本科生毕业论文 (设计)

题　　目　保险中的数据科学

学　　院　经济学院

专　　业　保险精算

学生姓名　蔡嘉文

学　　号　2012141013115　年级　2012

指导教师　李旸

教务处制表

二〇一六年一月七日

保险中的数据科学

经济学院

学生: 蔡嘉文　　指导教师: 李旸

[摘要]

大数据时代背景下，拥有客户多方面数据的保险公司迎来了机遇与挑战。通过运用数据科学的分析方法，可以挖掘潜藏在数据中的信息，并对信息进行加工利用,从而辅助优化保险产品的定价和理赔额的设计，并且可能产出更好的产品或提供更好的服务。

本文将根据数据科学的研究范式，即通过数据清洗到模型检验的一系列步骤，对几个保险公司相关的数据集进行探索和研究，并将结果与通过传统统计方法所得到的结果进行比较。

[主题词] LaTeX; 论文; 毕业设计; 模板

How to use this template to write thesis

Economy

Student:Kevin Cai Adviser:Li Yang

[Abstract]

The problem I am trying to solve in this paper is to introduce a new

The approach I adopt to solve the problem is ...

The results obtained in this research include ...

The impacts of our obtained results are ...

[Key Words] LaTeX; Thesis; Graduation Project; Template

目录

1 绪论

1.1 意义和背景

1.regularized logistic + feature engineer 2.twitter api + outlier detection + fraud detection

1.2 文献综述

1.2.1 数据科学在保险业中的应用

李娜娜（2013)，介绍了数据挖掘基本理论，对医疗保险进行了需求分析，介绍了数据仓库的设计以及数据的结构及存储，分别通过聚类进行区别定价，决策树进行客户风险控制，神经网络进行欺诈案件识别。Varun Chandola（2008）描述了健康保险理赔数据的类型和特征，指出在健康保险中的三种问题：欺诈，浪费，滥用，使用文本挖掘（LDA主题模型识别欺诈模式)，社交网络分析，序列分析识别欺诈并提高健康保险运作效率。

1.2.2 医疗风险评级

Moturu(2009)[1] 采用非随机抽样平衡敏感性 (Sensitivity) 与特异性（Specificity)，并在 Adaboost，SVM 等算法中实现以牺牲少量特异性的情况下大幅提高敏感性。

[1] 采用非随机抽样平衡敏感性 (Sensitivity) 与特异性（Specificity)

2 方法论基础

本文将首先从机器学习的角度介绍一些具有代表性的分类算法的运作步骤，并给出对应核心部分的 MATLAB 或 R 代码。然后将用小部分篇描述正则化，交叉验证等关于模型选择和模型评价方面的知识。

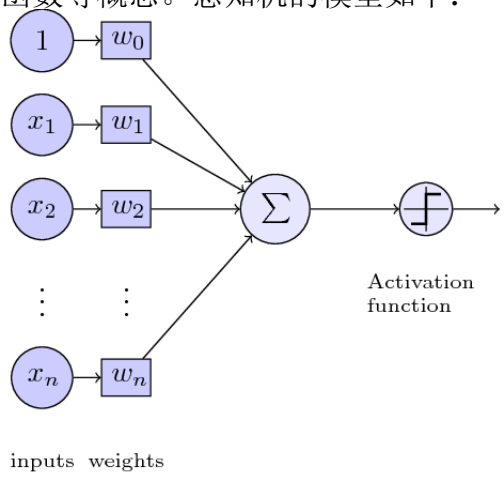
2.1 分类算法

在机器学习与统计中，分类是指识别出样本所属类别的问题。分类问题在机器学习中属于有监督学习的范畴，即通过一组包含自变量与因变量（有限种取值）的训练模型，并通过模型对其他没有因变量的样本进行预测。实务中的例子有垃圾邮件分类（通过邮件内的词语集合判断邮件是否为垃圾邮件），信用风险评级等。

