电子科技大学信息与软件工程学院

综合设计 I 课题任务书

课题名称	基于品友（iPinYou）广告数据日志的点击率预测方法				
课程名称	综合课程设计	**专业方向**	XXXXX	**选课年级**	XXXXX
指导教师		**教师电话**		**教师邮箱**	

主要任务（请注意内容与工作量要求并覆盖毕业要求相关指标点）：

互联网广告逐渐成为目前网站的主要收入来源之一，通过广告定向投放技术可以极大的提升广告营销的收益。定向广告投放的一个重要技术是分析将一个广告投放给一个用户后，该广告能够被用户点击、应用被安装，或者商品被购买的概率，这就是所谓的点击率预测技术。本课题提供“品友”的电商广告投放日志，要求学生设计实现一个点击率预测算法，其中的关键技术包括：数据清理及训练集和测试集划分，根据数据的特征，选择合适的学习模型，例如逻辑回归模型，通过训练集数据学习模型参数，并利用测试集数据对点击率预测模型的性能进行测试。指导老师将提供具体的技术方案指导。

预期成果或目标：

设计实现一个完整的点击率预测软件，包括数据清洗、训练集和测试集的自动划分、学习模型的选择，评价指标的展示等。

软件需求分析和设计文档。

涉及知识点：

1、 大数据的清理和训练集的划分方法；
2、 简单的机器学习模型，例如逻辑回归模型；
3、 点击率预测模型的主要特征分析及测试结果评价方法；
4、 需求分析的基本方法。

指导教师签名：_____

年 月 日

备注：此任务书必须双面打印。

理论知识部分介绍

1 点击率(转化率)预测模型

作为目前在线广告最热门的研究领域，点击率预测的研究吸引了大量来自工业界和学术界的研究者。本节首先给出了点击率预测问题的形式化描述，然后对已有的预测模型进行了梳理，并将其分为基于传统机器学习模型和基于深度学习模型两大类，最后对比了其中典型方案的特点，分析了点击率预测仍然存在的难点和需要解决的关键问题。

1.1 问题的形式化描述

点击率预测问题是一个典型的回归问题，如图所示，学习系统首先基于给定训练数据集，$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，构建预测模型，$Y=f(X)$，然后预测系统根据预测模型预测每个符合投放规则的广告对于新到达的广告展示机会 x_{N+1} 的点击率 y_{N+1}。在预测模型的学习过程中，需要确定最优化策略，即通过最小化损失函数 $L(\cdot)$ 来学习模型参数，如公式(1)所示，对于不同的预测模型，可以设计不同的损失函数，涉及到的求解算法包括牛顿法、拟牛顿法、梯度下降法、随机梯度下降法等。

$$\min_{\boldsymbol{\theta}} L(f(x_i),y_i) \quad (1)$$

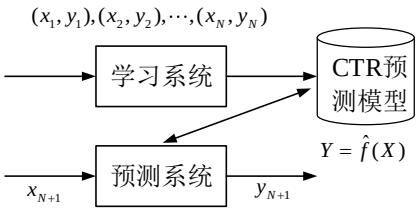

图 1 点击率预测问题

1.2 基于传统模型的解决方案

点击率预测问题最早的解决方案是利用逻辑回归(Logistic Regression, LR)来学习点击率预测模型。LR 中，定义点击率预测公式如(2)所示，其中 $\boldsymbol{x}$ 表示广告展示机会的特征向量，y_i 表示广告是否被点击的真值，$y_i=1$ 表示发生过点击，$y_i=-1$ 表示没有发生点击，$\mathbf{w}\in\mathbf{R}^{n+1}$ 表示模型中参数向量，n 表示特征向量的维度，因此该次广告展示机会未发生点击的概率公式如(3)所示，因此可将(2)和(3)合并为(4)。

$$P(y_i=1|\boldsymbol{x}_i)=\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}} \quad (2)$$

$$P(y_i=-1|\boldsymbol{x}_i)=1-\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}=\frac{1}{1+e^{\boldsymbol{w}^T\boldsymbol{x}_i}} \quad (3)$$

$$P(y_i=\pm1|\boldsymbol{x}_i)=\frac{1}{1+e^{-y_i(\boldsymbol{w}^T\boldsymbol{x}_i)}} \quad (4)$$

在逻辑回归模型中，损失函数通常使用对数损失函数(即负的对数似然函数)，如公式(5)所示，这里 m 表示训练集中的样本数，为了防止模型过拟合，通常在损失函数中加入 L2 正则化项，如(6)所示，λ 表示正则化参数，因此可以通过最小化损失函数来学习模型参数 $\mathbf{w}$。逻辑回归模型的参数求解算法非常多，除了牛顿法、拟牛顿法，还包括随机梯度下降法、坐标下降法等，也可以使用 FOBOS、RAD、FTRL、FTRL-Proximal 等在线算法求解。

$$L(\boldsymbol{w})=-\sum_{i=1}^{m}\log\left(\frac{1}{1+e^{-y_i(\boldsymbol{w}^T\boldsymbol{x}_i)}}\right)=\sum_{i=1}^{m}\log(1+\exp(-y_i(\boldsymbol{w}^T\boldsymbol{x}_i))) \quad (5)$$

$$\min_{\boldsymbol{w}}\{\sum_{i=1}^{m}\log(1+\exp(-y_i(f(\boldsymbol{w},\boldsymbol{x}_i)))+\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\} \quad \text{其中 } f(\boldsymbol{w},\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i \quad (6)$$

