

硕士研究生学位论文

|  |  |
| --- | --- |
| 题目： | 面向微博的在线广告点击 |
|  | 行为预测系统设计与实现 |

|  |  |
| --- | --- |
| 姓 名： | 孙超 |
| 学 号： | 1301221193 |
| 院 系： | 软件与微电子学院 |
| 专 业： | 软件工程 |
| 研究方向： | 互联网信息挖掘与处理 |
| 导师姓名： | 王海峰、刘宏志 |

二〇一 年 月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

# 摘要

广告点击行为预测是互联网计算广告的重要环节，通常的做法是先用机器学习算法在已有用户的特征数据和点击行为记录上训练出一个模型，然后用训练出的模型结合新的用户特征数据返回该用户点击行为的概率。以往工作的重点是训练模型参数选取和用户特征选取上，其中用户特征选取多集中于结构化特征，即如用户的性别、年龄、职业、教育程度等固有属性。

本文在用户结构化特征基础上，利用用户的微博行为数据（包括发表、转发、点赞）过的文章，通过特征工程转换为用户的非结构化特征，用GBDT和LR融合的方法做点击行为预估，并设计和实现了一个可商业化运营的线上分布式系统。实验和线上运行记录表明该方法较传统的只用结构化特征的方法在预估准确率方面有了明显的提高。

关键词：广告，点击率，预测

ENGLISH TITLE

Author Name ( Major )

Directed by your Supervisor

# ABSTRACT

In environmental economics, environmental resources including environmental quality are categorized as amenity resources. Due to its importance to human welfare, the amenity resources theoretical study and valuation is an ongoing issue at the academic frontier in the environmental economics area.

KEY WORDS: Key word 1, Key word 2, Key word 3, ……

# 目录

[摘要 I](#_Toc469264943)

[ABSTRACT II](#_Toc469264944)

[目录 III](#_Toc469264945)

[第一章 绪论 1](#_Toc469264946)

[1.1 项目背景 1](#_Toc469264947)

[1.2 广告点击率预测研究发展现状 1](#_Toc469264948)

[1.3 本文的研究内容和方法 2](#_Toc469264949)

[1.4 本文内容组织结构 3](#_Toc469264950)

[第二章 相关技术 4](#_Toc469264951)

[2.1 逻辑回归模型 4](#_Toc469264952)

[2.2 在线学习算法FTRL 4](#_Toc469264953)

[2.3 迭代决策树 GBDT 4](#_Toc469264954)

[2.4 GBDT结合LR用于CTR预测 4](#_Toc469264955)

[2.5 分布式计算平台 4](#_Toc469264956)

[2.6 本章小结 4](#_Toc469264957)

[第三章 面向微博的在线广告点击行为预测需求分析 5](#_Toc469264958)

[第四章 图表示例 7](#_Toc469264959)

[第五章 结论及展望 8](#_Toc469264960)

[参考文献 9](#_Toc469264961)

[附录A 附录示例 12](#_Toc469264962)

[致谢 13](#_Toc469264963)

[北京大学学位论文原创性声明和使用授权说明 14](#_Toc469264964)

注：目录从第1章开始，前边因页眉需要设置了标题，实际使用时更新后去掉前边部分。使用时请删除本注释。如本示例，更新目录后删除前边三项（摘要、ABSTRACT、目录）即可。

# 第一章 绪论

## 1.1 项目背景

当前我国互联网产业迅速发展，根据中国互联网络信息中心（CNNIC）的最新统计，截至2016年6月，我国网民规模达7.10亿，互联网普及率达到51.7%，与2015年底相比提高1.3个百分点，超过全球平均水平3.1个百分点，超过亚洲平均水平8.1个百分点[1]。与此同时，电子商务产业也发展迅速，十二五期间，中国的网络零售交易额规模跃居世界第一，网购在网民中的普及率高达55.7%，仅在2016年双11期间，天猫交易额就突破1200亿元，物流订单量超6.5亿[2]。

