}(x_i))^{1 - y_i}$$

 \triangle 内容举报

TOP 返回顶部

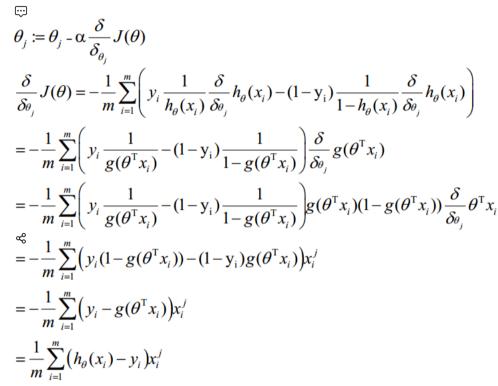
对数似然函数为:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{m} (y_i \log h_{\theta}(x_i) + (1 - y_i) \log(1 - h_{\theta}(x_i)))$$

最大似然估计就是求使l(heta) 取最大值时的heta,其实这里可以使用梯度上升法求解,求得的heta就是要求的最佳参数。但 是 , 在Andrew Ng的课程中将J(heta) 取为下式 , 即

$$J(\theta) = -\frac{1}{m}l(\theta)$$

因为乘了一个负的系数-1/m,所以取J(heta) 最小值时的heta为要求的最佳参数。 \Box





$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_\theta(x_i) - y_i) x_i^j$$

向量化Vectorization

Vectorization是使用矩阵计算来代替for循环,以简化计算过程,提高效率。

如上式, $\Sigma(...)$ 是一个求和的过程,显然需要一个for语句循环m次,所以根本没有完全的实现vectorization。

下面介绍向量化的过程:

约定训练数据的矩阵形式如下,x的每一行为一条训练样本,而每一列为不同的特称取值:





广告





⚠ 内容举报

$$x = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}, \theta = \begin{bmatrix} \theta_0 \\ \dots \\ \theta_n \end{bmatrix}$$

$$\begin{array}{c} \text{Pr} \\ \text{Pr}$$

$$E = h_{\theta}(x) - y = \begin{bmatrix} g(A_1) - y_1 \\ \dots \\ g(A_m) - y_m \end{bmatrix} = \begin{bmatrix} e_1 \\ \dots \\ e_m \end{bmatrix} = g(A) - y$$

g(A)的参数A为一列向量,所以实现g函数时要支持列向量作为参数,并返回列向量。由上式可知h heta(x)-v可由 g(A)-y 一次计算求得。

θ更新过程可以改为:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i) x_i^j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m e_i x_i^j = \theta_j - \alpha \frac{1}{m} x^T E$$

综上所述, Vectorization后θ更新的步骤如下:

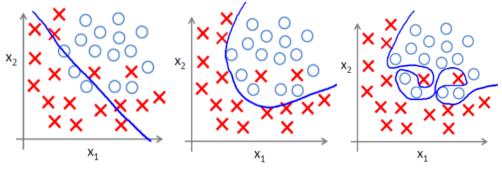
- (1)求 $A = x \bullet \theta$;
- (2)求 E = g(A) v;
- (3) $\Re \theta := \theta \alpha x^T E$.

正则化Regularization

过拟合问题

对于线性回归或逻辑回归的损失函数构成的模型,可能会有些权重很大,有些权重很小,导致过拟合(就是过分拟合了训 练数据),使得模型的复杂度提高,泛化能力较差(对未知数据的预测能力)。

下面左图即为欠拟合,中图为合适的拟合,右图为过拟合。



问题的主因

过拟合问题往往源自过多的特征。

- 1)减少特征数量(减少特征会失去一些信息,即使特征选的很好)
 - 可用人工选择要保留的特征;
 - 模型选择算法;
- 2)正则化(特征较多时比较有效)
 - 保留所有特征,但减少θ的大小