逻辑回归模型是一种广义线性模型，非常容易实现大规模实时并行处理，因此在工业界获得了广泛应用，但是线性模型的学习能力有限，不能捕获特征间的高阶特征(非线性信息)，为此文献[2]提出了 Poly2 模型，不仅考虑了单个特征携带的信息，而且考虑二阶的特征组合(feature conjunction)携带的信息，因此其 $f(\boldsymbol{w},\boldsymbol{x}_i)$ 改写为公式(7)所示，不仅考虑了一阶的特征对应的权重，而且考虑了二阶的特征组合对应的权重，

这里 $w_h(k,l)$ 表示样本的第 k 个特征和第 l 个特征组合对应的权重，x_i^k 表示样本 $\boldsymbol{x}_i$ 的第 k 个特征的值，因此 Poly2 模型需要为每个特征组合学习一个权重。Poly2 的问题在于，当样本的特征维度非常大时，二阶组合特征的权重计算复杂度将变得非常大，为 $O(\bar{n}^2)$，这里 $\bar{n}$ 表示样本中非 0 元素的平均值；此外，如果某个特征组合在训练集中没有出现，那么对应项的权重将不能得到充分学习，从而降低预测的准确性。

$$f(\boldsymbol{w},\boldsymbol{w}_h,\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(w_h(k,l)\cdot x_i^k\cdot x_i^l) \quad (7)$$

鉴于 Poly2 的问题，文献[3]提出了基于因子分解机(Factorization Machine, FM)的 CTR 预测模型，其基本思想是将(6)中的 $f(\boldsymbol{w},\boldsymbol{x}_i)$ 改写为公式(8)的形式，这里 $\boldsymbol{v}_k$ 和 $\boldsymbol{v}_l$ 分别表示特征 k 和特征 l 的维度为 D 的隐含向量。因此，在 FM 中每个特征用一个 D 维的隐含向量表示，从而使二阶的特征组合的权重分解为两个隐含向量的点积，将计算复杂度降低为 $O(D\cdot\bar{n})$，同时即使训练集中没有出现某个特征组合，由于两个特征的隐含向量是分开学习的，因此不会影响预测准确性。FM 的参数求解算法可以采用

$$f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(<\boldsymbol{v}_k\cdot\boldsymbol{v}_l>\cdot x_i^k\cdot x_i^l) \text{ 其中 } <\boldsymbol{v}_k\cdot\boldsymbol{v}_l>=\sum_{d=1}^{D}v_k^d v_l^d \quad (8)$$

FM 的缺陷在于每个特征都只学习一个唯一的隐含向量，在与其它不同特征进行组合时，同一个特征产生的影响力都是相同的，而事实上当与不同领域的特征组合时，特征可能表现出不同的隐含特征分布。例如有如下点击记录“当一个女性(Female)用户在发布媒体 Vogue 上浏览页面时，对投放的 Gucci 广告，发生了点击行为”。在 FM 中，学习二阶的特征组合时，只需要学习三个隐含特征向量，无论与特征 Gucci 还是与特征 Female 进行组合时，特征 Vogue 的隐含向量都是相同的，而事实上，更希望在与不同领域的特征进行组合时，考虑差异化的隐含特征向量。为此，文献[4]在 FM 模型的基础上引入了“领域”的概念，提出了 FFM(Field-aware Factorization Machines)模型，其基本思想是将特征分割为若干领域，例如在将特征 Gucci 划分为 Advertiser 领域，特征 Female 划分为 Gender 领域，特征 Vogue 划分为 Publisher 领域，每个特征将针对不同的领域学习不同的隐含向量，因此公式(8)中的 $f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)$ 改写为公式(9)的形式，这里 $\boldsymbol{v}_{k,f_l}$ 和 $\boldsymbol{v}_{l,f_k}$ 分别表示特征 k 在特征 l 所属的 f_l 领域以及特征 l 在特征 k 所属的 f_k 领域的隐含向量。FFM 的参数求解算法采用的是随机梯度下降，以及改进的 AdaGrad 算法。

$$f(\boldsymbol{w},\boldsymbol{v},\boldsymbol{x}_i)=\boldsymbol{w}^T\boldsymbol{x}_i+\sum_{k=1}^{n}\sum_{l=k+1}^{n}(<\boldsymbol{v}_{k,f_l}\cdot\boldsymbol{v}_{l,f_k}>\cdot x_i^k\cdot x_i^l) \quad (9)$$

