大作业与未来方向

续本达

清华大学 工程物理系

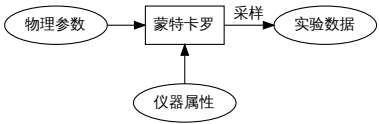
2023-08-03 清华

广义线性回归

1. 关系代数让回归分析变得极其直观，让我们专注于问题的本质
2. 模型的选择极其重要，应当使用客观标准
   - AIC 是最简单直接的客观标准
   - 找到“最佳模型”的过程充满曲折，是实验“研究”的主体过程
3. 广义线性回归，把误差分布从高斯替换为其它指数族分布
   - 把连接函数从恒等替换为非线性函数
   - 几乎可以解决所有日常工作中的非线性问题，惊喜！

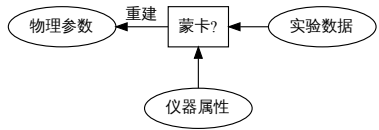
大作业安排

正向第一阶段



- 2023-07-21 – 2023-08-10
- 模拟实验测量

逆向第二阶段



- 2023-08-03 – 2023-08-24
- 分析实验数据
- 测量物理模型参数
- 发现物理规律
- 黑盒分数按排名
  - 在不同大作业之间归一化

实验测量的分析

输入 （模拟的）实验测量原始数据
- 如果不是模拟的，则无法评分
- 但是助教会尽可能把它做得和真的一样

输出 物理对象的信息

采分 与助教手中的模拟输入相比

第二阶段分组

- 同学们先联络好，组队信息在网络学堂提交
  - 到 `https://physics-data.meow.plus` 注册账号，一并提交
  - 可能需要重新组队
- 每队至多三人
  - 单人队：大作业得分 $\times 1.03$
  - 三人队：每人大作业得分 $=$ 队伍得分 $\times 0.95$
  - 不同队伍间请勿直接交换代码
- 如果大作业结果含有学术突破，总评保送 A+。

实验测量的分析

输入 （模拟的）实验测量原始数据
- 如果不是模拟的，则无法评分
- 但是助教会尽可能把它做得和真的一样

输出 物理对象的信息

采分 与助教手中的模拟输入相比

第二阶段分组

- 同学们先联络好，组队信息在网络学堂提交
  - 到 `https://physics-data.meow.plus` 注册账号，一并提交
  - 可能需要重新组队
- 每队至多三人
  - 单人队：大作业得分 $\times 1.03$
  - 三人队：每人大作业得分 $=$ 队伍得分 $\times 0.95$
  - 不同队伍间请勿直接交换代码
- 如果大作业结果含有学术突破，总评保送 A+。

https://physics-data.meow.plus 的使用

1. 注册账号：注意邮箱的“垃圾箱”等，如果收到不到注册邮件尝试更换邮箱
   - 遇到困难开 issue 提问
2. 找到大作业的位置
   注意平台上还有很多其它的作业，不要走错了
   - https://physics-data.meow.plus/challenges/pd2023-gamma
   - https://physics-data.meow.plus/challenges/pd2023-muon
   - https://physics-data.meow.plus/challenges/pd2023-spectroscopy
3. 前往 "submissions" 选项卡，点击 "CREATE SUBMISSION"
   - 上传你的解答文件。

大作业的可复现性

复现 原则的要求

提交到 `https://physics-data.meow.plus` 的结果必须可复现，否则无效。

思路和要点

- 组队完成后，将获得
  `https://git.tsinghua.edu.cn/physics-data/2023/project_2`
  之下的仓库一份，使用 GNU Make 构建整个分析流程，连同报告、程序整理到仓库中。
  - 注意小组分工中 Git 使用的规范
  - 善于使用 Git `branch`, Gitlab `merge request` 等团队协作功能
- 把流程系统化成输入、输出与过程三要素。
  - 而向数据编程，data-driven programming
- 系统表达输入数据、输出数据和中间结果的依赖关系，
  - 成为“可执行的说明文档”

具体说明

gamma 刘学伟

- 粒子物理方向：台山中微子实验的成像原理探索
- Ghost Hunter 2023 课赛结合

muontagging 刘明昊

- 宇宙射线缪子的测量与屏蔽

spectroscopy 吴致颉

黑客技能

科学数据处理的原则

黑客的审美： 复现 透明 一次 最佳工具

版本控制

Git 与队友分工协作，与明天的自己协作
Git 是“搬砖工地安全帽”，无头盔禁止上岗

关系代数

数据表示成关系，数据的操作表示成关系代数运算

数据格式

透明 CSV, HDF5, JSON, 数据库 SQL

黑客技能（二）

数据流水线

GNU Make 管理数据的依赖与转换，实现错误恢复和并行计算
实现数据层次的 Python/R/Bash/Scheme/SQL 多语言融合

正则表达式

描述字符串的微型语言，数学模型

命令环境

POSIX 环境中强大的小工具组合，开发与使用相融合

计算语言

Python 语法友好，工具丰富，统领 C/C++/Fortran/R/SQL 库

永远留在古老的计算环境

- SuperK 质子衰变和中微子实验，XMASS 暗物质实验
- 问题
  1. 数据处理技术发展停滞，违反“最佳工具”原则
  2. 新成员需要花精力学习旧技术，人力浪费
- 解决方案：test-driven development