在此背景下催生了互联网广告这一新兴产业，互联网广告经历了从最初铺天盖地的横幅广告到人群及兴趣精准定向的搜索广告与推荐引擎，进而催生了计算广告学这一新兴研究课题，这一课题涉及大规模搜索和文本分析、信息获取、统计模型、机器学习、分类、优化以及微观经济学等诸多领域的知识[3]。相比于基于传统媒体如电视、广播、报纸海量而无特定目标的广告投放，以互联网为媒介的广告投放可以根据用户特征（主要包括用户个人信息和上网行为）做到个性化精准投放。从而为大大提高了广告收益，为企业节省了推广成本。其中，广告点击率预测(Click Through Rate 简称CTR)是计算广告学的一个重要研究内容，是互联网企业广告竞价排名和流量变现业务的重要支撑。

## 1.2 广告点击率预测研究发展现状

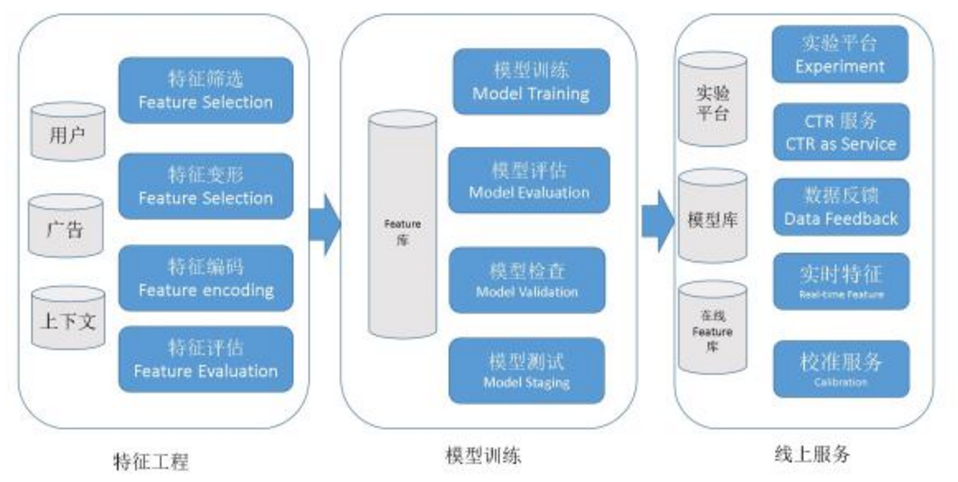
目前广告点击行为预测研究主要集中于特征学习、数据特点和用户行为这3个方面，本文将以这3个方面分别介绍已有的相关工作。

特征学习方面，影响CTR预估准确率的特征很多，不是选用的特征越多准确率就会越高，如何科学的选取又用的特征是不少学者和工程师的研究方向。在这方面，文献[4]研究了原始查询和重写后查询之间的相关性与广告点击率之间的关系，该文章考察了一些特征如次序、长度差异、编辑距离等对广告点击率的影响；文献[5][6]将广告显示位置和广告查询相关性作为特征，同时根据相似广告来解决非常见广告和新广告的点击率预测问题；文献[7]综合运用协同过滤、贝叶斯网络和特征工程等模型来预测点击率；文献[8]提出用协同过滤和张量分解来提取用户个性化特征用于点击率预测；

数据特点方面，广告数据多为稀疏数据，文献[4]提出利用相似广告来预估目标广告的点击率；针对新广告的历史数据缺失问题，文献[9]提出“竞拍词-广告主”矩阵，用层次聚类的方法解决历史数据不充分的问题；文献[10]设计了用于稀疏广告和新广告的点击率预测模型，提出了基于层级结构的预估模型和基于Time-Spatial的预估模型；文献[11]提出了基于经验贝叶斯的自然数据分层和基于数据一致性的两种平滑计算方法对层级模型做了改进。

用户行为建模，用假设检验和贝叶斯网络刻画用户浏览场景，进而估计出用户点击广告的概率。文献[12]基于用户的浏览行为假设提出了级联模型，如果用户点击一个文档，若该文档不能满足用户的查询需求，则该用户很可能继续查看后续搜索结果并有点击行为； 文献[13]扩展级联模型到多次点击。