现有的分类算法分为广义线性分类算法，支持向量机算法，决策树算法，神经网络算法几大类。下面对几种具体算法进行介绍。

2.1.1 感知机

感知机算法是由 **Rosenblatt**于 1957 年所发明的一种二类分类的线性模型，可以看作最简单的前向神经网络。在神经科学中，神经细胞的状态取决于从其它的神经细胞收到的输入信号量，及突触的强度（抑制或加强）。当信号量总和超过了某个阈值时，细胞体就会激动，产生电脉冲。**Rosenblatt**被生物神经细胞的运作方式所启发，提出权重，阈值，激活函数等概念。感知机的模型如下：



2.1.2 Logistic 回归

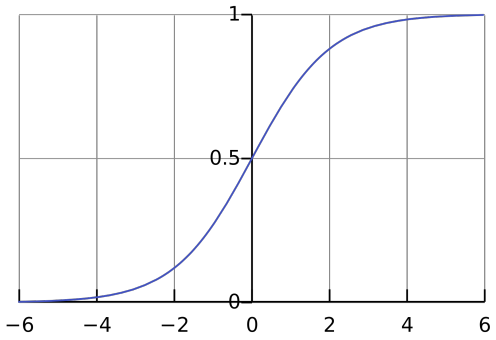
2.1.2.1 logit 函数的背景，形态及特性

logit 函数的数学表达式为：

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

其所具有的数学特性为：

$\frac{d}{dx} f(x) = f(x) * (1 - f(x))$



logit 函数的具有的生态学背景 https://www.zhihu.com/question/36714044

1. 生物种群在无约束情况下，数量呈指数增长
2. 增加约束，设立上限，则自然地得出有约束的公式

此外，logit 函数

```
import numpy as np

def incmatrix(genl1,genl2):
m = len(genl1)
n = len(genl2)
M = None #to become the incidence matrix
VT = np.zeros((n*m,1), int)  #dummy variable

#compute the bitwise xor matrix
M1 = bitxormatrix(genl1)
M2 = np.triu(bitxormatrix(genl2),1)

for i in range(m-1):
for j in range(i+1, m):
[r,c] = np.where(M2 == M1[i,j])
for k in range(len(r)):
VT[(i)*n + r[k]] = 1;
VT[(i)*n + c[k]] = 1;
VT[(j)*n + r[k]] = 1;
VT[(j)*n + c[k]] = 1;

if M is None:
```

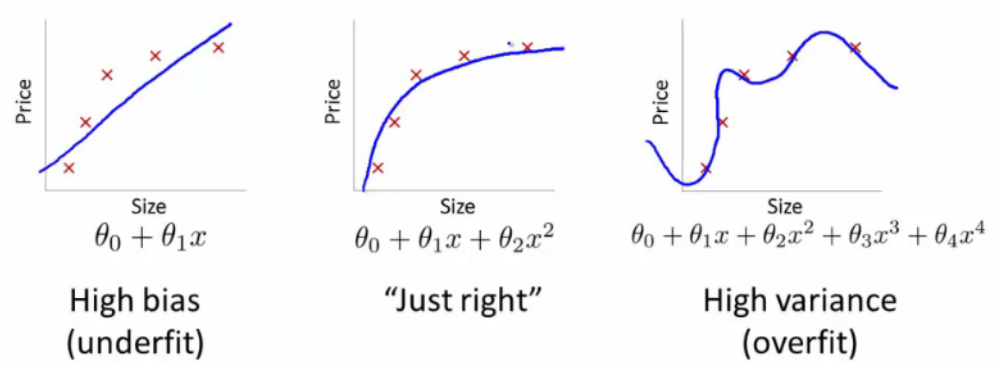
```
M = np.copy(VT)
else:
M = np.concatenate((M, VT), 1)

VT = np.zeros((n*m,1), int)

return M
```

2.2 模型选择与验证

我们首先借助一张图来说明欠拟合和过拟合，偏差和方差的概念。



我们可以看到，当用线性模型来拟合数据时造成了欠拟合，模型在训练集和测试集均有较大的误差。

而使用四次模型时，模型似乎完美地拟合了训练样本，但是由于模型过度追求拟合训练样本 (将样本的误差，噪音等一并进行了学习)，破坏了其一般化 (generalization) 的能力，最终在测试集上也不会有较好的表现，这样的现象被称为过拟合。

