# 西安电子科技大学 2019 年数学建模校内赛

# 承 诺 与 产 权 转 让 书

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们同意将参赛论文以及支撑材料中的所建模型、算法以及程序产权归属西安电子科技大学以及合作单位共有。特别的，B 题参赛论文以及支撑材料中的相应产权西安电子科技大学拥有 50%，合作单位享有 50%。2019 年数学建模校内赛竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号是（从 A/B 中选择一项填写）： A

参赛报名队号为 19B399

报名时所属学院（请填写完整的全名）： 网络与信息安全学院

参赛队员姓名与学号 (打印，用二号字，并签名) ：

1. 龚逸儒 17180288017

2. 冯雪阳 17180288003

3. 彭佐佳 17180288008

日期： 2019 年 5 月 3 日

西安电子科技大学 **2019** 年大学生数学建模校内赛

# 评 阅 专 用 页

| | 评 阅人 1 | 评阅人 2 | 评阅人 3 | 总评 |
|---|---|---|---|---|
| 成绩 | | | | |

# 近视的预测与预警

## 摘要

近年来，中国青少年近视问题不容乐观，进行近视的预测与预警迫在眉睫。本文讨论的是 21 世纪近视的预测与预警机制。以部分青少年数据作为参考资料，通过建立合理的回归模型来预测近视的概率和程度，最后结合智能护眼装置“云夹”探讨了物联网时代的近视预测与预警机制。

在给出眼睛视力与影响因素的量化模型时，我们选取屈光度作为衡量眼睛视力好坏的变量，从生理结构、用眼习惯和遗传基因三大方面选取多个影响因素并通过相关性分析对影响因素进行简化。然后使用主成分回归分析法建立屈光度与多个影响因素的回归模型，得到如下结论：①生理因素对视力的影响最大②户外运动能改善视力③近距离用眼会损害视力。

在研究视力演化机理模型的过程中，我们利用 logistic 回归分析将影响因素进一步简化，构建了近视的 logistic 回归模型。该模型能通过关键影响因素的数据，较为准确地预测出研究对象患有近视的可能性。在此基础上建立预警机制模型，对近视风险做了三个等级的预警，有针对性地对不同年龄的不同群体提供专业的建议。

最后，我们建立 logistic 机器学习模型，结合使用最大似然估计法和梯度上升法，通过不断训练数据集使近视的预测结果更加精确。结合随机生成的解释因素值，计算出结果，总结出最终的可供学习的信息系统（数据表）。

**关键词：主成分回归分析　Logistic 回归　最大似然估计法 梯度上升法 机器学习**

# 正文

### 问题重述

近年来，我国青少年的近视问题日益严重且低龄趋势明显，已成为重大社会公共卫生问题。近视会影响个人健康，限制个别对视力有要求的职业的选择。征兵时越来越多的人被视力这个门槛卡住，这甚至会危及到国家安全。近视的危害不容小觑，控制近视问题严重化迫在眉睫。

高中生物告诉我们：性状是由基因和环境共同决定的。近视作为一种性状表现，也有基因和环境的因素。眼球的生理构造、遗传因素、近距离用眼负荷、户外运动时间缺乏、不良饮食等都会导致近视。

请建立数学模型解决以下问题：

1.给出影响视力的关键因素，量化视力与关键因素的关系。搜集数据时扩宽思路，从多途径获取数据。

2.在问题一求解出视力与关键因素的量化模型的基础上，求解出视力的演化机理模型。

3.问题三要求在问题二的基础上，请给出眼睛视力的预警机制模型。预警机制模型要求如下：①这里的预警机制是开放的，总体原则是在什么时候，做出什么样的预警②要给出理由。③预警对象多样化④设计一种 AI 的预警机制⑤目标读者是父母、老师而不是学者。

4. 在上述基础上设计可供学习的信息系统（数据表），通过查阅资料和数据以及仿真等手段进行机器学习模型。

5.给出实现演化机理模型和机器学习模型的思路，例如收集哪些数据，请写一份方案。

# 问题分析

## 2.1 问题一的分析

我们想通过回归分析求解出视力与关键因素之间的函数关系作为本题的量化模型。我们从眼球结构生理因素、遗传因素和环境因素等方面选取了 14 个影响因素作为自变量。在选取衡量视力的变量作为因变量时，综合考虑，我们选取屈光度作为衡量视力的指标。屈光度是量度光从外界环境进入人眼屈光能力的单位，近视的原因是屈光度太大。近视的屈光度为负，近视度数（$MD$）是屈光度（$D$）的数值乘以 100。下面给出近视度数和屈光度的对应公式：

$$MD = -D \times 100$$

影响视力的因素很多，影响程度各不相同。求解出的回归模型中的回归系数的数值反映了影响程度的大小：回归系数数值越大，影响因素越是关键。当回归系数<0 时，屈光度与影响因素成负相关，视力与影响因素成正相关；反之同理。

### 2.2 问题二的分析

我们将眼睛视力的演化机理模型理解为求解眼睛视力随着环境因素（近距离用眼时间、运动时间等）变化而变化的动态过程。

### 2.3 问题三的分析

在本文使用的预警机制中，主要使用问题二中获得的模型和被测试者提供的数据进行近视可能性的分析，并且以 40%和 60%为界限，提供不同的建议方案。同时根据文献显示的儿童、青少年和成年人生理特征的不同，我们提供了不同的预警上限，比如运动时长下限分别为 2 小时和 1.5 小时，工作活动上限分别为 5 小时和 8 小时。

### 2.4 问题四的分析

我们采用最大似然估计和梯度上升算法来实现 AI 的 logistic 回归，首先使用数据对模型进行训练，梯度阈值为 0.0001。然后在每项数据的最大值与最小值之间随机取值，形成一张预表。再通过训练所得模型计算预表中的每行结果，定义为 MYOPIC，最终完成训练表。