正则化是结构风险最小化策略的实现,是在经验风险上加一个正则化项或惩罚项。正则化项一般是模型复杂度的单调递增 函数,模型越复杂,正则化项就越大。

从房价预测问题开始,这次采用的是多项式回归。左图是适当拟合,右图是过拟合。

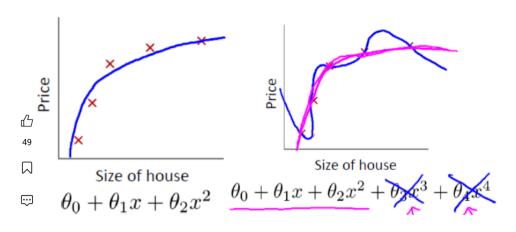








⚠ 内容举报





直观来看,如果我们想解决这个例子中的过拟合问题,最好能将 x^3, x^4 的影响消除,也就是让 $\theta 3 \approx 0, \theta 4 \approx 0$ 。假设我们对 θ_3, θ_4 进行惩罚,并且令其很小,一个简单的办法就是给原有的Cost函数加上两个略大惩罚项,例如:

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{n} (h_{\theta}(x_i) - y_i)^2 + 1000\theta_3^2 + 1000\theta_4^2$$

这样在最小化Cost函数的时候, $\theta 3 \approx 0.04 \approx 0$ 。

正则项可以取不同的形式,在回归问题中取平方损失,就是参数的L2范数,也可以取L1范数。取平方损失时,模型的损失函数变为:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{n} (\mathbf{h}_{\theta}(\mathbf{x}_{i}) - y_{i})^{2} + \lambda \sum_{j=1}^{n} \theta_{j}^{2}$$

lambda是正则项系数:

- 如果它的值很大,说明对模型的复杂度惩罚大,对拟合数据的损失惩罚小,这样它就不会过分拟合数据,在训练数据上的偏差较大,在未知数据上的方差较小,但是可能出现欠拟合的现象;
- 如果它的值很小,说明比较注重对训练数据的拟合,在训练数据上的偏差会小,但是可能会导致过拟合。

正则化后的梯度下降算法0的更新变为:

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (h_{\theta}(x_i) - y_i) x_i^j - \frac{\lambda}{m} \theta_j$$

正则化后的线性回归的Normal Equation的公式为:

$$\theta = \begin{pmatrix} X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \end{pmatrix}^{-1} X^T Y$$

其他优化算法

- Conjugate gradient method(共轭梯度法)
- Quasi-Newton method(拟牛顿法)
- BFGS method
- L-BFGS(Limited-memory BFGS)

后二者由拟牛顿法引申出来,与梯度下降算法相比,这些算法的优点是:

- 第一,不需要手动的选择步长;
- 第二,通常比梯度下降算法快;

但是缺点是更复杂。

⚠
内容举报

企 返回顶部

多类分类问题

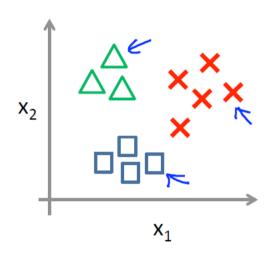
对于多类分类问题,可以将其看做成二类分类问题:保留其中的一类,剩下的作为另一类。 对于每一个类 i 训练一个逻辑回归模型的分类器 $h_{ heta}^{(i)}\left(x
ight)$, 并且预测y = i时的概率 ; 对于一个新的输入变量x, 分别对 每一个类进行预测,取概率最大的那个类作为分类结果

 $\max_{i} h_{\theta}^{(i)}(x)$

മ 49

Multi-class classification:







http://blog.csdn.net/dongtingzhizi/article/details/15962797

(http://blog.csdn.net/dongtingzhizi/article/details/15962797)

Coursera公开课笔记:

斯坦福大学机器学习第六课"逻辑回归(Logistic Regression)"

(http://52opencourse.com/125/coursera%E5%85%AC%E5%BC%80%E8%AF%BE%E7%AC%94%E8%A E%B0-