## 1.3 本文的研究内容和方法

广告点击率预测过程可分为3个步骤：特征工程、魔性训练和线上服务。

在这其中，特征工程是很重要的一个环节，特征选取的优劣对最终预测结果的准确率有很大的影响。对于用户而言，用户的固有信息如年龄、性别、地域、职业、教育程度等信息可作为用户特征，我们称之为结构化特征；而如用户的行为数据如浏览过的网页、购买过的物品、转发过的微博等非固有的，随时间和环境变化的特征，我们称之为非结构化特征。而什么样的特征适合做CTR预估，业界并没有统一的标准。往往靠工程科研人员结合具体情况和生活常识通过反复实验和以往经验来决定。以年龄特征为例：年轻人普遍喜欢运动类的广告，30岁左右的男人喜欢车，房子之类的广告，50岁以上的人喜欢保健品的广告。再以性别特征为例：化妆品的广告在女性上面的点击率就比在男性上面的点击率高很多，又如体育用品的广告在男性上面的点击率也比女性高，说明性别这个特征在化妆品和体育行业也是有预测能力的，经过多个行业的验证，就认为性别这个特征可以用了。在实际的使用中发现，性别这个特征比较有效，手机平台这个特征也比较有效，地域和年龄这两个特征有一定效果，但没有前两个那么明显，跟它们的使用方式可能有关。以上示例说明在CTR预估上特征选择往往根据具体应用场景的，主观的。

CTR预估中用的最多的是逻辑回归模型，又称LR(Logistic Regression)模型，是一种广义线性模型。LR使用了Logit变换将函数值映射到0~1区间，映射后的函数值就是CTR的预估值。LR模型的优点是效率高，容易并行化，处理上亿级别数据不是问题。但缺点是作为一种线性模型，其学习能力有限，需要大量特征工程预先分析出有效的特征、特征组合，从而去间接增强LR 的非线性学习能力[14]。上文中提到，特征选取往往依靠人工经验，耗时耗力且不一定能带来效果提升。如何自动发现有效的特征、特征组合，弥补人工经验不足，Facebook在2014年的文章介绍了通过GBDT(Gradient Boost Decision Tree)解决LR的特征组合问题[15]，随后Kaggle竞赛将此付诸实践[16]。在此之后GBDT与LR融合的方法开始引起了业界关注。

本文在结构化特征的基础之上，利用用户的微博行为记录(包括发表、转发、点赞等行为)，使用GBDT方法在用户有过行为的微博文章里提取非结构化特征，与结构化特征一起作为LR模型的训练数据训练出线上模型。实验证明，用了此方法后CTR预估准确率较之前只使用结构化特征有了明显的提升。此外，本文还详细介绍了基于此种方法的分布式在线实时广告点击率预测系统的设计和实现。

## 1.4 本文内容组织结构

第一章：绪论，介绍了广告点击率预测背景、发展现状和本文的主要工作，最后总结文章的组织结构。

第二章：相关技术，

# 第二章 相关技术

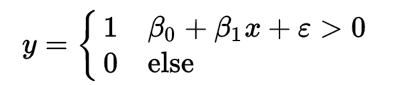
## 2.1 逻辑回归模型

逻辑回归(Logistic Regression)模型，是机器学习中的一种分类模型，由于算法的简单和高效，在社会学、生物统计学、临床、数量心理学、计量经济学、市场营销等领域得到广泛的应用。

逻辑回归算法最早由统计学家David Cox提出[17][18]，逻辑回归有二分类、多分类和顺序分类[19] 。二分类模型是应用最多的模型，其最终输出结果值只有两种，如“生”或“死”，“胜利”或“失败”，在本文的应用场景里就是用户“点击”或“未点击”广告，通常这两种结果值用“0”和“1”来表示[20] ，本文中“0”表示未点击，“1”表示点击。逻辑回归被用作预测基于相互独立变量基础上某一事件发生的几率，

和其他的回归分析方法类似，逻辑回归使用一个或多个预测变量，这些变量可以是连续的也可以是离散的。和一般线性回归不同的是，逻辑回归是用来预测相互依赖的二项变量而非连续值结果。而这违反了线性回归的假设前提，尤其是残差不服从正态分布。此外，线性回归可能对相互依赖的二项分布的变量做出无意义的预测。因此需要一种方法将二项分布值转换为连续实数值。要做逻辑回归首先用相互独立的变量表示不同等级事件发生的几率，然后计算出这些几率所占比率，然后再对这些计算出来的几率取对数，这一步称作对数变换，被看作是逻辑回归的连接函数。