### 2.5 问题五的分析

在实现上文演化机理模型和机器学习模型这两个模型时，我们的具体思路从搜集数据和学习眼科学相关知识两点切入。

## 问题假设

（1）假设找到的数据集真实可靠且数据间相互独立。

（2）假设屈光度可以很好地衡量视力好坏。

# 符号说明

| 符号 | 说明 |
|---|---|
| $x_i$ | 影响视力的第 i 个因素 |
| $D$ | 屈光度 |
| $\beta_i$ | 影响视力的第 i 个因素的回归系数 |
| $\emptyset(x)$ | sigmoid 函数 |
| $T$ | 青少年缓解视觉疲劳的周期 |
| $P$ | 近视的风险 |
| $\alpha$ | 梯度上升的阈值 |

注：其他符号将在文中说明。

# 模型的建立与求解

## 5.1 眼睛视力与关键因素的量化模型

### 5.1.1 搜集数据，给出影响因素

结合信息化时代的优势，我们准备通过以下途径搜集相关数据集：

①从智能穿戴装置公司找用户用眼习惯（包括近距离用眼和户外活动）与近视情况的数据集。智能穿戴装置如云夹是一款通过监测用户用眼习惯并进行护眼的智能护眼夹扣。云夹专业版监测用户的阅读时长、阅读距离、阅读环境光强、阅读角度和户外活动时间。

②从眼科医院找包含患者屈光度、身高、体重、年龄、性别等信息的电子病历作为数据集。

③做问卷调查，问卷调查应包括受访者性别、居住地、用眼习惯、护眼态度和方法等信息。

我们列出了以下 14 个影响视力的因素，以思维导图的形式给出这 14 个指标的符号和含义：


图 1-影响视力的因素

### 5.1.2 影响因素的简化

我们选取屈光度作为衡量视力的指标。

如图 1，我们给出了 14 个影响视力的因素。这 14 个变量间有相关性很高的变量，这 14 个变量的相关性矩阵$(x_{ij})_{14\times14}$如下图 2：①单元格是黑色时，$x_i$,$x_j$ 的相关性为零即变量间相互独立；②单元格颜色为橙色系时，$x_i$,$x_j$ 为正相关，且颜色越浅，正相关程度越大；③单元格颜色为蓝色系时，$x_i$,$x_j$ 为负相关，且颜色越浅，负相关程度越大。

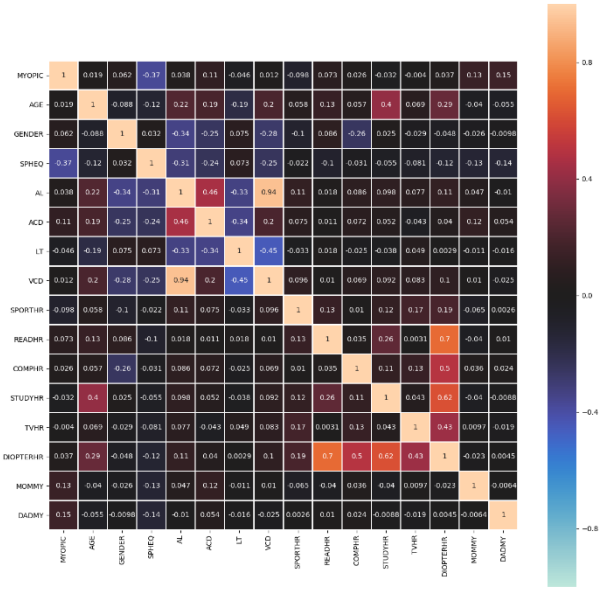


图 2-变量间的相关关系

分析图 2，我们对变量作出如下化简：①4 个生理因素的变量之间有相关性，仅保留 ACD（前房深度）②DIOPTERHR（每周近距离工作总和）与 READHR（每周读书

时长）、COMPHR（每周在电脑上工作时长）、STUDYHR（每周学习时长）和 TVHR（每周看电视时长）这四项有很强的相关性，作出以下调整来方便模型的求解：i 去掉 DIOPTERHR ii 将 READHR 和 STUDYHR 这两个变量整合成一个新的变量 WORKH（每周工作时长）iii 将 COMPHR 和 TVHR 这两个变量整合成一个新的变量 ELECTH（每周使用电子设备时长）。

其他变量保持不变，14 个因素化简为下列 8 个因素：

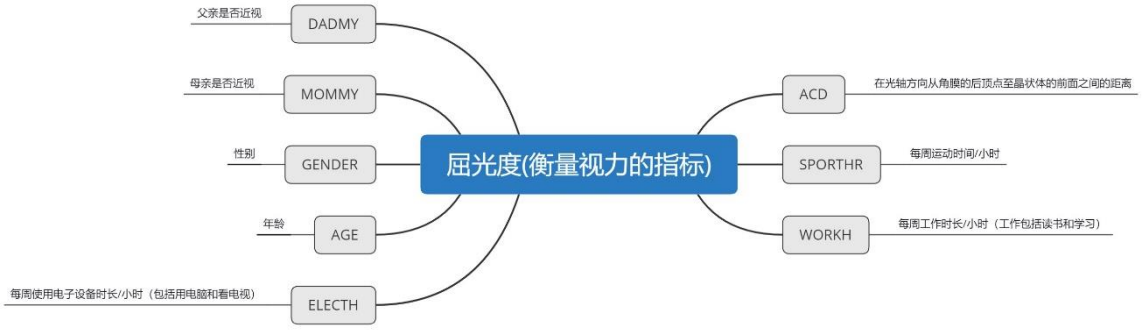


图 3-简化后的影响因素

### 5.1.3 多元线性回归模型的建立与求解

记自变量 ACD,SPORTHR,WORKH,ELECTRH,AGE,GENDER,MOMMY,DADMY 分别为 $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8$；因变量屈光度为 $D$。

假设屈光度与 8 个影响因素之间存在线性关系，建立屈光度($D$)与 8 个影响因素($x_1,x_2,\ldots x_8$)的多元线性回归模型如下：

$$\begin{cases} D=\beta_0+\beta_1x_1+\cdots+\beta_8x_8+\xi \\ \xi\sim N(0,\sigma^2) \end{cases}$$

现有 618 个独立观测数据$[b_i, a_{i1}, \cdots a_{i8}]$，其中 $b_i$ 为 $D$ 的观察值，$a_{i1}, \cdots a_{i8}$分别是 $x_1,x_2,\ldots x_8$ 的观察值，$i$=1,2,…618。

$$\begin{cases} b_i=\beta_0+\beta_1a_{i1}+\cdots+\beta_{14}a_{i8}+\xi_i \\ \xi_i\sim N(0,\sigma^2), i=1,\cdots,618 \end{cases}$$

