

# 法律声明

□ 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



# EM算法实践

---

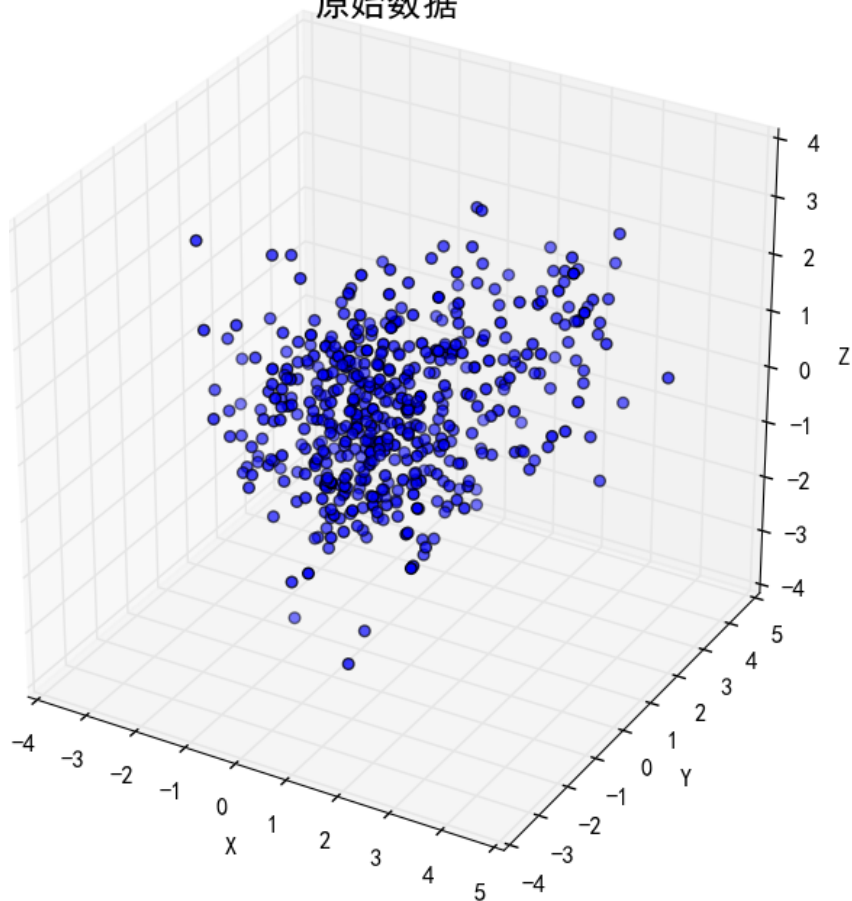


小象学院  
ChinaHadoop.cn

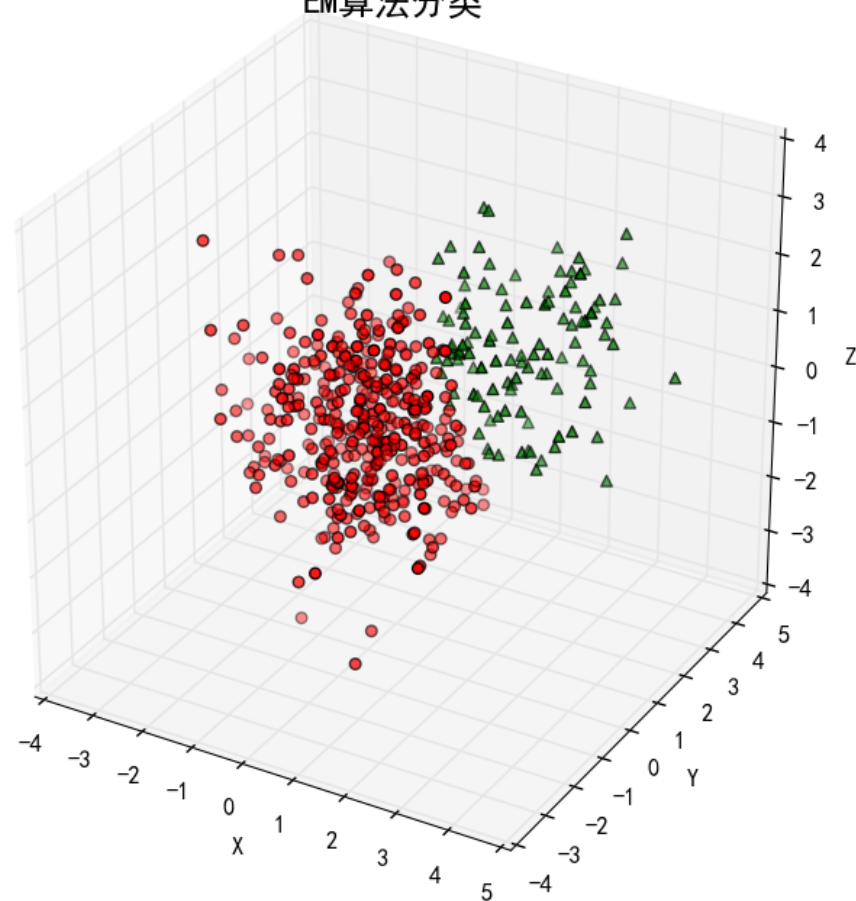
邹博

# 多维GMM聚类：EM算法

原始数据



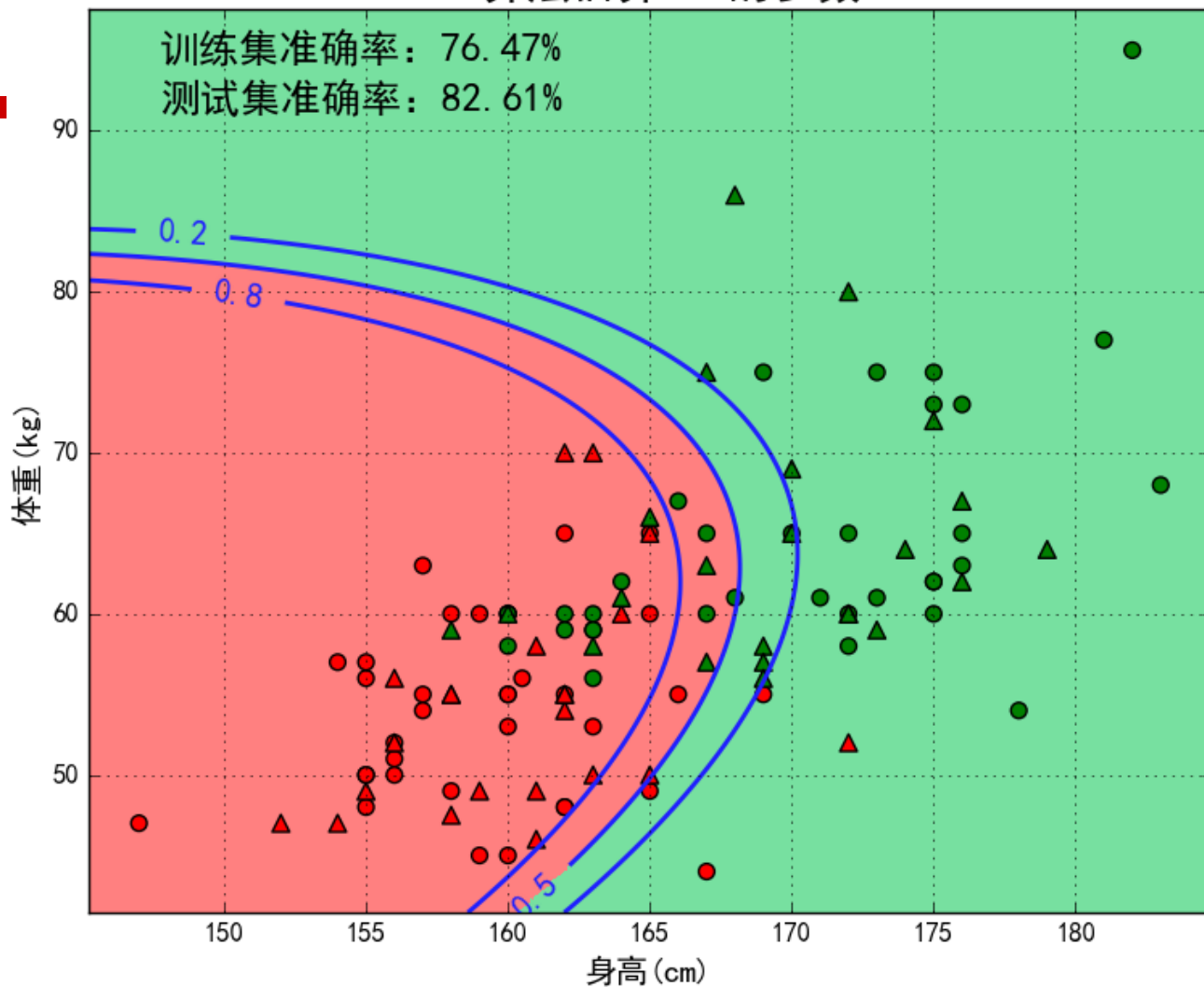
