CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

CDA 数据分析师 CERTIFIED DATA ANALYST

IT大咖说 知识分享平台

我的转行之路

ADD YOUR POWERPOINT TITLE HERE

演讲人：林术贺

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT

北京 中国大饭店 2017.07

自我介绍

姓名：林术贺

职业：数据分析师

就职于：ims新媒体商业集团

如果你过的不好，你不得不拼命改变现状

计算机专业

多职业经历

工作不稳定

经常失业

无一技傍身

人生已无别出路，坚定信念不回头

→CDA二期学员

一个转行者是如何学习的数据分析的

- 转行者应该怎样夯实基础，
- 转行者如何学习数学建模，
- 如何把知识变成能力

• 我的学习路程-几组数据

- CDA的SAS教程，累计观看6遍以上
- 读书量：累计读书 近40本
- 公开课：累计观看2遍以上
- 前期以理论知识为主，后期以机器学习和社交网络分析相关的实操书籍为主并兼顾复习旧内容

统计分析的定义

统计学习：以数据为研究对象，基于数据 构建概率统计模型并运用模型对数据进行预测与分析的一门学科， 是概率论，统计学，计算机等多个领域的交叉学科，扎实的统计学基础是做好数据分析 的前提；

数据分析师至少应该掌握那些模型

客户关系管理（CRM）的分析模型：RFM

T检验，方差分析

主成份法，因子分析

相关，卡方，对应分析

线性回归，逻辑回归

决策树，boosting，随机森林，SVM

一个数据分析师应该掌握那些模型

K-means聚类,系统聚类

关联规则

协同过滤

朴素贝叶斯

分析师至少要学习那些课程-基础篇

- 统计学习方法
- 统计学
- 统计学：从数据到结论
- 应用多元统计分析
- 概率论基础（第3版）
- 统计思维：程序员数学之概率统计（第2版）
- 应用多元统计分析（第二版）
- 统计学看穿一切数字的统计学
- 可汗公开课（统计学+线性代数）
- 谁说菜鸟不会数据分析（3本）

分析至少要学习那些课程-提升篇

- 线性回归分析基础
- 应用回归及分类
- 数据挖掘 概念与技
- 数据挖掘技术：应用于市场营销、销售与客户关系管理
- 数据挖掘与数据化运营实战思路、方法、技巧与应用
- 线性模型和广义线性模型（第3版）
- 图解机器学习
- 大数据与机器学习：实践方法与行业案例
- 机器学习
- 社交网络分析
- 机器学习实战

分析师至少要学习那些课程-工具学习篇

- SPSS统计分析基础 / 高级教程张文彤 第二版 （ 全2本 ）
- Python大战机器学习
- 机器学习与R语言
- r语言与数据挖掘
- sas变成演绎
- sas编程与数据挖掘商业案例
- 深入解析sas
- sas统计分析与应用实例
- 实用统计方法与sas系统
- 网络数据的统计分析与r语言
- 网络数据可视化与分析利器：Gephi

分析师应该掌握哪些统计分析软件

软件
1主攻 2个辅助

统计分析
SAS R SPSS Python

BI
excel taleau

软件学习不贪多，精通一个再学习另外一个

• 个人的软件使用情况：

- 做统计分析，数据挖掘用sas（熟练）；
- 机器学习：建议 r 和 python （基本应用）

- 不会建模的分析师一定不是个优秀的分析师
  - 一个入门的数据分析师至少能处理哪些问题？

• 数据分析师至少能解决哪些问题？

- **结构性问题:**
  - 1) 问：温度和小学生数量2个维度是如何影响冰激凌销量的？
  - 2) 问：女性粉丝数每增加一个单位，对广告效果产生什么影响（低cpm，高购买转化，高互动）？
  - 3）信用评分卡， 根据申请贷款的客户的各项数据进行打分

• 一个入门的数据分析师至少能处理哪些问题？

- **预测性问题**（回归 / 时序 / 分类）：分类问题要比回归更常见：
  - 1）用机器学习的方法找出哪些用户是作弊用户（根据数据中的知识对账号打标签1or0）
  - 2）预测哪些用户即将流失
  - 3）销量/库存预测