记

$$X=\begin{bmatrix} 1 & a_{11}\cdots & a_{18} \\ \vdots & \vdots & \vdots \\ 1 & a_{6181}\cdots & a_{6188} \end{bmatrix},\ Y=\begin{bmatrix} b_1 \\ \vdots \\ b_{618} \end{bmatrix},$$

$$\boldsymbol{\xi}=[\xi_1,\cdots,\xi_{618}]^{\mathrm{T}},\ \boldsymbol{\beta}=[\beta_0,\beta_1,\cdots,\beta_8]$$

$$\begin{cases} \boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\xi} \\ \boldsymbol{\xi}\sim N(0,\sigma^2E_8) \end{cases},$$

式中：$E_8$为 8 阶单位矩阵。

用最小二乘法估计参数$\beta_0,\beta_1,\cdots,\beta_8$。求解得到的回归方程如下：

$$\mathrm{D} = 3.8015 - 0.5393x_1 + 0.0021x_2 — 0.0028x_3 - 0.0015x_4 - 0.1451x_5 - 0.0821x_6 + 0.1615x_7 - 0.2814x_8$$

通过分析多元线性回归方程的回归系数的正负和数值可得出以下结论：①在环境因素、遗传因素和生理结构三类影响因素中，生理结构对视力的影响最大②运动可以降低屈光度的数值，即运动可以改善视力③使用电子设备会增加屈光度的数值即近距离用眼会损坏视力。

多元线性回归模型可以较好地量化眼睛视力与影响因素的关系，但从图 2 中可以看出影响因素间存在相关性。为克服最小二乘法估计在数据集存在多重共线性时表现出的不稳定性，我们引入主成分回归分析模型。

### 5.1.4 主成分回归分析模型的建立与求解

为克服最小二乘法估计在数据集存在多重共线性时表现出的不稳定性，我们采用主成分回归分析，将 8 个自变量变换到另一组变量，即主成分，选择其中一部分重要的主成分作为新的自变量，然后用最小二乘法对选取主成分后的模型参数进行估计，最后再变换为原来的模型求出参数的估计。

求解主成分的步骤如下：

（1）对原始数据进行标准化处理。将各指标 $a_{ij}$ 转换成标准化指标 $\bar{a}_{ij}$

$$\widetilde{a_{ij}} = \frac{a_{ij} - \mu_j}{s_j}, i = 1,2,\ldots 618, j = 1,2,\ldots 8$$

式中$\mu_j$，$s_j$是第 $j$ 个指标的样本均值和样本标准差。对应地,称

$$\tilde{x_j} = \frac{x_j - \mu_j}{s_j}, j = 1,2,\ldots,618$$

为标准化指标变量。

（2）计算相关系数矩阵 $R$。相关系数矩阵 $R=(r_{ij})_{8\times8}$,有

$$r_{ij} = \frac{\sum_{k=1}^{618} \widetilde{a_{ki}} \cdot \widetilde{a_{kj}}}{618 - 1}, i, j = 1,2,\cdots,8$$

式中：$r_{ii}=1$；$r_{ij}=r_{ji}$，$r_{ij}$ 为第 $i$ 个指标与第 $j$ 个指标的相关系数。

（3）计算特征值和特征向量。计算相关系数矩阵 $R$ 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_8 \geqslant 0$，即对应的标准化特征向量 $u_1,u_2,\ldots,u_8$,其中 $\boldsymbol{u_j}=[u_{1j},u_{2j},\ldots,u_{8j}]^{\mathrm{T}}$,由特征向量组成 8 个新的指标变量

$$y_1 = u_{11}\widetilde{x_1} + u_{12}\widetilde{x_2} + \cdots + u_{18}\widetilde{x_8}$$

$$y_2 = u_{21}\widetilde{x_1} + u_{22}\widetilde{x_2} + \cdots + u_{28}\widetilde{x_8}$$

$$\vdots$$

$$y_8 = u_{81}\widetilde{x_1} + u_{82}\widetilde{x_2} + \cdots + u_{88}\widetilde{x_8}$$

式中：$y_1$为第 1 主成分；$y_2$为第 2 主成分；…；$y_8$为第 8 主成分。

（4）计算特征值 $\lambda_j(j=1,2,\ldots,8)$的信息贡献率和累计贡献率。称

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{8}\lambda_k}, j = 1,2,\cdots,8$$

为主成分 $y_j$的信息贡献率；而且称

$$\alpha_{\mathrm{p}} = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{8}\lambda_k}$$