EM算法分类



# EM算法

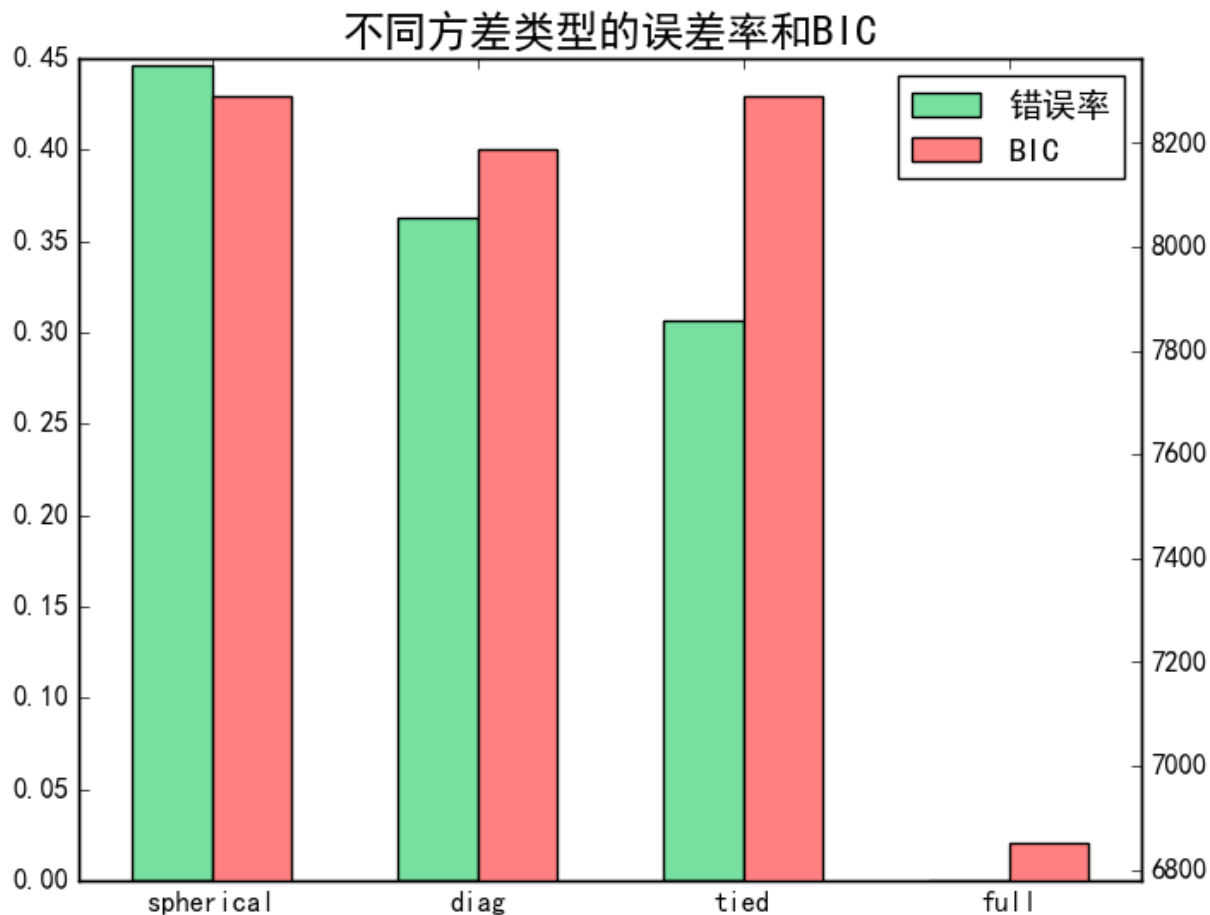
□ 副产品  
■ 等值线

EM算法估算GMM的参数



# GMM调参

□ 副产品  
■ 双y轴

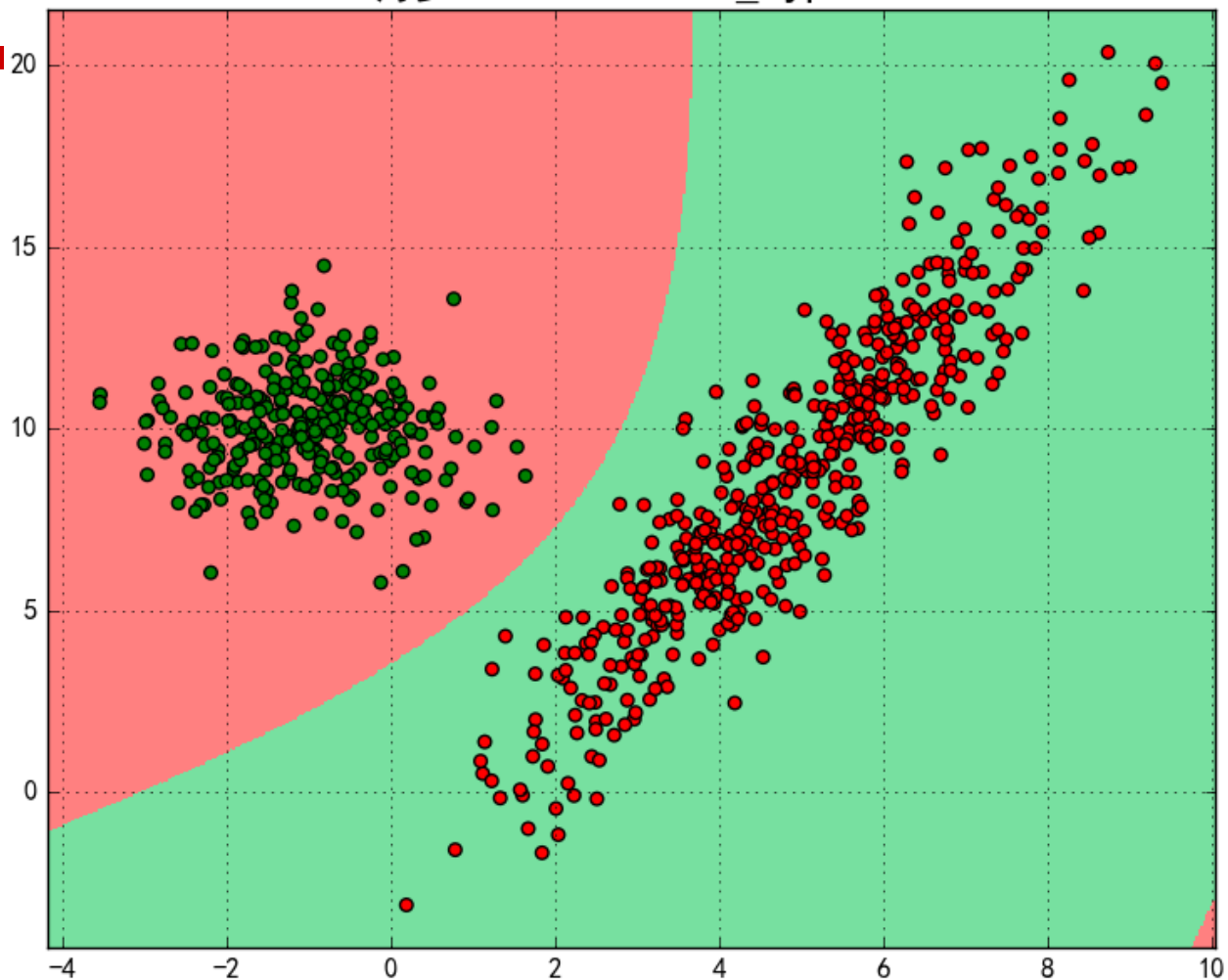


# 模型选择的准则

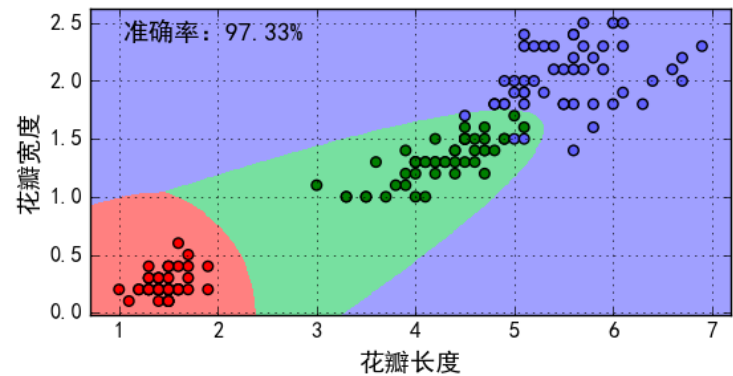