永远留在古老的计算环境

- SuperK 质子衰变和中微子实验，XMASS 暗物质实验
- 问题
  1. 数据处理技术发展停滞，违反“最佳工具”原则
  2. 新成员需要花精力学习旧技术，人力浪费
- 解决方案：test-driven development

自制 Python 驱动的批量处理流水线

- 症状问题
  1. 使用 Python 调用大量 shell 命令，程序可读性差，违反“最佳工具”原则

```
for name in args.name:
  if name not in productions:
    print('Unknown production: ' + name)
    continue
  for script in productions[name]:
    parg = arguments + ' --name %s ' % name
    print('python %s %s' % (script, parg))
    os.system('python %s %s' % (script, parg))
```

  2. 流水线验证 flag 文件是否存在来确定是否成功执行，误判多，难以 debug
- 解决方案：使用 GNU Make 构建流水线
  - make 默认使用 /bin/sh 执行命令。 SHELL=/bin/sh
  - 把 SHELL 换成提交任务给超级计算机集群的脚本

```
SHELL=lsf
export MAKE_TARGET=$@
export MAKE_SOURCE=$^
```

自制 Python 驱动的批量处理流水线

- 症状问题
  1. 使用 Python 调用大量 shell 命令，程序可读性差，违反“最佳工具”原则

```
for name in args.name:
  if name not in productions:
    print('Unknown production: ' + name)
    continue
  for script in productions[name]:
    parg = arguments + ' --name %s ' % name
    print('python %s %s' % (script, parg))
    os.system('python %s %s' % (script, parg))
```

  2. 流水线验证 flag 文件是否存在来确定是否成功执行，误判多，难以 debug
- 解决方案：使用 GNU Make 构建流水线
  - make 默认使用 /bin/sh 执行命令。 SHELL=/bin/sh
  - 把 SHELL 换成提交任务给超级计算机集群的脚本

```
SHELL=lsf
export MAKE_TARGET=$@
export MAKE_SOURCE=$^
```

make 对接集群的脚本

```
#!/home/jinping/gentoo/bin/bash -e
# Platform LSF wrapper to be used as GNU Make shell.

# GNU Make convention for the first argument.
[[ ${1} = '-c' ]] && shift

for j in ${MAKE_SOURCE}; do
    [[ -z $(bjobs -J ${j}) ]] && continue
    DEP+=" && done(${j})" # 把 make 中的依赖关系传递给调度系统
done

cat << EOF > ${MAKE_TARGET}.sh
#!/home/jinping/gentoo/bin/bash
$@
EOF

chmod +x ${MAKE_TARGET}.sh

bsub -q normal"${DEP}" -J ${MAKE_TARGET} -o ${MAKE_TARGET}.log ${MAKE_TARGET}.sh
```

建设状况

- https://git.tsinghua.edu.cn/physics-data/lecture

```
head -q -n1 lecture/l*.org
```

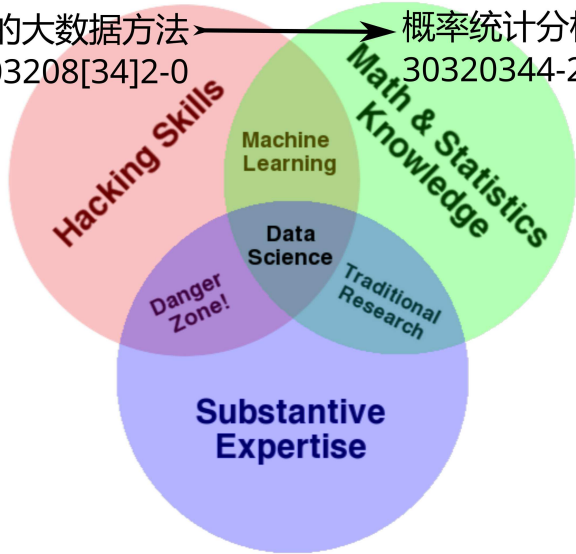
```
#+TITLE: 第一讲 实验物理的大数据方法总论 DONE
#+TITLE: 第二讲 Python基础 DONE
#+TITLE: 第三讲 复合类型与函数 DONE
#+TITLE: 第四讲 Python 模块 DONE
#+TITLE: 第五讲 数组 DONE
#+Title: 第六讲 数据格式 DONE
#+Title: 第七讲 数据绘图 TODO
#+Title: 第八讲 蒙特卡罗方法与大作业 TODO
#+Title: 第九讲 GNU 命令行 TODO
#+Title: 第十讲 GNU Make 数据生产线 TODO
#+Title: 第十一讲 正则表达式 TODO
#+Title: 第十二讲 bash 脚本 TODO
#+Title: 第十三讲 关系代数 TODO
#+Title: 第十四讲 DataFrame 表格数据结构 TODO
#+Title: 第十五讲 关系代数与回归分析 TODO
#+Title: 第十六讲 现实案例与未来方向 TODO
```