为主成分$y_1$，$y_2$，…，$y_p$作为 $p$ 个主成分的累计贡献率。

选取前七个主成分时，累计贡献度达到 94.38％。

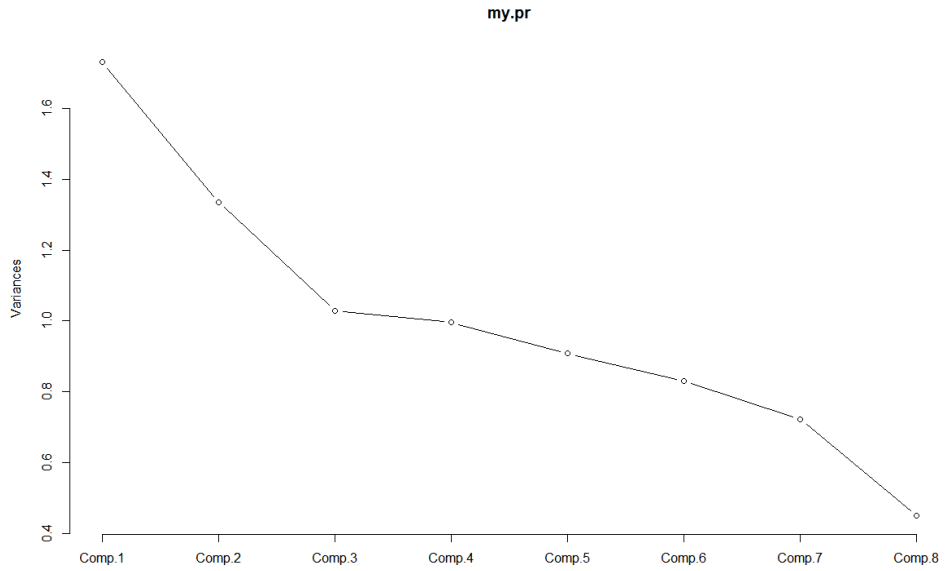


图 4-主成分碎石图

（5）依据贡献率计算出结果：

$$\widetilde{D} = 0.216y_1 + 0.167y_2 + \cdots + 0.090y_7$$

（6）逆变换求解出屈光率与 8 个影响因素的关系。

$$\widetilde{D} = \frac{D - E(D)}{stdev(D)}$$

$$\widetilde{x_j} = \frac{x_j - \mu_j}{s_j}$$

上式中：E()均值；stdev()标准差

$$D = 3.6856 - 0.5303x_1 + 0.0023x_2 - 0.0035x_3 + 0.0002x_4 - 0.1372x_5 - 0.0731x_6 - 0.1673x_7 - 0.2771x_8$$

通过分析主成分回归分析方程的回归系数的正负和数值可得出以下结论：①在环境因素、遗传因素和生理结构三类影响因素中，生理结构对视力的影响最大②运动可以降低屈光度的数值，即运动可以改善视力③使用电子设备会增加屈光度的数值即近距离用眼会损坏视力。

## 5.2 眼睛视力的演化机理模型(myopia logistic 模型)

### 5.2.1 myopia logistic 模型的建立

为解决青少年是否近视的二分类问题，我们想到 sigmoid 函数和 logistic 模型的相关特性。将 sigmoid 函数记作∅(x)：

$$\emptyset(\mathrm{x}) = \frac{1}{1 + e^{-x}}$$

ø(x)图像如下：

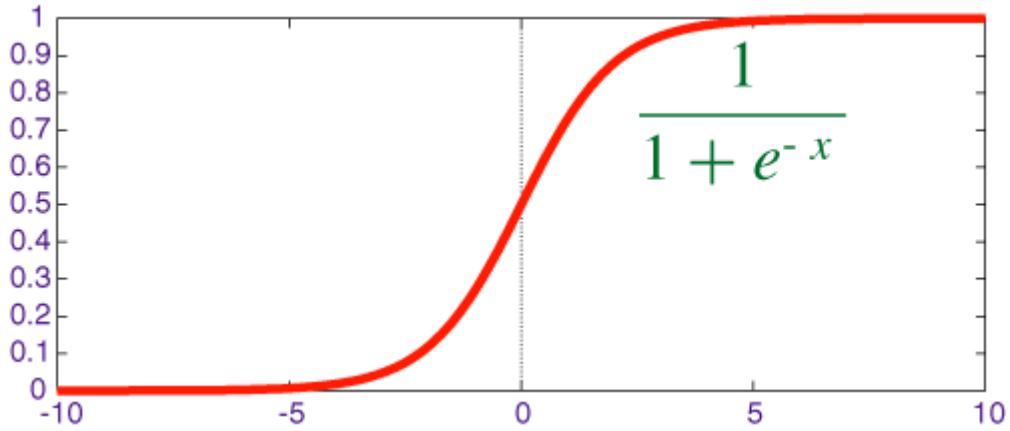


图 5-sigmoid 函数

从上图可以看到ø(x)是一个 s 形的曲线，它的取值在[0, 1]之间，在远离 0 的地方函数的值会很快接近 0 或者 1。这个特性对可以帮我们解决青少年是否近视的二分类问题。

逻辑回归的假设函数形式如下：

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

逻辑回归模型所做的假设是：

$$P(y=1|x;\theta)=\phi(\theta^T x)=\frac{1}{1+e^{\theta^T x}}$$

这表示在给定 $x$ 和 $\theta$ 的条件下 $y=1$ 的概率，即在给定样本数据和系数的条件下，研究对象近视的概率。

为了确定 $x$ 和 $\theta$，我们对最原始的 14 个因素分析。通过做 logistic 回归的显著性检验对 14 个具有内部相关性的影响因素做出删减，缩减并优化出最终的模型：

$$P(y=1|x;\theta)=\frac{1}{1+e^{\theta^T x}}$$

其中

$\theta=[-8.066,1.511,-0.075,0.086,0.776,0.946,0.050]^{\mathrm{T}}$

$x=[1,x_1,x_2,x_3,x_4,x_5,x_6]^T$，其中 $x_1,x_2,...,x_6$ 分别代表 ACD,SPORTHR, WORKHR,MUMMY,DADMY,SPORTHR:GENDER 这六类影响因素的数值。