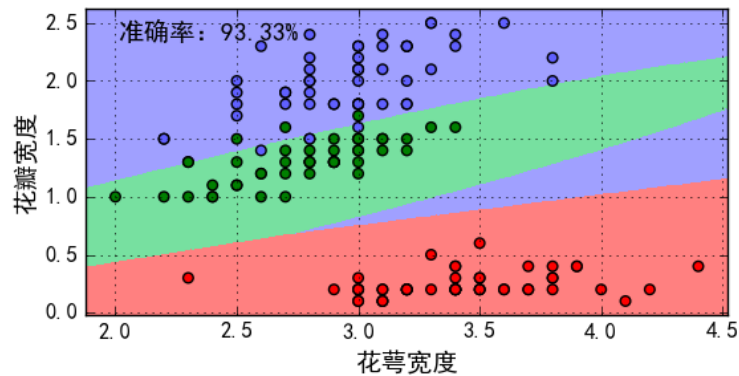
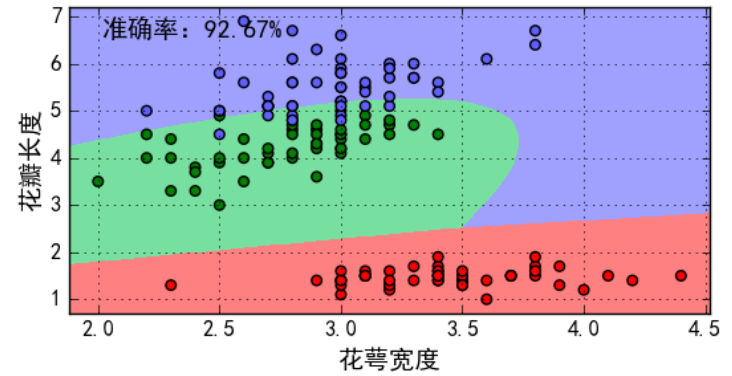
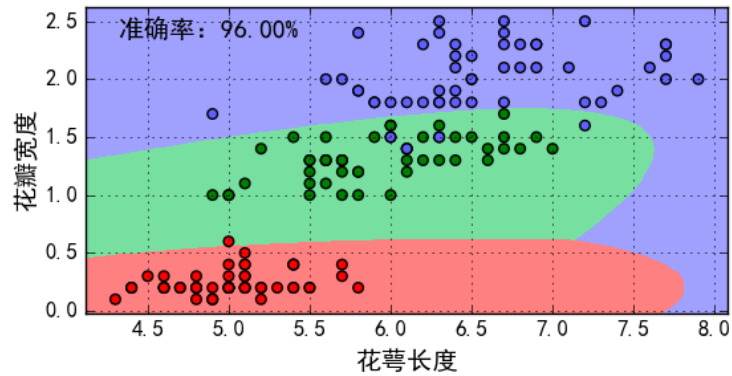
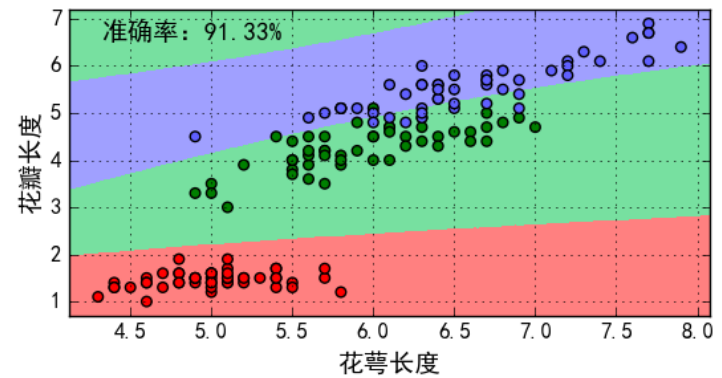
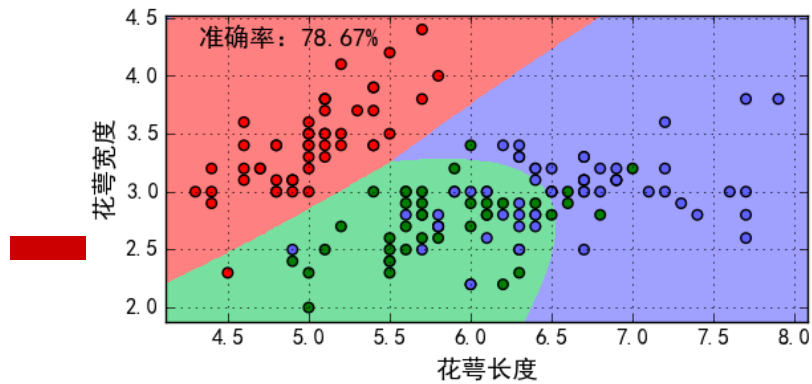
- 记： $L$ 为某模型下样本的似然函数值， $k$ 为模型中未知参数的个数(维度)， $n$ 为样本个数，则：
- $AIC = -2 \ln L + 2k$ 
  - akaike information criterion
  - 日本统计学家赤池弘次(Akaike)于1973年提出
- $BIC = -2 \ln L + (\ln n)k$ 
  - Bayesian Information Criterion/Schwarz criterion
  - Akaike于1976年通过改进AIC得到
  - Gideon E. Schwarz在1978年根据Bayesian理论重新发现

