法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

逻辑回归模型在申请评分卡中的应用

目录

逻辑回归的概述

伯努利分布

- 从概率的角度来看，“逾期”是一个随机事件。如何刻画它的随机性？
- 伯努利分布：一种离散分布，用于表示0-1型事件发生的概率

例：$P(\text{逾期}) = \text{p}$，$P(\text{不逾期}) = 1 - \text{p}$

合并起来，可以是

$$P(\text{Y} = \text{y}) = p^{y}(1-p)^{1-y}$$

$$y = \begin{cases} 1, & \text{逾期} \\ 0, & \text{不逾期} \end{cases}$$

逻辑回归的概述

伯努利分布(续)

似然函数和对数似然函数

一组申请者在表现期的逾期状态为$\{y_1, y_2, \ldots, y_n\}$, $y_i \in \{0,1\}$,似然函数和对数似然函数是

$$L(p) = \prod P(Y = y_i) = \prod p^{y_i}(1-p)^{1-y_i}$$

$$l(p) = \log\big(L(p)\big) = \log\left\{\prod P(Y = y_i) = \prod p^{y_i}(1-p)^{1-y_i}\right\}$$

$$= \sum y_i \log(p) + (1-y_i)\log(1-p) \qquad (1)$$

参数估计

$$\hat{p} = argmax\ l(p)$$

对(1)求关于p的一阶导数并等于0，有

$$\hat{p} = \frac{\sum y_i}{n}$$

逻辑回归的概述

逻辑回归的基本概念

- 不同的申请人，逾期概率不同。即：

$$p = f(x_1, x_2, \ldots, x_k)$$

其中$\{x_1, x_2, \ldots, x_k\}$是申请人的个人资质

- p的特点
  - 有界
  - 不可直接观测
- 如何定义f()?可以用线性回归吗？

逻辑回归的概述

逻辑回归的基本概念(续)

线性回归

$$p = \beta_0 + \sum \beta_i x_i + \epsilon,\ \ \epsilon \sim N(0, \sigma^2)$$

等价于

$$E(p) = \beta_0 + \sum \beta_i x_i\ ,\ \ \mathrm{var(p)}=\sigma^2$$

优点

- 形式简单

缺点

- p无界
- 不利于通过对数似然函数求解参数

逻辑回归的概述

逻辑回归的基本概念(续)

