李　萍

目标岗位：**算法工程师**

lipingict@gmail.com · (+86) 130-200-48987 · http://www.lipingict.com

教育背景

中国科学院计算技术研究所，计算机，免试保送，北京　　2014 – 2017

吉林大学，计算机科学与技术，1%，长春，吉林　　2010 – 2014

自我评价

踏实认真，有团队协作的能力和沟通协调能力。有扎实的编程功底以及较丰富的大数据处理经验。熟悉大规模数据挖掘和机器学习，能熟练使用 Hadoop, Spark 等大数据平台。

技能

- 编程语言: Python = C++ > Scala > Shell
- 语言: 英语 - 熟练 (六级) , 日语 (JLPT N2)

实习/项目经历

百度凤巢　　2015 年 5 月 – 2015 年 8 月

个人职责 数据挖掘工程师

项目概述 在 FCR_Model 实习期间主要负责模型评估、特征挑选以及特征调研.

- *特征分析工具*: 凤巢没有统一的特征分析工具，特征的筛选是自由组合，筛选效率较低，基于凤巢特征抽取系统 Adfea 和模型训练系统 Platform 开发一个特征评价分析用来分析特征从而助力特征挑选.
- *sug 模型评测*: 线上使用的 sug cpm 预估模型效率较低，使用时间衰减策略来预测 suggestion query 的 CPM 可以极大地提高效率. 在百度半年的 sug 数据集上评估该策略在不同衰减因子以及时间窗口上的效果.
- *特征调研*: 调研 Kaggle CTR prediction 比赛中优胜选手使用的特征、连续特征值的处理以及使用的模型，并写出详细的调查报告.

图形化大数据机器学习平台 BDA Studio　　2015 年 8 月 – 至今

个人职责 算法工程师

项目概述 图形化机器学习平台由 BDA Studio 以及 BDALib 构成，分别是大数据机器学习库和可拖拽大数据机器学习平台 BDA Studio.

- *图算法*: 开发三个图算法 (Pagerank, ICmodel, KShell), 相比于 graphx 原生算法相比可收敛，具有可扩展性、速度快等特点，可以支持上 10 亿规模顶点的图数据挖掘.
- *推荐算法*: 实现Factorization Machine 和NMF算法的 local 版本、spark shared 版本、spark graphx 版本, 在 movie-lens 数据集上这些算法在测试数据集上 RMSE 都在 0.80 左右.
- *ETL*: 完成 BDA Studio 的 ETL 功能，支持 Mysql, Hive 等异源数据的导入.

天翼大数据算法应用大赛　　2015.12 – 2016.03

- *概述*: 使用前 7 周用户每天点击 10 个视频网站的统计数据，预测用户第八周每天点击视频网站的数量。
- *职责*: 特征调研，特征抽取，特征评价
- *成绩*: 比赛历时两个月，第一赛季排行榜第九名, 第二赛季在 **1111 名**队伍中斩获第一。
- *技术*: Spark, Xgboost, GBDT

论文及获奖情况

分布式算法实现比较研究：数据分布与模型分布, CCIR 在投　　2016 年 6 月

Ease the Process of Machine Learning with Dataflow, CIKM　　2016 年 5 月

Sesssion Segmentation Method Based on COBWEB, EI 检索　　2012 年 6 月

连续三年国家奖学金, 吉林大学　　2011 年 – 2013 年