Poly2、FM、FFM 都是在 LR 基础上增加对二阶的特征组合的权重自动学习的模型。除此之外，Facebook 的研究人员提出了另一种筛选特征和特征组合的方式，称为 GBDT+LR 方案[5，ADKDD 2014]，该方案利用 GBDT(Gradient Boost Decision Tree)来帮助筛选有区分度的特征和特征组合，作为 LR 模型的输入，从而增强 LR 的非线性学习能力。GBDT 是一种非线性模型，它基于集成学习中 boosting 的思想，每次迭代都在减少残差的梯度方向新建立一颗回归树，每个叶子结点作为一个取值为 0/1 的输入特征，因此新特征向量的长度等于 GBDT 模型里所有回归树包含的叶子结点总数。在图 2 的例子中，有两颗回归树，五个叶子结点，因此输入到 LR 的新的特征向量的维度为 5。当样本 $\mathbf{x}$ 输入 GBDT，假设它在左边的树中落入第一个结点，在右边的树中落入第二个结点，则新的特征向量编码为[1,0,0,0,1]，再将其输入到 LR 中进行模型学习。除了 GBDT+LR，2014 年 Kaggle CTR 的冠军队使用了 GBDT+FFM 的融合方案，也取得了非常好的预测效果。

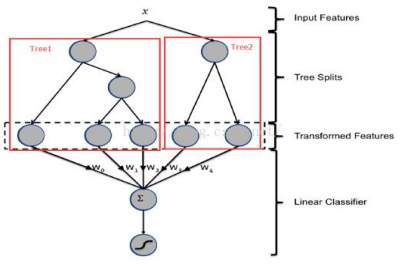

图 2 GBDT 模型的示意图

2 点击率(转化率)预测模型的评价指标

精度和**召回率**以及其 **F1 度量**是考察预测精度的基本指标，它能够考虑到广告历史日志中点击事件稀少导致的类不平衡问题。而这些指标是针对分类的，而广告的点击率预测中常用的是回归，回归转化为分类依据是判定阈值，通常为 0.5。而考虑到类不平衡问题后，若没对预测数值进行校准，其最佳的阈值是很难判断的。阈值的设定不同会导致同一次回归分析得出不同的精度召回率以及 F1 度量，这显然给结果评估造成了干扰。

ROC 曲线描述了灵敏度随着特异性的关系，**AUC 是曲线与坐标轴的面积**，AUC 对于广告点击率预测是一个优良的指标。AUC 求得预测器在阈值可能的取值范围内的累积值，因此避免了阈值的选取对结果的干扰，可适用于任何连续值回归模型。AUC 的取值范围为[0,1]，0.5 是随机猜测的取值，而完美的模型为 1，因此很容易判断模型的好坏，同时适用于在线和离线两种模型的建立。另外 AUC 与精度召回率之间存在相关性，再考虑到之前的阈值干扰问题，因而舍弃使用精度召回率指标。

为评价每次预测的误差，可以计算预测值与实际值之间的损失。损失函数表示为：$loss(y_i, y_i')$，y_i为实际值，y_i'为预测值。对于使用逻辑回归预测的点击率问题，应该使用对数损失（Logarithmic Loss）函数：

$$logloss(y_i, y_i') = -y_i \log y_i' - (1 - y_i)\log(1 - y_i')$$

对数损失也是逻辑回归噪声的最大似然函数的负对数。且当实际值为 1 时，预测趋近于 1，则损失趋近于 0，若预测错误则趋近于 0，则遭受趋向于无穷大的损失，当实际值为 0 时与之同理。且在逻辑回归中使用对数损失函数作为目标函数能保证求解函数的凸性，因而具有全局最优解。

参考文献

[1] O. Chapelle, E. Manavoglu, and R. Rosales, “Simple and scalable response prediction for display advertising,” *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 4,pp. 61:1–61:34, Dec. 2014.

[2] Y.-W. Chang, C.-J. Hsieh, K.-W. Chang, M. Ringgaard, and C.-J. Lin, Training and testing low-degree polynomial data mappings via linear SVM, Journal of Machine Learning Research, vol. 11, pp. 1471-1490, 2010.

[3] R. Steffen, “Factorization machines with libFM,” *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 57:1–57:22, May 2012.

[3] Ta A P. Factorization machines with follow-the-regularized-leader for CTR prediction in display advertising[C]// IEEE International Conference on Big Data. IEEE, 2015:2889-2891.

[4] Juan Y, Zhuang Y, Chin W S, et al. Field-aware Factorization Machines for CTR Prediction[C]// ACM Conference on Recommender Systems. ACM, 2016:43-50

[5] Bowers S, Bowers S, Bowers S, et al. Practical Lessons from Predicting Clicks on Ads at Facebook[C] // Eighth International Workshop on Data Mining for Online Advertising. ACM, 2014:1-9.