逻辑回归

$$E(p) = f(\beta_0 + \sum \beta_i x_i)$$

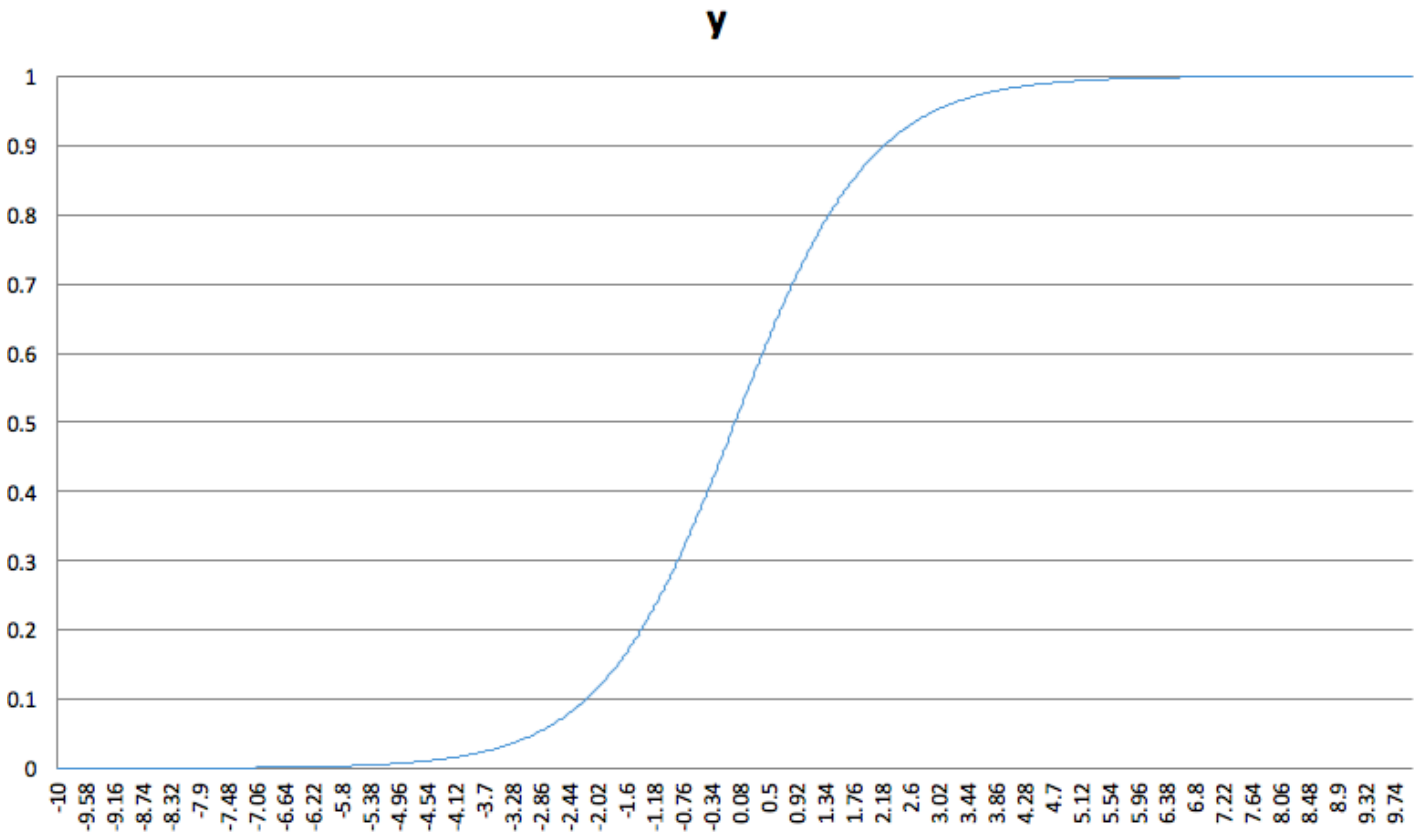
$$f(x) = \frac{\exp(x)}{1 + \exp(x)} \qquad (2)$$

逻辑回归函数的特点

- x取值于$(-\infty, +\infty)$, p取值于$(0,1)$, $f(x)$处处可导
- $\lim_{x \to +\infty} f(x) = 1$, $\lim_{x \to -\infty} f(x) = 0$

逻辑回归的概述

- 逻辑回归的函数图像



逻辑回归的概述

逻辑回归的参数估计

通过伯努利分布的对数似然函数和逻辑回归函数，我们有：

$$l(p) = \sum y_i \log(p_i) + (1 - y_i) \log(1 - p_i) = \sum y_i \log\left(\frac{\exp(\beta_0 + \sum \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum \beta_j x_{ij})}\right) + (1$$

$$- y_i) \log\left(\frac{1}{1 + \exp(\beta_0 + \sum \beta_j x_{ij})}\right)$$

$$= \sum \{y_i \left(\beta_0 + \sum \beta_j x_{ij} - \log\left(1 + \exp\left(\beta_0 + \sum \beta_j x_{ij}\right)\right)\right) - (1$$

$$- y_i) \log(1 + \exp\left(\beta_0 + \sum \beta_j x_{ij}\right))\}$$

$$= \sum \{y_i \left(\beta_0 + \sum \beta_j x_{ij}\right) - \log(1 + \exp\left(\beta_0 + \sum \beta_j x_{ij}\right))\}$$

x_{ij}:第i组样本的第j个特征的值

逻辑回归的概述

逻辑回归的参数估计(续)

考虑l对x_q的系数的偏导数

$$\frac{\partial l}{\partial \beta_q} = \sum \{y_i - \frac{1}{\exp(-\beta_0 - \sum \beta_j x_{ij})}\} x_{iq}$$

$\frac{\partial l}{\partial \beta_q} = 0$没有显式解！

通过梯度上升法求出β_q的估计：

$$\beta_q^{r+1} = \beta_q^r + h \times \delta, \delta = \frac{\partial l}{\partial \beta_q}|_{\beta_q = \beta_q^r}$$