- 一个入门的数据分析师至少能处理哪些问题？

关联分析：

购物篮分析：

那些商品会一起被购买？
那几个页面之间存在重要关联？

商品最优结构特征优化

物品摆放
电商推荐

• 一个入门的数据分析师至少能处理哪些问题？

- **市场细分**：客户分群

- **移动电话卡为什么分为“全球通” “神州行” “动感地带”**

- 给自己打分，如下问题你目前能解决几个？

  - 结构性问题
  - 预测性问题
  - 客户分群
  - 关联分析

学习中的困惑是什么

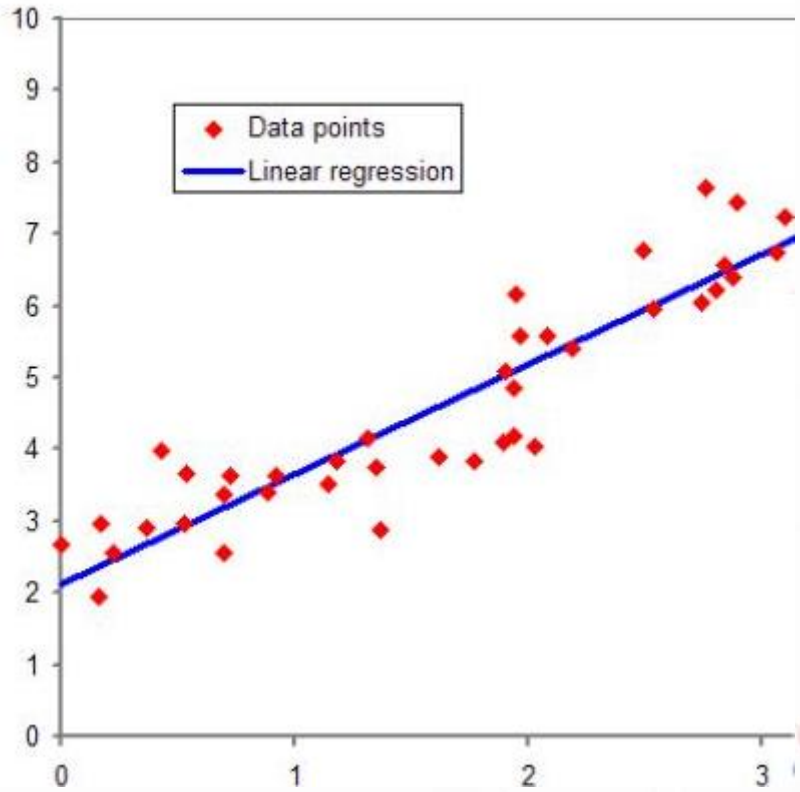
- 知识点记不住，概念无法理解
- 遇到模型就发懵
- 学新忘旧
- 如何把知识变成能力

学习中的困惑-概念无法理解

- 陌生的数学符号
- 被算法的推导过程迷惑
- 拗口的定义
- 剩下的实在无法理解的，手超三遍，总有一天你会开窍

- 当年我没懂的数学问题，在这2年的学习中全部懂了，时间会给你答案

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_{ki} X_{ki}$$



学习中的困惑

- 遇到模型就发懵,无从下手

• 如何学习模型-熟练掌握数据转换的方法

- 当数据无法满足模型的适用条件怎么办？

1）如果线性回归分析的y不符合正态分布 怎么办？

2）量刚问题是否处理如何处理？

数据的变换方法、标准化方法必会

• 如何学习模型

- 熟练掌握模型的原假设和适用条件

- 想得到正确可信的分析结果，请务必遵循模型的适用条件

#线性回归模型的假设:

a、线性:因变量与自变量间的线性关系。

b、正态性:因变量的正态性(残差服从正态性)。使用残差图诊断

c、独立同分布:残差间相互独立，且遵循同一分布，要求方差齐性

d、正交假定:误差项与自变量不相关，其期望为零。

- 想得到正确可信的分析结果，请务必遵循模型的适用条件

• 如何学习模型

每个模型能解决什么问题

他们都能解决什么问题？

线性回归

逻辑回归

方差分析，t检验

决策树

自行总结每个模型能解决的问题并汇集成文档

• 如何学习模型

模型的差别

模型优缺点

模型差别：在实际工作中，可以帮你快速的定位到适合的模型；

	T检验	方差分析	线性回归	逻辑回归	决策树
T检验					
方差分析					
线性回归					
逻辑回归					
决策树					

做一个模型矩阵，在交叉处写上模型的差别

举例：当我们知道模型的差别及优缺点后如下问题可以顺利解决

- 做分类器的时候，38响应变量每个都有小程度的缺失，如果删除缺失值后建模的话，样本量会降低请问你选择那个模型？

  - 支持向量机，逻辑回归，神经网络，决策树or决策树的集成算法你选那个？

我应该从那几个维度对比模型的优缺点

1 是否能解决结构性问题

2 对缺失值/异常值/量刚 是否敏感

3 是不是黑箱模型

4 能否处理连续（离散）的Y

在代码集合上，写上模型的优缺点

• 如何学习模型-如何把知识变成能力

尴尬问题：学得一身本事但无法融合到业务问题中，最终变为大表哥 大表姐

我是如何突破这个问题的：

1）论文是个好东西，业务与算法相结合最完美的学习资料；

2）案例丰富的书籍 是你的首选 如：ibm spps数分析与挖掘实战案例精粹

建议：借鉴别人的经验看别人是如何用知识解决问题的

- 学新忘旧怎么办
  - 记忆力差，命令容易混淆
  -
- 创建属于自己的代码本

• 学新忘旧-创建属于自己的代码本

- 创建并不断丰富自己的代码集合：由三层组成
  - 第一层： 模型的原假设，适用条件，核心算法公式
  - 第二层： 代码，一段代码解决一个问题，qq图怎么画？异方差如何检验,调整？等
  - 第三层： 写上你对改模型的总结和认识，优缺点是什么？
  - 重点：一个命令一段备注

```
#线性回归模型的假设：
a、线性：因变量与自变量间的线性关系。
b、正态性：因变量的正态性（残差服从正态性）。使用残差图诊断
c、独立同分布：残差间相互独立，且遵循同一分布，要求方差齐性，
   可以通过残差的独立性检验Durbin-Watson.
d、正交假定：误差项与自变量不相关，其期望为零。

notes:正态性与异方差在构建模型中的作用较小，对推论影响大。*/

 *输出PP图和QQ图：;
 proc reg data=data_anl.performance;
    model gcharacteristic=jaim jhonour jpromotion jcompetency/stb;
    *plot jaim*NQQ. jaim*NPP.;
    *plot jaim*NQQ.;
    *plot jaim*obs.;
    plot residual.*predicted.;
 run; quit;

 ods graphics on;
 proc reg data=data_anl.performance;
    model gcharacteristic=jaim jhonour jpromotion jcompetency/acov
thod=3 dw dwprob;
    output out=result r=gcresid student=studentr;
 run;quit;
 ods graphics off;
 *spec：异方差检验，检验模型第一矩和第二矩,即H0：误差的同质性和与解释变量不相关;
 *acov提供调整后的异方差标准误，及其显著性;
 *acov与hccmethod配合使用，hccmethod=提供4种修正的协方差矩阵；0适用于较大样本，1,2
用于小样本（小于250样本）;
 * H0：dw=2；p=1-1/2dw;
```

• 学新忘旧-对时间的充分把握

- **环境：不在电脑前**
- 学新念旧：我个人离公司单程1.5小时，所以我有充足的时间在路上学习；
  - 早晨：大脑清醒，学习全新的知识，读未读过的书
  - 晚上：复习上一本书，防止学新忘旧，做到念念不忘；