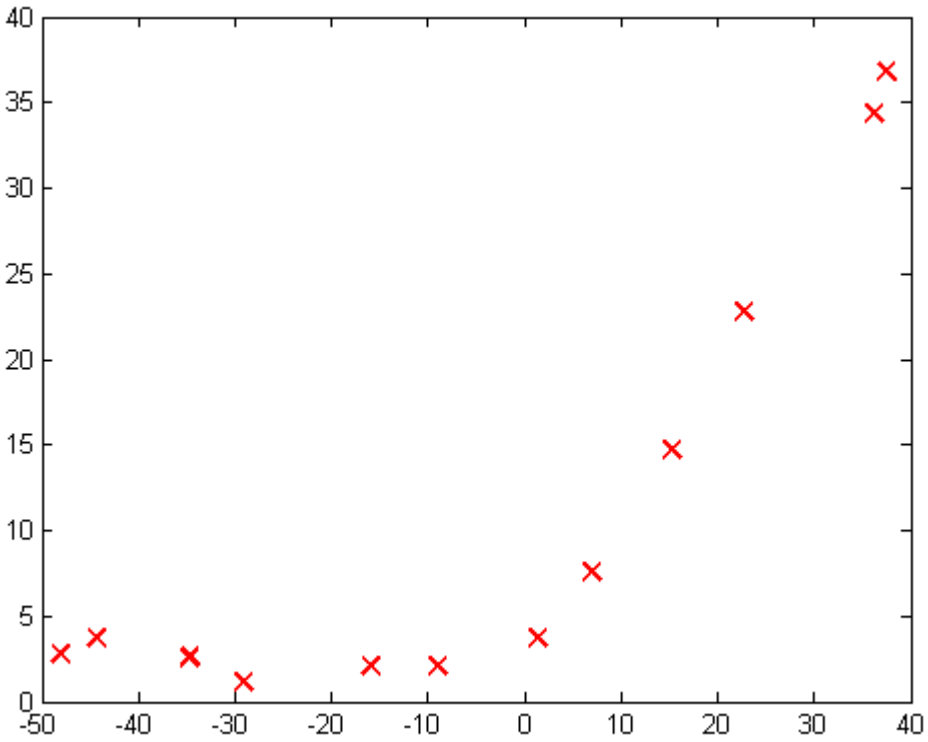
下面用一个例子来进一步说明上面的问题并引入学习曲线的概念。

```
%% ============ Part 1: Loading and Visualizing Data =============
X =  [-15.9368;-29.1530;36.1895;37.4922;-48.0588;-8.9415;15.3078;-34.7063;1.3892;-44.

y = [2.1343;1.1733;34.3591;36.8380;2.8090;2.1211;14.7103;2.6142;3.7402;3.7317;7.6277;

% m = Number of examples
m = size(X, 1);

% Plot training data
plot(X, y, 'rx', 'MarkerSize', 10, 'LineWidth', 1.5);
```

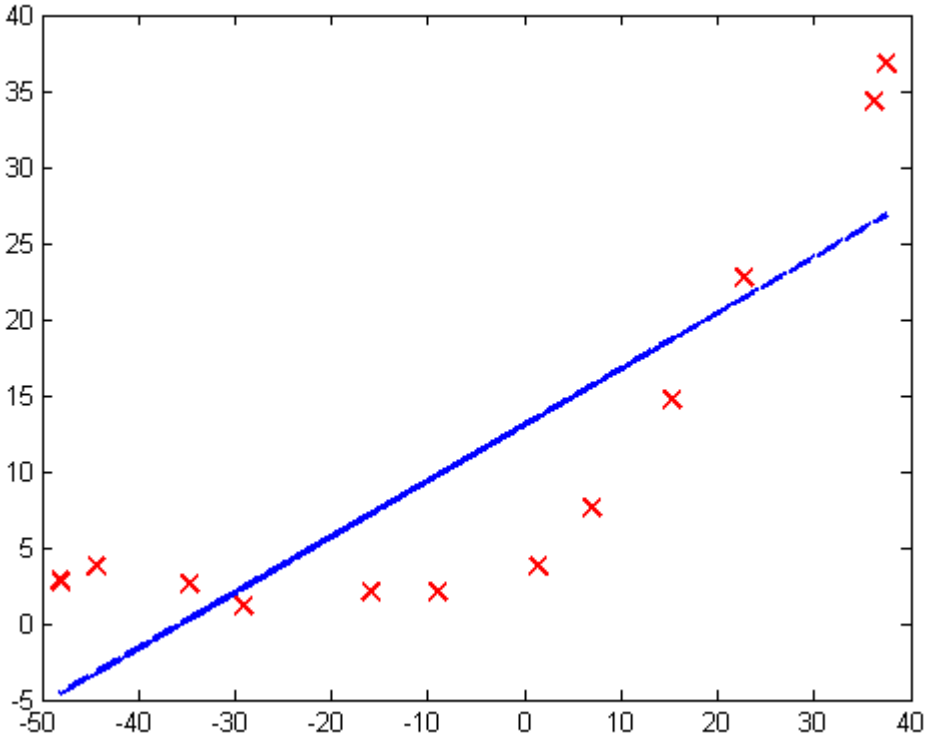



```
%% ========== Part 2: Perform Linear Regression =============

%Add a column to X to compute theta0
X1 = [ones(m,1) X];

%Use Normal Equation to compute theta
theta = pinv((X1'*X1))*X1'*y;

%Plot the regression line
plot(X, y, 'rx', 'MarkerSize', 10, 'LineWidth', 1.5);
hold on;
plot(X, [ones(m, 1) X]*theta, '--', 'LineWidth', 2)
hold off;
```