随机梯度上升法

批量梯度上升法

目录

逻辑回归中的变量挑选

逻辑回归中的变量挑选

变量挑选的作用和目的

- 剔除掉跟目标变量不太相关的特征
- 消除多重共线性的影响
- 增加解释性

变量挑选和降维

- 变量挑选是降维的一种手段，反之不是
- 主成分分析法：降维，但没有剔除变量

变量挑选的常用手段

- LASSO
- 逐步回归法
- 随机森林法

逻辑回归中的变量挑选

LASSO

定义

Least absolute shrinkage and selection operator，对回归模型特征的压缩估计

原理

在损失函数中增加模型参数的L1正则约束

$$\min -l(\beta), \text{ s.t. } |\beta| < t$$

$|\beta|$: β的一阶范数，$\sum |\beta_j|$

t: 压缩因子

上述带约束的优化函数等价于

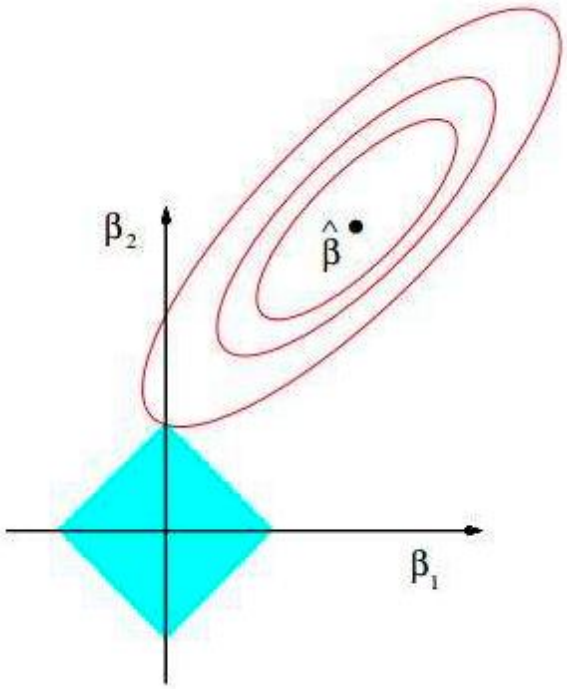
$$\min -l(\beta) + \lambda|\beta|$$

逻辑回归中的变量挑选

LASSO(续)

LASSO的一个几何解释

(二维情形下的){β}的等高线有较大的概率和约束函数的顶点(位于坐标轴上)相交



逻辑回归中的变量挑选

LASSO(续)

- 通过控制λ的值来控制选进模型的特征的个数
  - $\lambda \to 0$: 没有正则化约束，不会剔除特征
  - $\lambda \to +\infty$：所有特征都不会挑选进模型
  - 可以用交叉验证法选择最合适的λ

Group LASSO

- 可以指定一组变量同时被选进或者选出
- 适用于dummy encoding和one hot encoding

逻辑回归中的变量挑选

□ 逐步回归法

向前挑选

I. 初始化时模型里没有特征

II. 每次挑选“最好”的变量放到模型里，评估模型性能的改善

III. 重复(II)直到模型性能不能进一步提升

向后挑选

I. 初始化时把所有特征放到模型里

II. 每次剔除“最差”的变量，该变量的剔除使得模型效果的变化最不显著；评估模型性能的改善

III. 重复(II)直到没有变量被剔除后，模型效果的变化不显著

双向挑选

向前向后法的结合

逻辑回归中的变量挑选

□ 随机森林法

随机森林(Random forest，RF)

一种集成机器学习方法,利用bootstrap和节点随机分裂技术构建多棵决策树,通过投票得到最终分类结果。RF的变量重要性度量可以作为高维数据的特征选择工具

生成随机森林的步骤

I. 从原始训练数据集中,应用bootstrap方法有放回地随机抽取 K个新的自助样本集,并由此构建 K棵分类回归树,每次未被抽到的样本组成了 K个袋外数据(Out-of-bag,OOB)。

II. 设有 n个特征,则在每一棵树的每个节点处随机抽取 m_{try} 个特征($m_{try} \leq n$),通过计算每个特征蕴含的信息量,在 m_{try}个特征中选择一个最具有分类能力的特征进行节点分裂。

III. 每棵树最大限度地生长,不做任何剪裁。

IV. 将生成的多棵树组成随机森林,用随机森林对新的数据进行分类,分类结果按树分类器的投票多少而定。

逻辑回归中的变量挑选

随机森林法(续)