- 给讲义仓库提 issue 和 merge request ，可获得伍至捌分。
- 一次原则的应用：把讲议与课件写到一起
  - 把课堂口述内容通过 speech-to-text 引擎合成文字

数据时代的物理技能

实验物理的大数据方法
夏 403208[34]2-0

概率统计分析及量测技术
30320344-2 秋



概率是逻辑的扩展 – Cox 定理

- Laplace: probability theory is nothing but common sense reduced to calculations.
- R. T. Cox, E. T. Jaynes，两位对统计学有重大贡献的物理学家
- Logical interpretation of probability
  1. Divisibility and comparability – The plausibility of a proposition is a real number and is dependent on information we have related to the proposition.
  2. Common sense – Plausibilities should vary sensibly with the assessment of plausibilities in the model.
  3. Consistency – If the plausibility of a proposition can be derived in many ways, all the results must be equal.
- 概率论的唯一性：Any system for plausible reasoning that satisfies certain qualitative requirements intended to ensure consistency with classical deductive logic and correspondence with commonsense reasoning is isomorphic to probability theory.

  参考：Van Horn, K.S., 2003. Constructing a logic of plausible inference: a guide to Cox’s theorem. International Journal of Approximate Reasoning 34, 3 – 24.

以本课程为起点

- 函数式编程：一切都是函数
  - 无状态，从而容易从错误中恢复
- MapReduce：分布式大数据系统的开始
  - 在之上建立了关系代数：
  - Apache Hive，SparkQL，etc.
- 机器学习：与回归分析有同样的程序接口
  - 算法上更一般，需要更多的“调参”

竞赛

- gamma 大作业 $\rightarrow$ Ghost Hunter 2023 中微子数据分析排位赛
  - 概率统计分析及量测技术课赛结合

技术问题：TUNA 协会

- 清华大学学生开源软件与网络技术协会
- TUNA 主页 `https://tuna.moe/`
- TUNA 技术群，黑客（广义）技术问题探讨
- 为课程提供了
  - Gentoo 的镜像支持
    `https://mirrors.tuna.tsinghua.edu.cn/gentoo/`
  - Debian 的镜像支持
    `https://mirrors.tuna.tsinghua.edu.cn/debian/`
- 为大家提供了
  - 清华大学学位论文 LaTeX 模版 `thuthesis`

暑期实习

如果你觉得配环境是一个非常快乐的事，并且经常帮助同学配环境。
计算环境的可复现性是而容易被忽视，但也最重要的环节。

- Google Summer of Code：Google
  - Gentoo、Debian 操作系统相关项目
- 开源之夏活动：中科院软件所、华为、TUNA
  - Gentoo、Debian 操作系统相关项目
- 脱去资本枷锁，为人类数字化的自由而战 (bs，有点中二

暑期实习

如果你觉得配环境是一个非常快乐的事，并且经常帮助同学配环境。
计算环境的可复现性是而容易被忽视，但也最重要的环节。

- Google Summer of Code：Google
  - Gentoo、Debian 操作系统相关项目
- 开源之夏活动：中科院软件所、华为、TUNA
  - Gentoo、Debian 操作系统相关项目
- 脱去资本枷锁，为人类数字化的自由而战 (bs，有点中二

暑期实习

如果你觉得配环境是一个非常快乐的事，并且经常帮助同学配环境。
计算环境的可复现性是而容易被忽视，但也最重要的环节。

- Google Summer of Code：Google
  - Gentoo、Debian 操作系统相关项目
- 开源之夏活动：中科院软件所、华为、TUNA
  - Gentoo、Debian 操作系统相关项目
- 脱去资本枷锁，为人类数字化的自由而战 (bs，有点中二