对数函数的结果被用作线性回归分析预测，预测的结果通过自然对数的反函数（指数函数）转换为事件发生几率。因此，尽管观察到的逻辑回归预测结果是0-1值，但逻辑回归实在预测一个连续值得概率。在有些应用场合中，这个连续的概率值就是最终所需；在其他应用场景则需要一个或是或非的布尔值，通过定义一个阈值将连续值转换为离散的布尔值。

逻辑回归可被看作是寻找一组最有的β参数以满足：  
这里ε是逻辑斯谛分布误差。这里定义了一个隐变量 ，误差项ε未被测量，因此变量y’也是不可估的，因此被称作隐变量。与普通回归分析不同的是，β参数不可以直接用y和x的方程式表示，相反，需要用反复迭代过程寻找这样一个表达式，这正是逻辑回归训练算法要做的工作。

### 2.1.1 逻辑函数

逻辑函数是逻辑回归模型的重要组成部分，逻辑函数之所以有用是因为它的输入可以是任意实数，而输出总是在(0,1)区间上，因此比较适合用来表示概率。逻辑函数定义如下：  
函数图形如下图所示：  
假设t是x的线性函数，t可以用x表示为：  
这样，逻辑函数就可以写成：  
F(x)可被看作最终预测的概率值。  
逻辑函数的反函数定义为：

两边取乘方得到：

其中g是对数函数，g(F(x))表明这个对数函数等价于线性回归表达式。F(x)表示因变量所代表事件发生的概率等于线性回归表达式的逻辑函数值。值得注意的是线性回归式的值域是全体实数，但是经过F(x)映射之后变为(0,1)；是线性回归方程式的截距；是回归系数。

因变量所表示的概率等同于线性回归式的指数函数。这恰好证明了对数算子是概率和线性回归式的连接函数。因为对数函数的值域是全体实数，因此它提供了足够的尺度使得在之基础之上进行线性回归。因此定义因变量的几率为：

对于连续的自变量，定义几率比(odds ratio)为：

这种指数关系解释了的作用：当x每增加1几率便乘以倍。如果因变量服从二项分布，那么几率比可定义为，这里a,b,c,d是2x2列联表的元素[21]。

### 2.1.2 模型调整

回归系数的预估一般用最大似然估计，最大似然估计也称最大概率估计，是用来估计一个概率模型参数的一种方法，其基本原理是：由已知实验结果（样本）出发，寻找一组满足这组样本分布的参数，把可能性最大的那组作为真实的参数。

设为随机变量，是多维参数向量，若随机变量相互独立且都服从，则可以得到概率函数为：

在固定时，以上函数式表示的概率；当已知时，它又变成的函数，记为：

称此函数为似然函数。似然函数值的大小意味着该样本值出现的可能性大小，由于已经得到样本值，可以估计似然函数值是比较大的，即出现上述样本可能性是高的，因此最大似然估计就是选择使达到最大值得那个作为真实模型的参数的估计。要寻找使最大的，只需在上对求导，令导函数等于0，为了求导方便，对式两边取对数得到：

对求导得到关于的k个方程组成的方程组：

解此方程组得到使函数最大的一组。

与带正态分布残差的线性回归不同的是，要寻找一个系数值的闭合表达式去最大化似然函数是不可能的，因此必须用多次迭代过程，如牛顿法[22]，该方法从一个尝试解开始在每次迭代过程中做修正看是否比之前有准确度提升直到准确度达到稳定值，称之为收敛[23]。

## 2.2 在线学习算法FTRL

## 2.3 迭代决策树 GBDT

## 2.4 GBDT结合LR用于CTR预测

## 2.5 分布式计算平台

参考“计算广告 互联网商业变现的市场与技术”9.5章节

## 2.6 本章小结

# 第三章 面向微博的在线广告点击行为预测需求分析

# 第四章 图表示例[[1]](#footnote-1)



图3.15 全球NAT-CO2-2007清单与PKU-CO2-2007比较的空间示意图

表3.5 室外细菌气溶胶香农-维纳指数（H）和均匀性指数（E）

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Stage 1 (>7.1 μm)** | | | | **Stage 2 (4.8-7.1 μm)** | | | | **Stage 3 (3.2-4.7 μm)** | | | |
| Con | Low | Medium | High | Con | Low | Medium | High | Con | Low | Medium | High |
| **H** | **2.52** | 2.58 | 2.57 | ***2.24*** | **2.48** | 2.21 | 2.21 | ***2.36*** | **2.66** | 2.65 | 2.64 | ***2.53*** |
| **E** | 0.87 | 0.88 | 0.93 | 0.85 | 0.9 | 0.86 | 0.86 | 0.85 | 0.9 | 0.9 | 0.85 | 0.88 |