# GMM调参

GMM调参: covariance\_type=full



# EM算法无监督分类鸢尾花数据



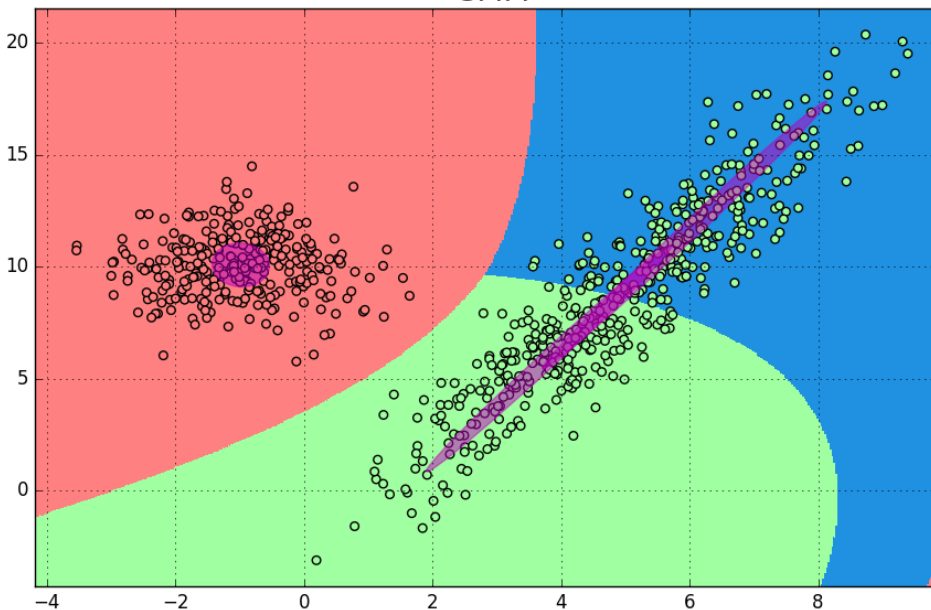


# DPGMM

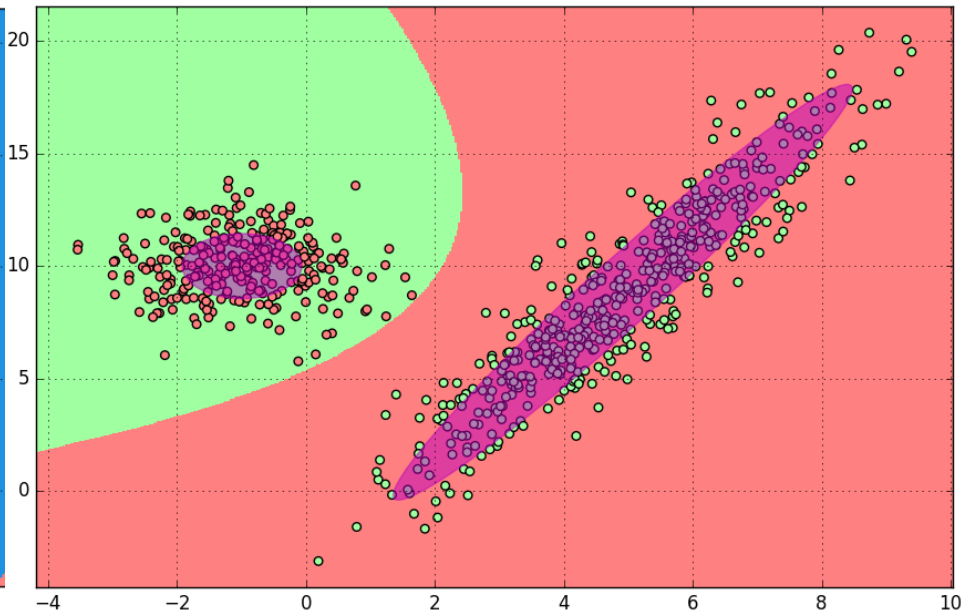
## □ Dirichlet Process Gaussian Mixture Model

■ 先验分布

GMM



DPGMM



# 复习：二项分布的最大似然估计

- 投硬币试验中，进行N次独立试验，n次朝上，N-n次朝下。
- 假定朝上的概率为p，使用对数似然函数作为目标函数：

$$f(n | p) = \log(p^n (1-p)^{N-n}) \xrightarrow{\Delta} h(p)$$