### 5.2.2 myopia logistic 模型的检验

用 ROC 曲线分析模型的合理性：

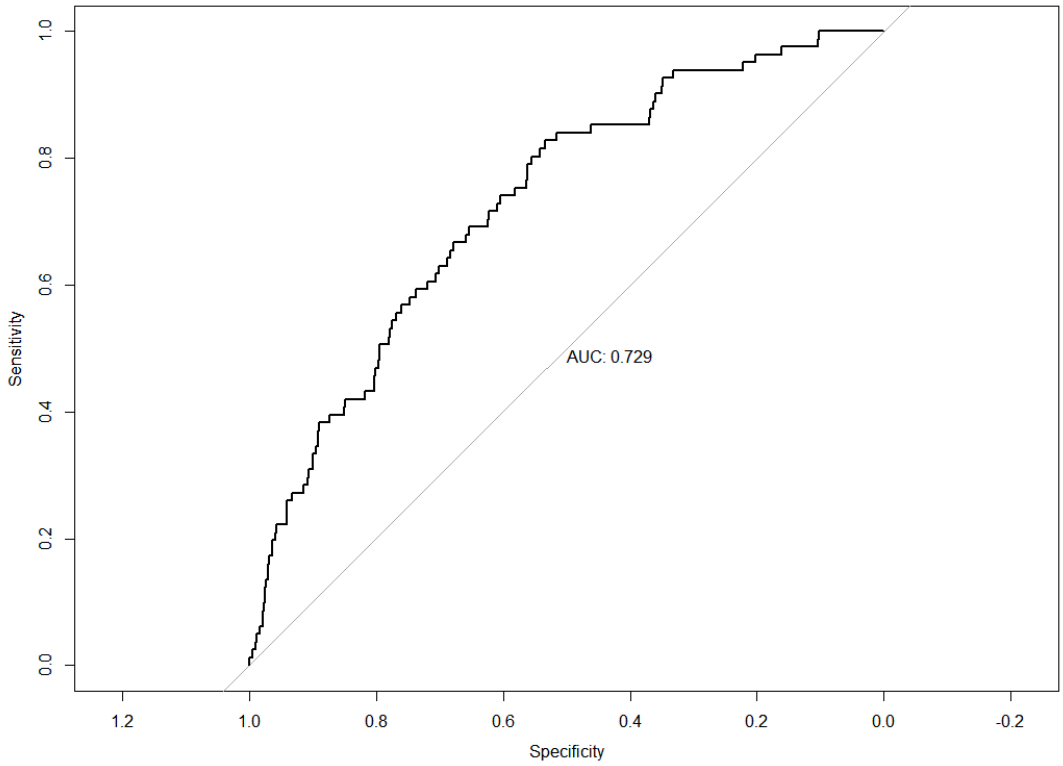


图 6-logistic 模型的 ROC 曲线

AUC 值为 0.729，接近 0.8，可认为上述模型是合理的。

### 5.2.3 myopia logistic 结果分析

根据 logistic 回归模型，可以发现，近视的概率由前房深度(ACD)、年龄(AGE)、户外运动时长(SPORTHR)、近距离工作时长(WORKHR)和父母遗传(MOMMY、DADMY)决定。

我们无法改变青少年的前房深度与遗传基因，在此只考虑环境因素(SPORTHR 和 WORKHR)对视力的影响。用控制变量法分析视力的演化机理模型，并画出近视可能性关于户外运动时长、近视可能性关于近距离工作时长的图像如下图。