变量重要性

OOB数据特征发生轻微扰动后的分类正确率与扰动前分类正确率的平均减少量

计算

对于每棵决策树，利用袋外数据进行预测，将袋外数据的预测误差将记录下来。其每棵树的误差是$\{error_i\}$

随机重排每个特征，从而形成新的袋外数据，再利用袋外数据进行验证，其每个变量的误差是$\{error_i'\}$

对于某特征来说，计算其重要性是变换后的预测误差与原来相比的差的均值$\{error_i' - error_i\}$。

特征挑选

将特征按照重要性从高到低排列，选取前N个特征

逻辑回归中的变量挑选

- 随机森林法(续)

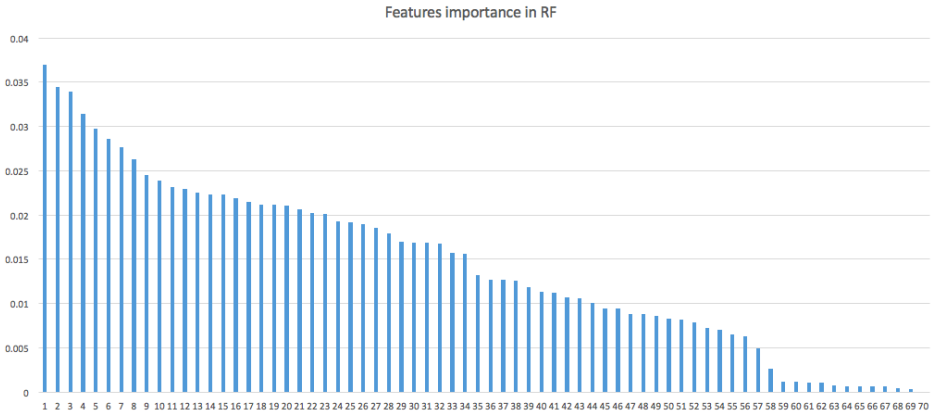


图1：所有变量的重要性

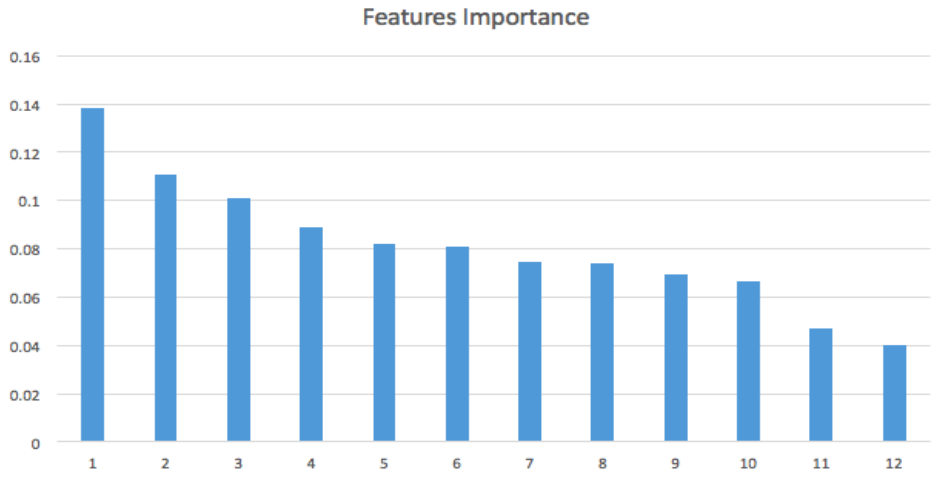


图2：单、多因子变量分析后所有变量的重要性

目录

逻辑回归中的权重问题

逻辑回归中的权重问题

两类错误

- 第一类错误：将逾期人群预测成非逾期
- 第二类错误：将非逾期人群预测成逾期
- 两种误判的代价不一样！

增加逾期类样本的权重

逻辑回归中的权重问题

逻辑回归中的权重问题(续)

设$\{y_i\}$对应的权重向量是$\{w_i\}$，则带权重的对数似然函数是

$$weighted\ l(p) = \sum\{w_i y_i \left(\beta_0 + \sum \beta_j x_{ij}\right) - \log(1 + \exp\left(\beta_0 + \sum \beta_j x_{ij}\right))\}$$

用梯度上升法求出带权重的参数估计

评分卡模型中

- 逾期样本的权重总是高于非逾期样本的权重
- 可以用交叉验证法选择合适的权重
- 也可以跟业务相结合：权重通常跟利率有关。利率高，逾期样本的权重相对低

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：大数据分析挖掘
- 新浪微博：ChinaHadoop