40
35
30
25
20
15
10
5
0
-5
-50
-40
-30
-20
-10
0
10
20
30
40

3 案例 1：D2Hawkye 客户健康风险评级

3.1 探索性数据分析

本模板需要依赖于LaTeX 2ε、XeTeX 以及 CTeX，因此在你使用前请确保这些发行版已经安装妥当。本文件全部使用 UTF-8 编码，并使用 XeLaTeX 编译，以支持国际化和 TrueType 技术字体。

3.2 模型训练

本模板由以下文件构成：

- main.tex - LaTeX 基本框架，你可以在此添加你需要的 Package
- make.bat - 运行在 Windows 上的编译脚本，双击即可执行，他可以免去你敲代码编译的麻烦 ☺
- Makefile - 运行在 Linux 上的编译脚本，使用 `make` 命令完成编译
- scuthesis.sty - 川大毕设论文格式样式包，你不需要了解这个文件（除非本模板板式不合符你的需求）
- src/basic_info.tex - 定义论文作者基本信息
- src/prologue.tex - 包含了封面、中英文摘要以及目录的定义
- src/epilogue.tex - 包含了论文参考文献、附录等信息
- src/ch*.tex - 论文每一章具体内容
- ref/refs.bib - bibtex 文献库，推荐使用 JabRef 维护

3.3 模型评价

首先说明一下，本教程是一篇 self-contained 的文章，本文章是直接编译本 LaTeX 模板得到，你可以具体参考模板源代码内容以学习如何使用。但是为了阐明脉络，下面我将以一次完整使用的形式展示如何使用本模板。

3.3.1 一个三级标题

首先，你需要填写自己的基本信息，例如姓名、学号之类，你需要打开 basic_info.tex 文件将其填写进去。然后你需要书写你的摘要，在 prologue.tex 里面是中英文摘要的定义处，你可以在其中编写摘要。假设你是 LaTeX 的老用户，你可能需要自己包含一些 Package，那么你可以在 main.tex 中添加 usepackage 命令。

3.3.1.1 四级标题

另外，随着章节数目增多，你可以自行新建 chxx.tex 文件，并将其在 main.tex 中用 include 指令包含进来。最后，为了编译你的论文，你需要使用 xelatex 命令。不过目前 Windows 用户可以直接双击 make.bat 生成论文。

3.4 问题

1. 中英文摘要分别不能超过一页，否则第二页的板式会有问题（由于本人精力有限，且该问题出现几率较小，目前暂未打算修复这个问题）。
2. 所有文件必须是 UTF-8 编码，否则编译不能通过。

参考文献

[1] Moturu S T, Liu H, Johnson W G. Healthcare Risk Modeling for Medicaid Patients - The Impact of Sampling on the Prediction of High-Cost Patients Learning from Imbalanced Data [J]. (B1) Biomedical Engineering Systems and Technologies. 2009: 126–133.