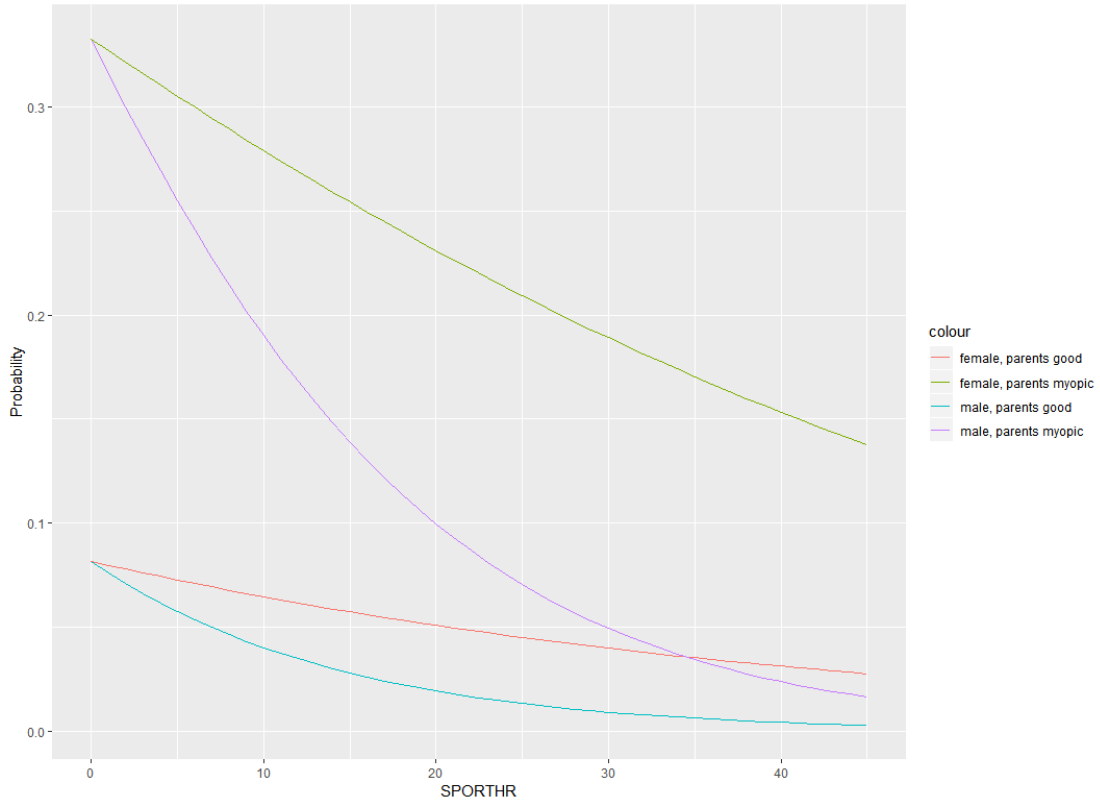


图 7-近视可能性随户外运动时长的变化

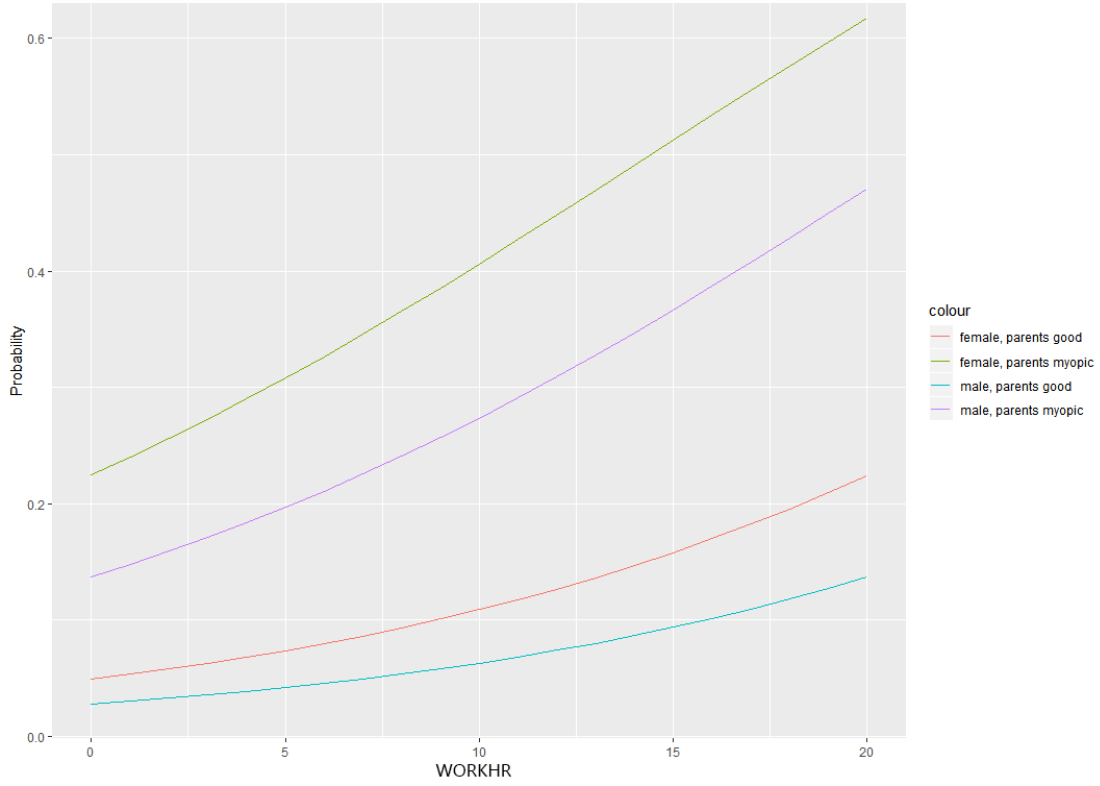


图 8-近视可能性随近距离工作时长的变化 1

分析结果显示:①双亲近视的女性患有近视的风险最高，双亲正常的男性患有近视的风险非常低。②无论是男性还是女性，近视可能性随着户外运动时间增加而降低。户外运动对双亲近视的男性降低近视风险的作用尤为显著③阅读时间过长会大大增加双亲近视的人患近视的风险。

## 5.3眼睛视力的预警机制模型

在模型二中，我们通过收集青少年的数据来计算出其患近视的可能性。我们将这些可能性作为近视的风险，设置近视的预警机制模型。

### 5.3.1 目标数据的收集

收集的目标数据除了上文提到的性别(GENDER)、前房深度(ACD)、户外运动时长(SPORTHR)、近距离工作时长(WORKHR)、父母近视情况(MOMMY、DADMY)以外，还可以通过智能设备采取工作时光照强度、工作时眼与桌面的距离等不易测量的数据。

### 5.3.2. 目标数据的处理

假设青少年缓解视觉疲劳的周期为$T$，$t_{W_i}$、$t_{S_i}$分别是数据收集第$i$天$(1\le i\le T)$的近距离工作时长和户外运动时长。

统计出近$i$天的平均近距离工作时长$\bar{t}_W$和平均户外运动时长$\bar{t}_S$：

$$t_W=\sum_{j=1}^{i}t_{W_i}/T\ ,\quad t_S=\sum_{j=1}^{i}t_{S_i}/T$$

将$\bar{t}_W$与$\bar{t}_S$代入问题二眼睛视力的演化机理模型亦即 myopia logistic 回归模型，计算出罹患近视的风险$P$。

### 5.3.3 预警机制的设置

我们将风险$P$分为三个等级：低风险$(0<P\le 40\%)$、中等风险$(40\%<P\le 60\%)$和高风险$(60\%<P<100\%)$。当$P$为低风险时，不作任何提醒，仅仅反馈近视风险，说明用眼在正常范围之内；当$P$为中等风险时，采用黄色预警并反馈近视风险，提醒用户注意用眼；当$P$为高风险时，采用红色预警并反馈高亮后的近视风险，提醒用户用眼过度，注意休息。

优化：我们参考医学研究中，青少年和成人生理特征的区别，将设置不同的最低运动时长，和最高工作时长；并且根据不同个人的最小规定日常工作时长来规划非规定工作时间的作息。

### 5.3.4 专家建议系统

向专业机构咨询适度用眼方式并结合从大数据中统计出所有低风险用户的平均数据，拟出一系列正常指标。若判断用户为近视高风险，则将该用户的各项统计值（$t_{W_i}$ 、 $t_{S_i}$ 等）与正常指标做比对，判断具体是哪一项用眼活动不达标。再结合专业机构的咨询意见，给出用户纠正用眼习惯的指导性建议。