%E6%96%AF%E5%9D%A6%E7%A6%8F%E5%A4%A7%E5%AD%A6%E6%9C%BA%E5%99%A8%E5% AD%A6%E4%B9%A0%E7%AC%AC%E5%85%AD%E8%AF%BE-

%E9%80%BB%E8%BE%91%E5%9B%9E%E5%BD%92-logistic-regression)

Coursera公开课笔记: 斯坦福大学机器学习第七课 "正则化(Regularization)"

(http://52opencourse.com/133/coursera%E5%85%AC%E5%BC%80%E8%AF%BE%E7%AC%94%E8 %AE%B0-

%E6%96%AF%E5%9D%A6%E7%A6%8F%E5%A4%A7%E5%AD%A6%E6%9C%BA%E5%99%A8%E5 %AD%A6%E4%B9%A0%E7%AC%AC%E4%B8%83%E8%AF%BE-

%E6%AD%A3%E5%88%99%E5%8C%96-regularization)



发表你的评论

(http://my.csdn.net/forever2008250)



linkin_punk (/linkin_punk) 2017-08-04 10:41

23楼

(/link敝党@nks上所述, Vectorization后θ更新的步骤如下"上面那个公式最右边的x^T是不是漏了个下标 x^T_j 回复

qftju (/qftju) 2017-08-01 16:50

22楼

(/qftj请问为什么把i=0时的 $\theta0$ 都给略去了?

回复









 \triangle 内容举报

qq_26229875 (/qq_26229875) 2017-05-12 11:01

21楼

(/qq_**262298作**素原创。原文见如下链接:

http://download.csdn.net/detail/lewsn2008/6547463

且本文由误: "θ更新过程可以改为:"下面的式子中的下标。应该去掉。

മ

查看 26 条热评~

49

相关文章推荐

对线性回归、逻辑回归、各种回归的概念学习 (http://blog.csdn.net/viewcode/article/det...

回归问题的条件/前提:1)收集的数据2)假设的模型,即一个函数,这个函数里含有未知的参数,通过学习,可以估计出 参数。然后利用这个模型去预测/分类新的数据。 1. 线性回归 假设 特征 和 ...

🥥 viewcode (http://blog.csdn.net/viewcode) 2013年04月17日 16:33 👊219443

机器学习系列(1)_逻辑回归初步 (http://blog.csdn.net/yaoqiang2011/article/details/49...

1、总述逻辑回归是应用非常广泛的一个分类机器学习算法,它将数据拟合到一个logit函数(或者叫做logistic函数)中,从而 能够完成对事件发生的概率进行预测。 2、由来 要说逻辑回归,我们...



🙆 yaoqiang2011 (http://blog.csdn.net/yaoqiang2011) 2015年10月14日 13:18 🛚 🕮51627



不止20K, Python薪酬又飙升了??

2017年 Pytyhon薪酬曝光啦!看完后薪资报告后,同事说了一句:人生苦短,不学Python算白活......

(http://www.baidu.com/cb.php?c=IgF_pyfqnHmknjnvPjc0IZ0qnfK9ujYzP1f4PjDs0Aw-

5Hc3rHnYnHb0TAg15HfLPWRznjb0T1Y4P1bLn1F-

YpAq8nWqdIAdxTvqdThP-

5yF_UvTkn0KzujYz0AFV5H00TZcgn0KdpyfqnHRLPjnvnfKEpyfqnHc4rj6kP0KWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6KWThnqn1cvnWm)

机器学习算法与Python实践之(七)逻辑回归(Logistic Regression) (http://blog.csdn....

机器学习算法与Python实践之(七)逻辑回归(Logistic Regression)zouxy09@qq.comhttp://blog.csdn.net/zouxy09 机器学习算法...

🎁 zouxy09 (http://blog.csdn.net/zouxy09) 2014年03月02日 23:49 🕮 187498

机器学习系列(2)_从初等数学视角解读逻辑回归 (http://blog.csdn.net/yaoqiang2011/arti...