# 第五章 结论及展望

# 参考文献[[2]](#footnote-2)

1. 第38次中国互联网络发展状况统计报告 <http://www.cnnic.net.cn/>
2. 新浪财经频道 <http://finance.sina.com.cn/roll/2016-11-12/doc-ifxxsfip4572557.shtml>
3. 刘鹏, 王超, 计算广告 互联网商业变现的市场与技术, 人民邮电出版社, 2015
4. Zhang W V, Jones R. Comparing click logs and editorial labels for training query rewriting, Proceedings of the WWW 2007 Workshop on Query Log Analysis: Social and Technological Challenges. Banff , Canada, 2007
5. Richardson M, Dominowska E, Ragno R. Predicting clicks: Estimating the click-through rate for new ads, Proceedings of the 16th International conference on World Wide Web, Banff, Canada, 2007: 521-530
6. Guo F, Liu C, Kannan A, et al. Click chain model in web search, Proceedings of the 18th International conference on World Wide Web. Madrid, Spain, 2009: 11-20
7. Jahrer M, Toscher A, Lee J Y, et al. Ensemble of collaborative filtering and feature engineered models for click through rate prediction, Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDDCup Workshop, Beijing, China, 2012
8. Shen S, Hu B, Chen W, et al. Personalized click model through collaborative filtering, Proceedings of the 5th ACM International Conference on Web Search and Data Mining, Seattle, USA. 2012: 323-332
9. Regelson M, Fain D. Predicating click-through rate using keyword clusters, Proceedings of the second workshop on Sponsored Search Auctions, Ann Arbor, USA, 2006: 9623-9628
10. Agarwal D, Border A Z, Chakrabarti D, et al. Estimating rate of rare events at multiple resolutions, Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, USA, 2007: 16-25
11. Wang X, Li W, Cui Y, et al. Click-through rate estimation for rare event in online advertising, Hua Xian-Sheng, Mei Tao, Hanjialic A eds. Online Multimedia Advertising: Techniques and Technologies. Hershey Pennsylvania, USA: IGI Global, 2010: 1-12
12. Kempe D, Mahdian M, A cascade model for externalities in sponsored search, Proceedings of the 4th International Workshop on Internet and Network Economics, Chicago, USA, 2008: 585-596
13. Guo F, Liu C, Wang Y M, Efficient multiple-click models in web search, Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 2009: 124-131
14. 腾讯大数据：CTR预估中GBDT与LR融合方案 <http://www.cbdio.com/BigData/2015-08/27/content_3750170.htm>
15. He X, Pan J, Jin O, et al. Practical lessons from predicting clicks on ads at facebook[C]. Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1-9

1. <https://github.com/guestwalk/Kaggle-2014-criteo>
2. Cox, DR. "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20, 1958: 215–242.
3. Walker SH, Duncan DB. "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54, 1967: 167–178.
4. <https://en.wikipedia.org/wiki/Logistic_regression#Logistic_function.2C_odds.2C_odds_ratio.2C_and_logit>
5. Hosmer, David W.; Lemeshow, Stanley. Applied Logistic Regression (2nd ed.). Wiley. ISBN 0-471-35632-8, 2000
6. Everitt, Brian. The Cambridge Dictionary of Statistics. Cambridge, UK New York: Cambridge University Press, 1998
7. Tjalling J. Ypma, Historical development of the Newton-Raphson method, SIAM Review 37 (4), 531–551, 1995.
8. Menard, Scott W, Applied Logistic Regression (2nd ed.). SAGE, 2002

[1] Intel Corperation, http://download.intel.com/pressroom/kits/IntelProcessorHistory.pdf.

[2] Intel Corperation, http://www.intel.com/content/dam/www/public/us/en/documents/presentation/  
revolutionary-22nm-transistor-technology-presentation.pdf.