实现眼睛视力的预警机制模型的思路用如下流程图体现：

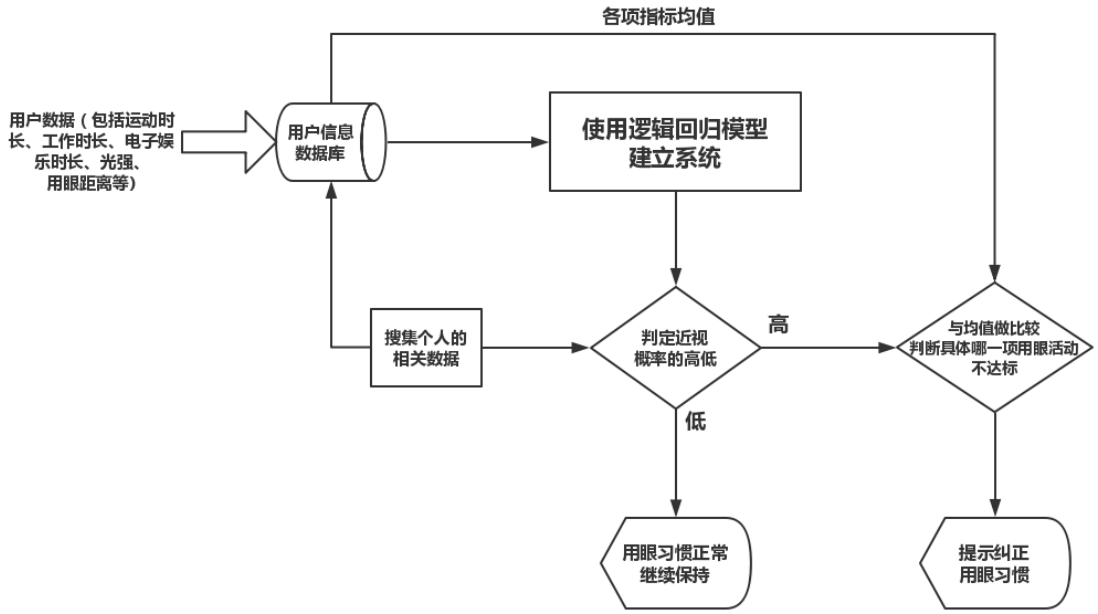


图 9-预警机制的流程图

## 5.4 机器学习模型

在使用 AI 算法中，为了求得 logistic 回归中的模型参数，可以将最大似然估计法和梯度上升法结合使用，具体求解如下：

### 5.4.1 建立似然函数

$$\mathrm{L(w)} = \mathrm{P(y|x;w)} = \prod_{i=1}^{n} P\left(y^{(i)}\middle|x^{(i)};w\right) = \prod_{i=1}^{n}((\emptyset(x^{(i)}))^{y^{(i)}}(1-\emptyset(x^{(i)}))^{1-y^{(i)}})$$

式中∅()为在问题二定义的 sigmoid 函数。

为了方便计算，取对数：

$$\ln \mathrm{L(w)} = \ln \mathrm{P(y|x;w)} = \ln \prod_{i=1}^{n} P\left(y^{(i)}\middle|x^{(i)};w\right) = \prod_{i=1}^{n}\left(y^{(i)} \ln(\emptyset(x^{(i)})) + (1-y^{(i)}) \ln(1-\emptyset(x^{(i)}))\right)$$

### 5.4.2 梯度上升法

可知当 w 使 L(w)最大的时候，w 最合理，所以采用梯度上升法来求解参数。

梯度上升的主要思想就是沿着函数的梯度方向寻找函数的最值。

首先计算函数的梯度：

$$\frac{\partial \ln L(\overline{w})}{\partial w_k} = \sum_{i=1}^{n} x_{ik}(y_i - \frac{1}{1+e^{\overline{w}\cdot\overline{x}}})$$

使用矩阵乘法直接表示梯度：

$$\nabla \ln L(\overline{w}) = \overline{x} \cdot \overline{\pi(\overline{x})} = \overline{x} \cdot \overline{error}$$

设步长为α(阈值)，则迭代得到的新的权重参数为：

$$\overline{w} = \overline{w} + \alpha \nabla lnL(\overline{w})$$

实现机器学习模型的思路用如下流程图体现：

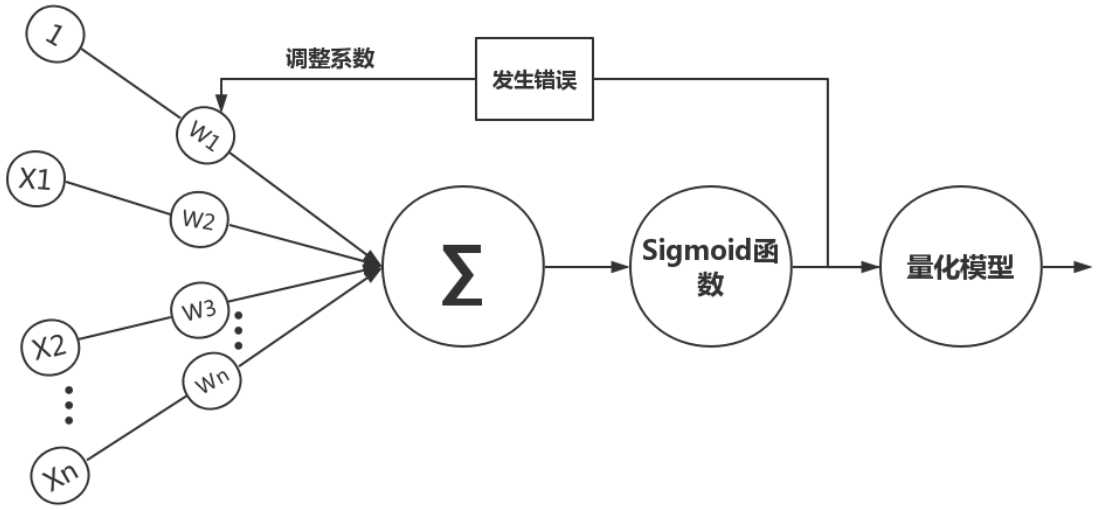


图 10-机器学习的流程图

## 5.5 实现模型

实现上文所提到的机理模型和学习模型，需要做到以下两点：

①搜集大量数据。在本题中我们收集了大量包含性别、年龄、生理结构(屈光度、轴向长度、前房深度、玻璃体腔深度等)、环境因素(运动时长、工作时长、电子器件使用时长等)和遗传因素(父母是否为近视)的数据。同时，在我们已有模型之上，我们可以通过收集更多的环境信息，如阅读距离、阅读环境光强、阅读角度等，对模型进行改进。

②了解眼科学相关知识。如屈光度和近视程度的关系，青少年和成年人分别的平均轴向长度区间、平均前房深度区间和平均玻璃体腔深度。我们的 AI 预警机制将会据此建议，比如年龄稍大些的成年人前房长度变化较小，而青少年的前房长度更易改变，因此建议 5 到 9 岁儿童不使用或少电子器件，并且建议每日的阅读时长不超过 3

小时；9 到 16 岁的青少年应尽量少用电子器件，并且每日的阅读时长不应超过 6 小时。

# 六、模型的评价和推广

## 6.1 模型的优点

模型简洁，运算量小，拟合度高，通过对数据的分析不断纠正线性系数，最终获得不同影响对最终视力的影响。同时模型可扩展性高，只需要将数据集扩充即可。

## 6.2 模型的缺点

本模型选取的数据集有待完善，具体问题有以下三点：①数据集不均衡，据统计，数据集中只有 15%的学生是近视的②实验年份是 1990-1995，与现在差异较大，可移植性不强。那个时代电子行业发展方兴未艾，电脑、电视的普及度远不及现在，因此，与现在的孩子相比，当时的孩子在电脑上工作的时间（COMHR）、看电视的时间（TVHR）普遍更少，户外体育活动时间（SPORTHR）普遍更长。③受访群体范围过小，不具有普适性。受访群体是 5-8 岁儿童，模型的适用群体应为 1-18 岁青少年群体。