作者: 龙心尘 && 寒小阳 时间: 2015年10月。 出处: http://blog.csdn.net/han_xiaoyang/article/details/49123419。 声明:版权所有...

🌘 yaoqiang2011 (http://blog.csdn.net/yaoqiang2011) 2015年10月22日 13:24 🛚 🕮22195

机器学习算法--逻辑回归原理介绍 (http://blog.csdn.net/chibangyuxun/article/details/5...

本博文针对逻辑回归的原理进行了介绍:逻辑回归就是这样的一个过程:面对一个回归或者分类问题,建立代价函数,然后通 过优化方法迭代求解出最优的模型参数,然后测试验证我们这个求解的模型的好坏。Logistic...

💮 chibangyuxun (http://blog.csdn.net/chibangyuxun) 2016年11月13日 11:23 🔲 6745

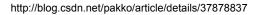
⚠ 内容举报

TOP 返回顶部



AI 工程师职业指南

我们请来商汤、杜邦、声智、希为、58同城、爱因互动、中科视拓、鲁朗软件等公司 AI 技术一线的专 家,请他们从实践的角度来解析 AI 领域各技术岗位的合格工程师都是怎样炼成的。













(http://www.baidu.com/cb.php?c=IgF_pyfqnHmknjfzrj00IZ0qnfK9ujYzP1f4Pjnd0Aw-

5Hc4nj6vPjm0TAq15Hf4rjn1n1b0T1dhujDsP1RkrAfkuWIBPhD30AwY5HDdnHcsnH63PHD0IqF 5y9YIZ0lQzqMpqwBUvqoQhl rA7Wuj0YmhP9PARvujmYmH0vm1qdIAdxTvqdThP-

5HDknWF9mhkEusKzujYz0AFV5H00TZcqn0KdpyfqnHRLPjnvnfKEpyfqnHnsnj0YnsKWpyfqP1cvrHnz0AqLUWYs0ZK45HcsP6K

Logistic Regression(逻辑回归)详细讲解 (http://blog.csdn.net/joshly/article/details/5...

Logistic Regression(逻辑回归)以前在学校学到Logistic Regression的时候,虽然最后会使用,但是对于许多地方有很多的 疑惑,今天在这里详细梳理一下Logistic Re...

wipshly (http://blog.csdn.net/joshly) 2016年01月10日 23:31 27474







(··)

逻辑斯蒂回归模型 (http://blog.csdn.net/hechenghai/article/details/46817031)

第一节中说了, logistic 回归和线性回归的区别是:线性回归是根据样本X各个维度的Xi的线性叠加(线性叠加的权重系数wi 就是模型的参数)来得到预测值的Y,然后最小化所有的样本预测值Y与真实值y'的...

🧑 hechenghai (http://blog.csdn.net/hechenghai) 2015年07月09日 14:57 🖫 14653

【机器学习】线性回归和逻辑回归的理解 (http://blog.csdn.net/chenoo0539/article/det...

最近学习斯坦福大学Andrew Ng教授的machine learning课程,才学到第4课,感觉对线性回归 (linear regression) 和逻 辑回归 (logistic regression) 总...

逻辑回归模型介绍和程序实现 (http://blog.csdn.net/weixin 38208741/article/details/7...

逻辑回归原理及实现 虽然叫做"回归",但是这个算法是用来解决分类问题的。回归与分类的区别在于:回归所预测的 目标量的取值是连续的(例如房屋的价格);而分类所预测的目标变量的取值是离散的(例如判断...



🥠 weixin_38208741 (http://blog.csdn.net/weixin_38208741) 2017年04月23日 10:50 🖫1327

逻辑回归、决策树和支持向量机的直观理解 (http://blog.csdn.net/yuxiangyunei/article/d...

分类问题是我们在各个行业的商业业务中遇到的主要问题之一。在本文中,我们将从众多技术中挑选出三种主要技术展开讨 论,逻辑回归(Logistic Regression)、决策树(Decision T...