$$\frac{\partial h(p)}{\partial p} = \frac{n}{p} - \frac{N-n}{1-p} \xrightarrow{\Delta} 0 \Rightarrow p = \frac{n}{N}$$

## 二项分布与先验举例

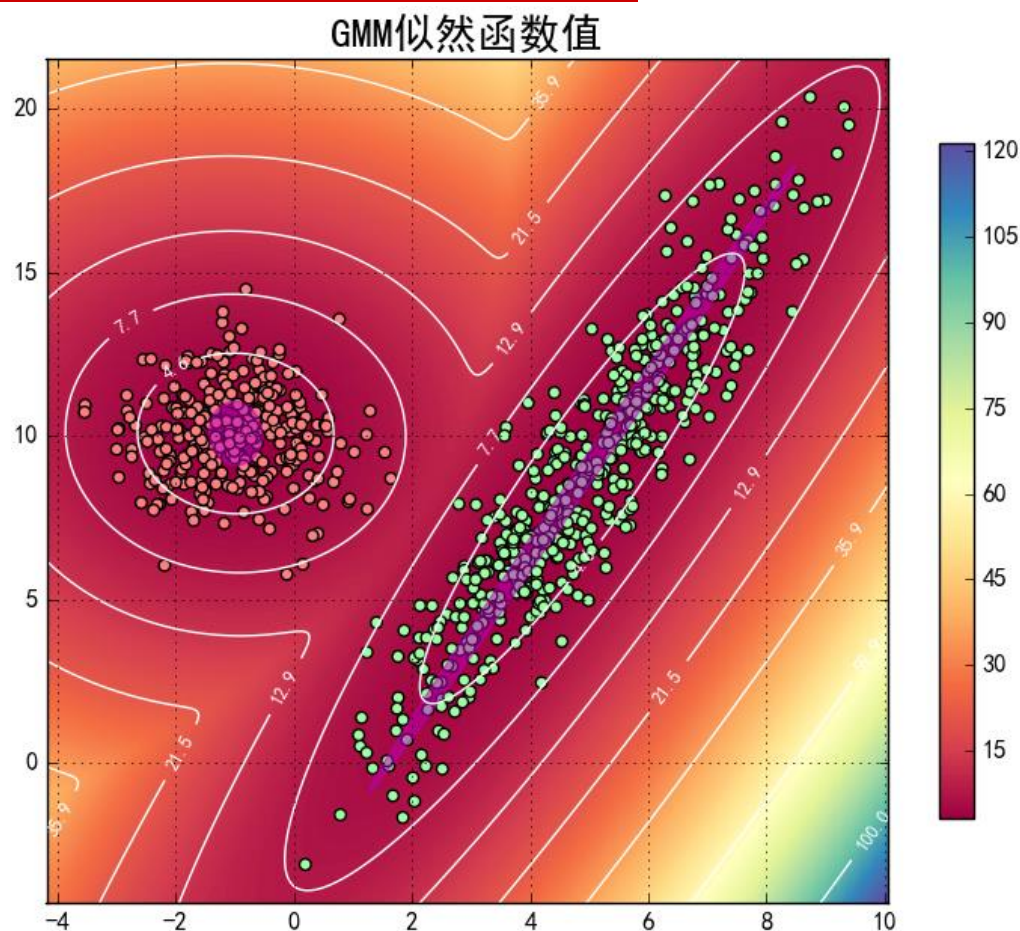
□ 在校门口统计一定时间段内出入的男女生数目分别为 $N_B$ 和 $N_G$ ，估算该校男女生比例。

$$\begin{cases} P_B = \frac{N_B}{N_B + N_G} \\ P_G = \frac{N_G}{N_B + N_G} \end{cases}$$