[3] I. Žutić, J. Fabian and S. Das Sarma, Spintronics: Fundamentals and applications, Reviews of Modern Physics 76 (2), 323-410 (2004).

[4] R. Hanson, L. Kouwenhoven, J. Petta, S. Tarucha and L. Vandersypen, Spins in few-electron quantum dots, Reviews of Modern Physics 79 (4), 1217 (2007).

[5] D. Loss and D. P. DiVincenzo, Quantum computation with quantum dots,Physical Review A 57 (1), 120 (1998).

注：以上是“顺序编码制”索引文献时参考文献著录法（对应第1章示例）。各项著录信息未核准，仅为样式参考。“著者—出版年”制索引文献著录方法如下（对应第二章示例）：

段凤魁, 贺克斌, 刘咸德, 董树屏, 杨复沫. 2007. 含碳气溶胶研究进展：有机碳和元素碳. 环境工程学报, 1: 1-8.

Bond T.C.; Bergstrom R.W. 2006. Light absorption by carbonaceous particles: an investigative review. Aerosol Science and Technology, 40: 27-67.

Bond, T. C.; Streets, D. G.; Yarber, K. F.; et al. 2004. A technology-based global inventory of black and organic carbon emissions from combustion. Journal of Geophysical Research, 109, D14203.

Cao, G. L.; Zhang, X. Y.; Zheng, F. C. 2006. Inventory of black carbon and organic carbon emissions from China. Atmospheric Environment, 40: 6516-6527.

Klimont, Z.; et al. 2009. Projections of SO2, NOx and carbonaceous aerosols emissions in Asia. Tellus, 61B, 602-617.

Lu, Z.; Zhang, Q.; Streets, D. G. 2011. Sulfur dioxide and primary carbonaceous aerosol emissions in China and India, 1996-2010. Atmospheric Chemistry and Physics, 11, 9839-9864.

Penner, J. E.; Eddleman, H.; Novakov, T. 1993. Towards the development of a global inventory for black carbon emissions. Atmospheric Environment, 27 (A): 1277-1295.

Streets, D. G.; Bond, T. C.; Carmichael, G. R.; et al. 2003. An inventory of gaseous and primary aerosol emissions in Asia in the year 2000. Journal of Geophysical Research, 108, 8809.

Streets, D. G.; Bond, T. C.; Lee, T.; Jang, C. 2004. On the future of carbonaceous aerosol emissions. Journal of Geophysical Research, 109, D24212, doi:10.1029/2004JD004902.

Streets, D.G.; Shalini, G.; Waldhoff, S.T.; et al. 2001. Michae Black carbon emissions in China. Atmospheric Environment, 35, 4281- 4296.

Zhang, Q.; Streets, D. G.; Carmichael, G. R.; et al. 2009. Asian emissions in 2006 for the NASA INTEX-B mission. Atmospheric Chemistry Physics, 9, 4081- 4139.

# 附录A 附录示例

# 致谢

本论文是在xx老师的悉心指导下完成的。xx老师作为一名优秀的、经验丰富的教师，具有丰富的xx知识和xx经验，在整个论文实验和论文写作过程中，对我进行了耐心的指导和帮助，提出严格要求，引导我不断开阔思路，为我答疑解惑，鼓励我大胆创新，使我在这一段宝贵的时光中，既增长了知识、开阔了视野、锻炼了心态，又培养了良好的实验习惯和科研精神。在此，我向我的指导老师表示最诚挚的谢意！

……

(仅为网络示例，可根据论文实际进行撰写，使用时把模板示例内容尽皆删除即可)

# 北京大学学位论文原创性声明和使用授权说明

**原创性声明**

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

**学位论文使用授权说明**

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

* 按照学校要求提交学位论文的印刷本和电子版本；
* 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
* 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
* 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名： 导师签名：

日期： 年 月 日

1. 图标题在图下方，表标题在表上方。图表序号分章设置，如图3.15表示第三章第15幅图。 [↑](#footnote-ref-1)
2. 全文参考文献索引方式只能选用“顺序编码制”或“著者—出版年制”其中之一，文献列表也应选择相对应的著录方法，此处作为示例列举了两种方式，实际撰写论文时不得混用。 [↑](#footnote-ref-2)