## 6.3 模型的推广

通过与智能护眼公司（如云夹）的合作搜集光强数据，被测试者家中光源颜色，被测试者阅读姿势(比如阅读角度、阅读时书与面的距离)、日常工作规定时长、日常学习规定时长等数据，扩充解释变量，使用主成分回归分析对数据分析，使得模型更加精确，同时引入日常规定工作学习时长，来修正非规定工作学习时间外的被测试者行为。

# 七、参考文献

[1]司守奎 孙兆亮. 数学建模算法与应用[M]. 第 2 版. 国防工业出版社, 2017.

[2]郭呈全 陈希镇. 主成分回归的 SPSS 实现[EB/OL]. [2019.5.3]. https://wenku.baidu.com/view/b5b0ce0003d8ce2f0066233e.html.

[3]Terence Zhi Liu Chaoyang Wang. Myopia Study with Logistic Regression[EB/OL]. [2019.5.3]. http://astro1.panet.utoledo.edu/~terencezl/projects/myopia.html.

[4]Little_Rockie. 如何在 R 语言中使用 Logistic 回归模型[EB/OL]. [2019.5.3]. https://www.cnblogs.com/nxld/p/6170690.html.

# 附录

**梯度上升与最大似然算法的代码实现：**

```
import copy
import csv
import random

import numpy as np
import pandas as pd
import pylab as pl
import statsmodels.api as sm

# 数据预处理，AGE 变为虚拟变量，加入回归所需的 intercept 变量
df=pd.read_csv("myopia.csv")
# 将离散值变为布尔值
dummy_ranks=pd.get_dummies(df['AGE'],prefix='AGE')
tablist=df.columns.values.tolist()

# 选出关键的因素
tablist.pop(1)
tablist.pop(2)
tablist.pop(2)
tablist.pop(3)
tablist.pop(3)
tablist.pop(6)
tablist.pop(6)
data=df[tablist].join(dummy_ranks.ix[:,'AGE_2':])

# 设置初始常量
data['intercept']=1
```

```
## 进行逻辑回归
train_cols=data[data.columns[1:]]

# 定义 sigmoid 函数
def sigmoid(inX):
    return 1.0/(1+np.exp(-inX))

# 梯度上升求最优参数
def gradAscent(dataMat, labelMat):
    dataMatrix=np.mat(dataMat)
    classLabels=np.mat(labelMat).transpose()
    m,n=np.shape(dataMatrix)
    alpha=0.0001
    maxCycles=10000
    weights=np.ones((n,1))
    for k in range(maxCycles):
        h=sigmoid(dataMatrix*weights)
        error=(classLabels-h)
        weights=weights+alpha*dataMatrix.transpose()*error
    print(weights)
    return weights
weights=gradAscent(train_cols,data['MYOPIC']).getA()

#检验系统预测的成功概率
test_data = copy.deepcopy(data)
test_data['intercept']=1.0
predict_cols= test_data[test_data.columns[1:]]
predict=[]
test=np.mat(predict_cols)
```

```
for i in test:
    sum=sigmoid(i*np.mat(weights))
    if sum<0.5:
        predict.append('0')
    else:
        predict.append('1')
test_data['predict']=predict

length=len(data['MYOPIC'])
test_result={'0 0':0,'0 1':0,'1 0':0,'1 1':0}
for i in range(0,length):
    string=str(test_data.loc[i,'MYOPIC'])+' '+str(test_data.loc[i,'predict'])
    test_result[string]+=1

predict_right=0
for i in range(0,length):
    if int(test_data.loc[i,'MYOPIC'])==int(test_data.loc[i,'predict']):
        predict_right+=1

print('准确率为:%.5f'%(predict_right/length))
for i in test_result.keys():
        print(i+' '+str(test_result[i]))

# 生成供机器学习的表格(前系统成功率 64%)
# 在最大最小范围内生成随机
def rand_csv(df):
        max_list=[]
        min_list=[]
        for i in df.columns.values.tolist():
                max_list.append(df[i].max())
```

```
                min_list.append(df[i].min())
        with open('rand_result.csv','w',newline='') as f:
                writer=csv.writer(f)
                writer.writerow(df.columns.values.tolist())
                for i in range(10000):
                        write_row=[]
                        for j in range(len(df.loc[1])):
                                if max_list[j]==1:

write_row.append(random.randint(0,1))
                                elif max_list[j]==9:
                                        write_row.append(random.randint(5,9))
                                else:

write_row.append(random.uniform(min_list[j],max_list[j]))

                        writer=csv.writer(f)
                        writer.writerow(write_row)

        f.close()

rand_csv(df)

# 数据预处理，AGE 变为虚拟变量，加入回归所需的 intercept 变量
df=pd.read_csv("rand_result.csv")
dummy_ranks=pd.get_dummies(df['AGE'],prefix='AGE')
tablist=df.columns.values.tolist()
tablist.pop(1)
tablist.pop(2)
tablist.pop(2)
```

```
tablist.pop(3)
tablist.pop(3)
tablist.pop(6)
tablist.pop(6)
data=df[tablist].join(dummy_ranks.ix[:,'AGE_2':])
data['intercept']=1
#检验系数
test_data = copy.deepcopy(data)
test_data['intercept']=1.0
predict_cols= test_data[test_data.columns[1:]]
predict=[]
test=np.mat(predict_cols)
zer=0
one=0
cnt=0
for i in test:
    cnt+=1
    sum=sigmoid(i*np.mat(weights))
    if sum<0.5:
        test_data.loc[cnt,'MYOPIC']=0
        zer+=1
    else:
        test_data.loc[cnt,'MYOPIC']=1
        one+=1
# 展示结果
print(zer)
print(one)
# 写入
test_data.to_csv('data_df.csv',index=False,header=True)
```