□ 若观察到4个女生和1个男生，可以得出该校女生比例是80%吗？

□ 修正公式：
$$\begin{cases} P_B = \frac{N_B + 5}{N_B + N_G + 10} \\ P_G = \frac{N_G + 5}{N_B + N_G + 10} \end{cases} \Rightarrow \begin{cases} P_B = \frac{1 + 5}{1 + 4 + 10} = 40\% \\ P_G = \frac{4 + 5}{1 + 4 + 10} = 60\% \end{cases}$$

# 似然函数值：复习Matplotlib的绘图



```
0.bankpay服务器端支付通知,211.103.171.164,1451608146,info
28761,登录成功,112.64.60.28,1451608269,info
26241,登录成功,27.19.182.171,1451608918,info
12717,登录成功,49.65.132.141,1451608940,info
1002,登录成功,122.234.233.79,1451609652,info
29131,登录成功,114.246.155.33,1451609663,info
29131,gongfeng观看课时《第八课:最大熵模型》,114.246.155.33,1451609676,info
1002,zixu4728观看课时《广告行业实例-用户行为分析、归类》,122.234.233.79,1451609682,info
29131,gongfeng观看课时《第八课:最大熵模型》,114.246.155.33,1451609766,info
27131,用户二维码登录,114.111.167.101,1451609789,info
49136,登录成功,171.216.73.17,1451610088,info
定时任务(#3)开始执行!,221.194.176.18,1451610251,info
定时任务(#3)执行结束!,221.194.176.18,1451610251,info
3098,登录成功,118.26.176.5,1451610297,info
5706,登录成功,61.135.152.208,1451610382,info
27733,登录成功,110.255.176.51,1451610484,info
27733,登录成功,110.255.176.51,1451610485,info
4965,登录成功,111.207.1.140,1451610533,info
4965,登录成功,111.207.1.140,1451610533,info
25890,登录成功,114.255.40.29,1451610710,info
25890,定时任务(#4)开始执行!,114.255.40.29,1451610711,info
25890,定时任务(#4)执行结束!,114.255.40.29,1451610711,info
29100,登录成功,118.244.254.23,1451610829,info
28130,登录成功,118.244.254.23,1451611188,info
1,登录成功,110.255.176.51,1451611566,info
28011,小明熊哥哥观看课时《kafka架构及spooop基础》,124.205.212.46,1451721813,info
0,定时任务(#3)开始执行!,125.39.112.12,1451721814,info
0,定时任务(#3)执行结束!,125.39.112.12,1451721814,info
19516,登录成功,101.224.28.47,1451721980,info
19516,登录成功,101.224.28.47,1451721980,info
7769,登录成功,101.224.28.47,1451721931,info
19516,登录成功,101.224.28.47,1451721980,info
29216,登录成功,118.244.254.23,1451721981,info
29216,登录成功,118.244.254.23,1451721981,info
19516,登录成功,101.224.28.47,1451722002,info
19516,登录成功,101.224.28.47,1451722002,info
8425,登录成功,110.255.176.51,1451722003,info
19516,登录成功,101.224.28.47,1451722288,info
19516,登录成功,101.224.28.47,1451722315,info
19516,登录成功,101.224.28.47,1451722408,info
19516,登录成功,101.224.28.47,1451722465,info
26168,登录成功,106.121.66.244,1451722477,info
29305,登录成功,210.73.8.130,1451722481,info
19516,登录成功,101.224.28.47,1451722528,info
985,登录成功,116.227.48.64,1451722540,info
28325,Normal观看课时《第六课:回归》,218.76.28.110,1451722575,info
28325,Normal观看课时《第八课:最大熵模型》,218.76.28.110,1451722578,info
28325,Normal观看课时《第九课:神经网络和推荐》,218.76.28.110,1451722583,info
28325,Normal观看课时《第十课:图论》,218.76.28.110,1451722583,info
19516,登录成功,101.224.28.47,1451722648,info
19516,登录成功,101.224.28.47,1451722648,info
```