SRT 与大创：Scheme

如果你认同实验物理与形式逻辑是文明的两大支柱，并喜欢课程的内容，可以考虑与我继续探索：

- 分析力学的 Scheme 描述：
  - Structure and Interpretation of Classical Mechanics, MIT 本科课程
  - 分析力学完成后，计划拓展至量子力学和电动力学
- 自动推理机：
  - 利用第一阶段大作业成果，直接生成第二阶段大作业的解答程序
  - 语言为 Anglican ，为 Clojure 语言的一种
  - Clojure 语言是 Scheme 语言在 Java 平台上的实现

SRT 与大创：Scheme

如果你认同实验物理与形式逻辑是文明的两大支柱，并喜欢课程的内容，可以考虑与我继续探索：

- 分析力学的 Scheme 描述：
  - Structure and Interpretation of Classical Mechanics, MIT 本科课程
  - 分析力学完成后，计划拓展至量子力学和电动力学
- 自动推理机：
  - 利用第一阶段大作业成果，直接生成第二阶段大作业的解答程序
  - 语言为 Anglican ，为 Clojure 语言的一种
  - Clojure 语言是 Scheme 语言在 Java 平台上的实现

SRT 与大创：Scheme

如果你认同实验物理与形式逻辑是文明的两大支柱，并喜欢课程的内容，可以考虑与我继续探索：

- 分析力学的 Scheme 描述：
  - Structure and Interpretation of Classical Mechanics, MIT 本科课程
  - 分析力学完成后，计划拓展至量子力学和电动力学
- 自动推理机：
  - 利用第一阶段大作业成果，直接生成第二阶段大作业的解答程序
  - 语言为 Anglican ，为 Clojure 语言的一种
  - Clojure 语言是 Scheme 语言在 Java 平台上的实现

致谢

顾问 杨鲁懿教授、高飞教授、陈晟祺、陈嘉杰

助教 王宇逸、刘晓义、刘学伟、刘明昊、盘笛、徐闯、陶嘉燊、孙迅、吴致颉、刘逸祺

小助教 李卓航、温欣洋、胡楚坤、陈诗洋、许威、徐一恺、胡宇阳、卢一鸣、宋明卓

科协 物理系科协环境配置培训，担任助教；工物系科协提供算力支持；计算系科协提供 saiblo 平台，第二阶段大作业平台，担任助教。

克服困难上课的同学们 谢谢！

致谢

顾问 杨鲁懿教授、高飞教授、陈晟祺、陈嘉杰

助教 王宇逸、刘晓义、刘学伟、刘明昊、盘笛、徐闯、陶嘉燊、孙迅、吴致颉、刘逸祺

小助教 李卓航、温欣洋、胡楚坤、陈诗洋、许威、徐一恺、胡宇阳、卢一鸣、宋明卓

科协 物理系科协环境配置培训，担任助教；工物系科协提供算力支持；计算系科协提供 saiblo 平台，第二阶段大作业平台，担任助教。

克服困难上课的同学们 谢谢！

致谢

顾问 杨鲁懿教授、高飞教授、陈晟祺、陈嘉杰

助教 王宇逸、刘晓义、刘学伟、刘明昊、盘笛、徐闯、陶嘉燊、孙迅、吴致颉、刘逸祺

小助教 李卓航、温欣洋、胡楚坤、陈诗洋、许威、徐一恺、胡宇阳、卢一鸣、宋明卓

科协 物理系科协环境配置培训，担任助教；工物系科协提供算力支持；计算系科协提供 saiblo 平台，第二阶段大作业平台，担任助教。

克服困难上课的同学们 谢谢！

致谢

顾问 杨鲁懿教授、高飞教授、陈晟祺、陈嘉杰

助教 王宇逸、刘晓义、刘学伟、刘明昊、盘笛、徐闯、陶嘉燊、孙迅、吴致颉、刘逸祺

小助教 李卓航、温欣洋、胡楚坤、陈诗洋、许威、徐一恺、胡宇阳、卢一鸣、宋明卓

科协 物理系科协环境配置培训，担任助教；工物系科协提供算力支持；计算系科协提供 saiblo 平台，第二阶段大作业平台，担任助教。

克服困难上课的同学们 谢谢！

致谢

顾问 杨鲁懿教授、高飞教授、陈晟祺、陈嘉杰

助教 王宇逸、刘晓义、刘学伟、刘明昊、盘笛、徐闯、陶嘉燊、孙迅、吴致颉、刘逸祺

小助教 李卓航、温欣洋、胡楚坤、陈诗洋、许威、徐一恺、胡宇阳、卢一鸣、宋明卓

科协 物理系科协环境配置培训，担任助教；工物系科协提供算力支持；计算系科协提供 saiblo 平台，第二阶段大作业平台，担任助教。

克服困难上课的同学们 谢谢！