🦱 yuxiangyunei (http://blog.csdn.net/yuxiangyunei) 2015年12月03日 13:45 🛚 🕮 3248

Stanford公开课机器学习---week3-1.Logistic Regression 逻辑回归 (http://blog.csdn.ne...

1分类 Classification在分类问题中,我们尝试预测的是结果是否属于某一个类(例如正确或错误)。 分类问题的例子有: 判断一 封电子邮件是否是垃圾邮件; 判断一次金融交易是否是欺诈等等。 ...

🍢 muzilanlan (http://blog.csdn.net/muzilanlan) 2015年09月30日 20:00 🛛 🖺 1431

Stanford机器学习---第三讲. 逻辑回归和过拟合问题的解决 logistic Regression & Regulari...

本栏目(Machine learning)包括单参数的线性回归、多参数的线性回归、Octave Tutorial、Logistic Regression、Regul arization、神经网络、机器学...



🦱 abcjennifer (http://blog.csdn.net/abcjennifer) 2012年07月07日 20:45 🛮 🕮 167921

内容举报

⚠

TOP 返回顶部

机器学习经典算法详解及Python实现---Logistic回归(LR)分类器(http://blog.csdn.net/...

Logistic回归可以也可以用于多分类的,但是二分类的更为常用也更容易解释。所以实际中最常用的就是二分类的Logistic回 归。LR分类器适用数据类型:数值型和标称型数据。其优点是计算代价不高,易于...



🥋 suipingsp (http://blog.csdn.net/suipingsp) 2014年12月09日 12:56 CC 22773



Delphi7高级应用开发随书源码 (http://download.csdn.net/detail/chen...

/http://downloa

2003年04月30日 00:00 676KB

逻辑回归模型的两种定义与参数估计思路 (http://blog.csdn.net/DawnRanger/article/det...

逻辑斯谛回归 (logistic regression) 是统计学习中的经典分类方法,属于判别模型。1. 逻辑斯谛回归模型定义在 Andrew N G 的 Machine Learning 课程和李航的统计...

Logistic Regression (逻辑回归)原理及公式推导(http://blog.csdn.net/programmer_w...

Logistic Regression(逻辑回归)是机器学习中一个非常非常常见的模型,在实习生环境中也常常被使用,是一种经典的分 类模型(不是回归模型)。本文主要介绍了Logistic Regressi...

逻辑回归Logistic Regression 模型简介 (http://blog.csdn.net/GarfieldEr007/article/de...

逻辑回归 (Logistic Regression)是机器学习中的一种分类模型,由于算法的简单和高效,在实际中应用非常广泛。本文作 为美团机器学习InAction系列中的一篇,主要关注逻辑回归算法的数学模...

🚇 GarfieldEr007 (http://blog.csdn.net/GarfieldEr007) 2016年04月15日 12:22 皿2965

Logistic Regression (逻辑回归)简介 (http://blog.csdn.net/fengjiancangyue/article/...

逻辑回归模型 (Logistic Regression) 是一个重要的有监督二类分类模型,在计算广告学中有着非常广泛的应用。在求职过 程中发现各大互联网公司凡是涉及广告工程师的岗位面试几乎都会问到逻辑回归,...

🌑 fengjiancangyue (http://blog.csdn.net/fengjiancangyue) 2014年12月16日 15:32 叫1377

逻辑回归模型及LBFGS的Sherman Morrison(SM) 公式推导 (http://blog.csdn.net/zhiro...

模型简介 二项逻辑回归模型: (1.1) (1.2) 其中,Y为标签,X为样本,W为模型参数,exp为指数函数。 根据 (1.1) 和 (1.2)可以得到P(Y=1|X)和P(Y=...

🥥 zhirom (http://blog.csdn.net/zhirom) 2014年08月01日 10:20

29076



Delphi7高级应用开发随书源码 (http://download.csdn.net/detail/chen...

(http://downloa

2003年04月30日 00:00 676KB

⚠ 内容举报