# 附：小象学院可疑账号检测

□ 小象学院是近年来非常活跃的大数据机器学习在线培训机构，拥有十余年在线教育经验，注册账号数十万，2017年2月底，收费学员人数近十万。

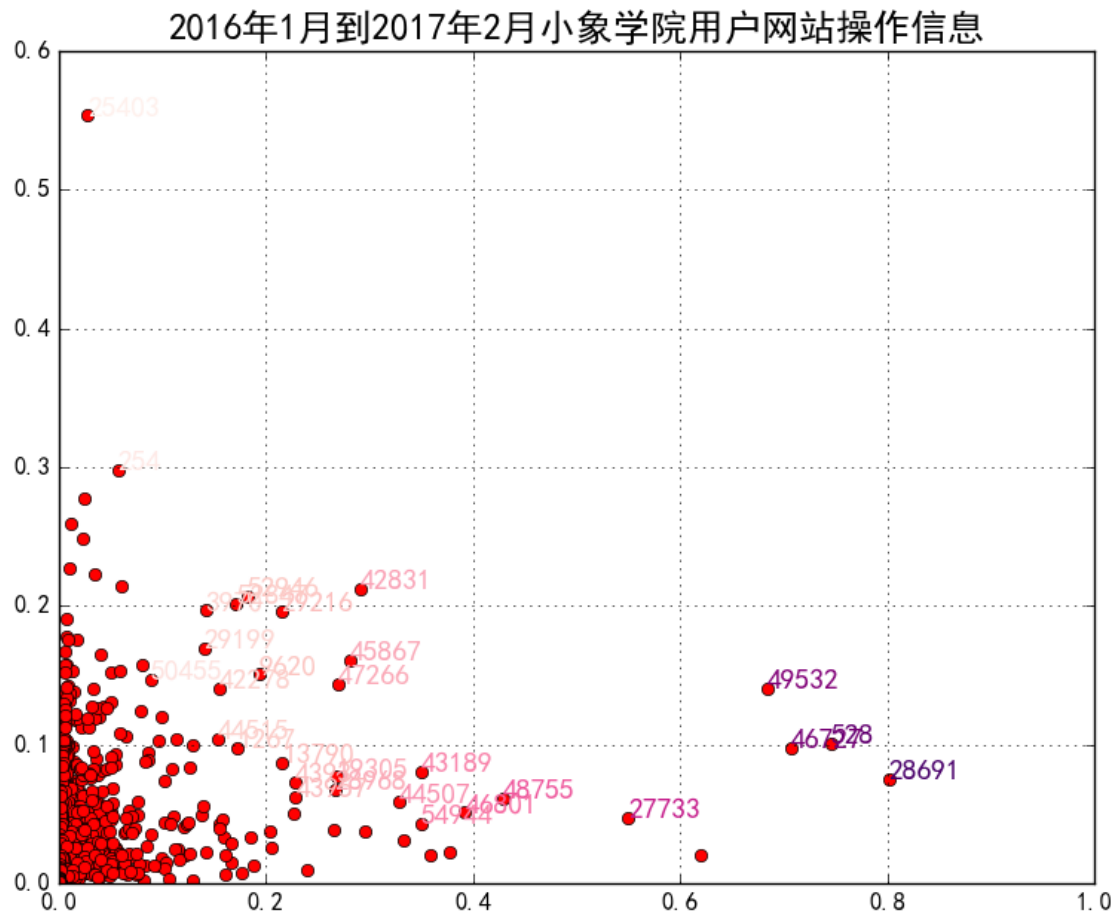
□ 由于在线课程、视频、数据、代码的复制相对容易，发现有账号倒卖现象。从用户的网站操作日志(登陆、观看、下载、社区等)，试发现可疑账号ID，以备后续处理。

# 可考虑的特征

---

- 选择2016年1月到2017年2月的用户数据，日志共240M，清洗和规则化后，考虑的特征如下：
  - 用户登陆次数
  - 用户登陆使用的IP个数
  - IP的变化次数
  - 观看视频次数
  - 社区留言次数
  - 用户报名的课程数目
  - 每个视频的观看时间/视频总时间
  - 奇异值(离群点)

# 特征选择和降维后的处理结果



# 参考文献

---

- [https://en.wikipedia.org/wiki/Bayesian\\_information\\_criterion](https://en.wikipedia.org/wiki/Bayesian_information_criterion)



# 我们在这里

□ <http://wenda.ChinaHadoop.cn>

■ 视频/课程/社区

□ 微博

■ @ChinaHadoop

■ @邹博\_机器学习

□ 微信公众号

■ 小象

■ 大数据分析挖掘

The screenshot shows the website [wenda.chinahadoop.cn/explore/](http://wenda.chinahadoop.cn/explore/). The page features a navigation bar with a search box and a '发现' (Discover) button circled in red. Below the navigation bar, there are tabs for '全部', '招聘求职', '机器学习', '大数据平台技术', 'DCon', '大数据行业应用', 'NoSQL数据库', '数据科学', and '江湖救急'. The main content area displays a list of questions and answers, including:

- yarn运行时一直重复这个info...好像没找到资源, 应该从哪里检查呢?
- 两种不同的相关推荐列表
- 如何在Linux下配java的JDK?
- sqoop把mysql数据导入Hbase报如图错误
- 泛化误差公式推导
- kafkaOffsetMonitor打开页面以后无法显示内容?
- markdown公式编辑\$符号不起作用
- hadoop-2.6.2-src源码编译成功之后找不到native下如图一所示文件, 执行图三所示搜索命令也没有找到, 进入源码编译之后的目录如图二! 这个文件找不到怎么解决呢? 是编译没产生?
- opentsdb安装时出现72个warning, 是正常的么?
- 关于在线广告和个性化推荐区别的一点浅见

On the right side, there are sections for '专题' (Topics) including '招聘求职', '大数据行业应用', '数据科学', '系统与编程', and '云计算技术'. There is also a '热门话题' (Popular Topics) section with '机器学习' (193 questions, 86 followers), 'spark' (200 questions, 91 followers), '算法' (55 questions, 65 followers), 'linux' (179 questions, 47 followers), and 'hbase' (224 questions, 62 followers). A '热门用户' (Popular Users) section lists users like 'gongfc', 'Hagrid', 'yanglei', '天然下雨', and 'hiveman'.

---

感谢大家!

恳请大家批评指正!