多以理论知识和算法理解为主

• 我的学习方法-对时间的充分把握

- **环境：电脑前**
- 正所谓百看不如一练，请不要放弃任何与统计软件打交道的机会，前几页推荐的工具类书籍是你练习的良好导师；
- 优点1：你无须为数据源担心，统计软件都自带数据源

工具实操类书籍

• 如何应用模型-根据业务目标进行模型组合

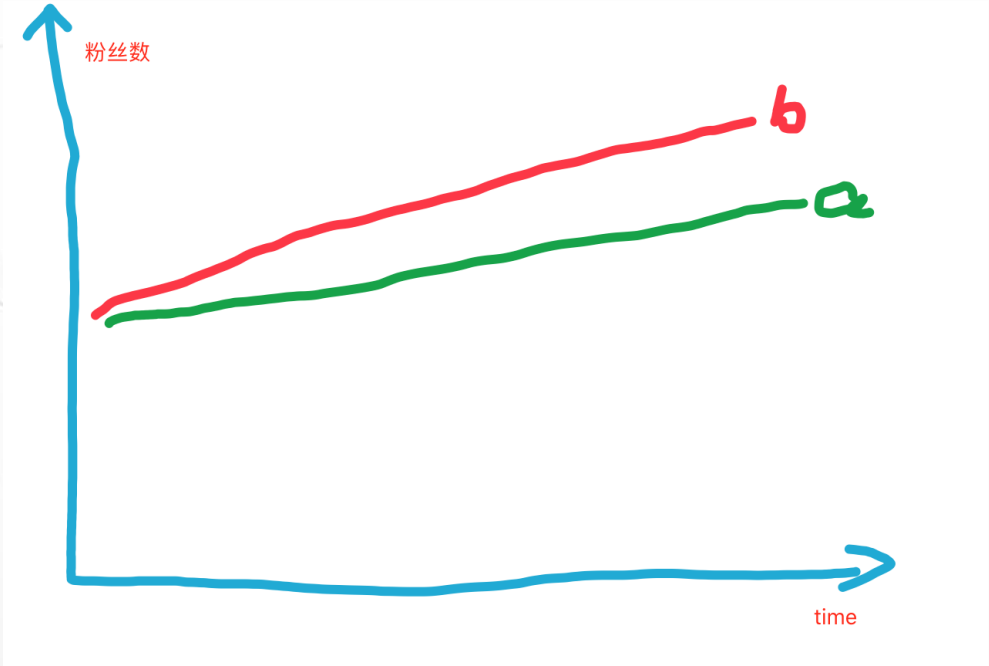
- K-means聚类 配合 系统聚类
  - 海量数据的聚类 既有速度又提高了精度

- 借助聚类分析和woe组合使用可以探索不同属性用户群组的响应率

算法组合可以有效的解决你的业务问题

• 如何应用模型-模型可用的不仅仅是模型

- 模型可用的不仅仅是模型， 还有模型的基算法
  - 例如：线性回归的最小二乘，可以用来衡量微博账号一段时间的粉丝增长速度（斜率）



• 如何应用模型-模型可用的不仅仅是模型

- **领悟算法精髓，你的人生不只一个RFM（时间，频率，金额）**

- 借助该模型的思想， 把如上三个维度 进行替换，就可以生成属于你的独有模型

  如：衡量广告主的价值 （rank最后一次投递广告时间，rank投放频次，rank投放金额）

  如：衡量自媒体的价值（rank阅读数，rank互动数，rank广告链接产生的点击数）

• 心得技巧：回归问题与分类问题相互转换

回归问题转换为分类问题，同样可以解决业务需求而且预测难度还变低了；

- 心得技巧：我们还应该掌握什么技能

指数排名算法

数据脱敏的方法：

- 除了会建模 我们还应该掌握什么技能

CHINA DATA ANALYST SUMMIT

CDA 数据分析师 www.cda.cn

THANKS

跨界互联 数聚未来

第四届中国数据分析师行业峰会

CHINA DATA ANALYST SUMMIT