法律声明

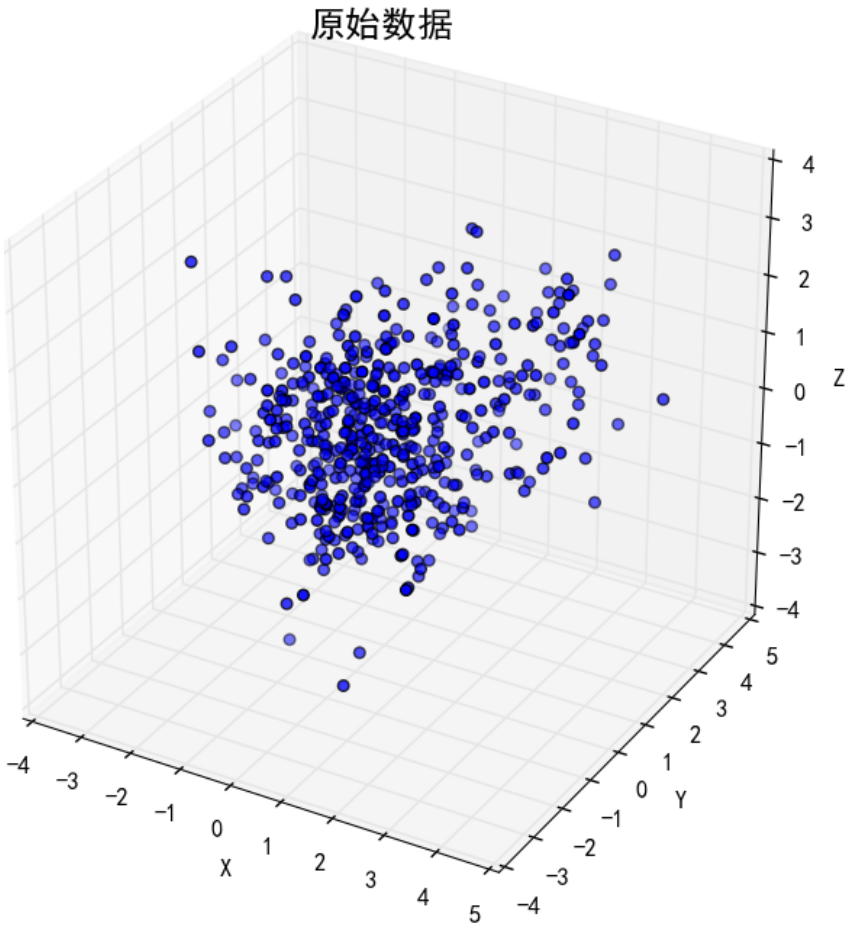
- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

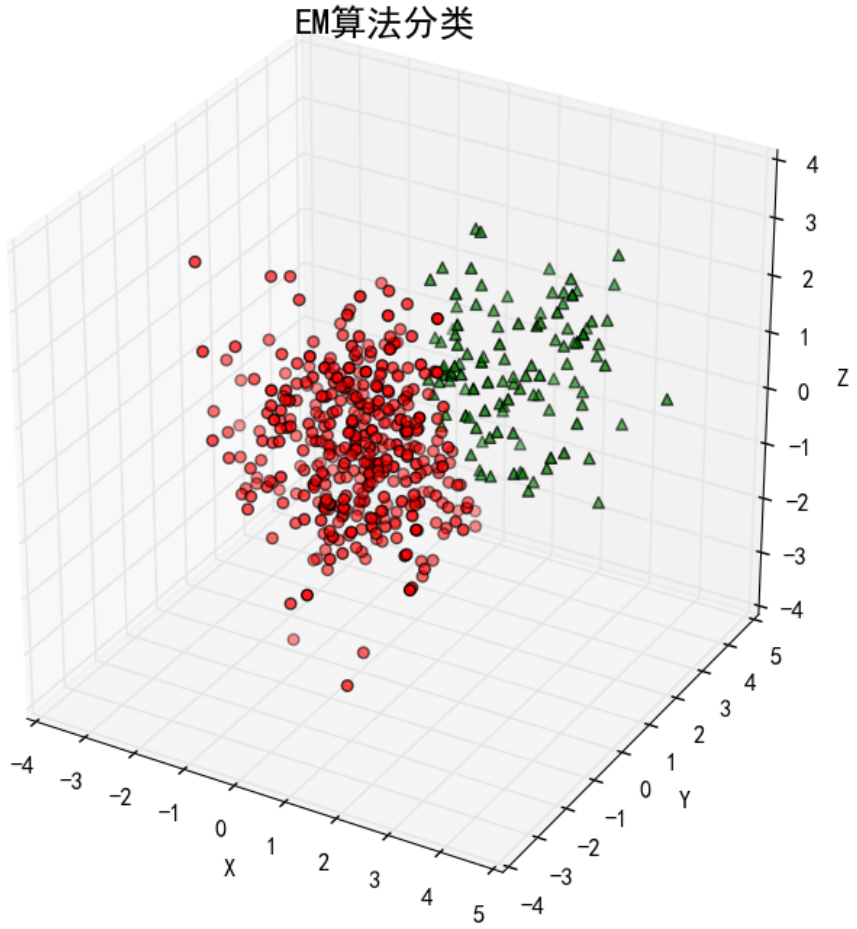
- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

EM算法实践

邹博

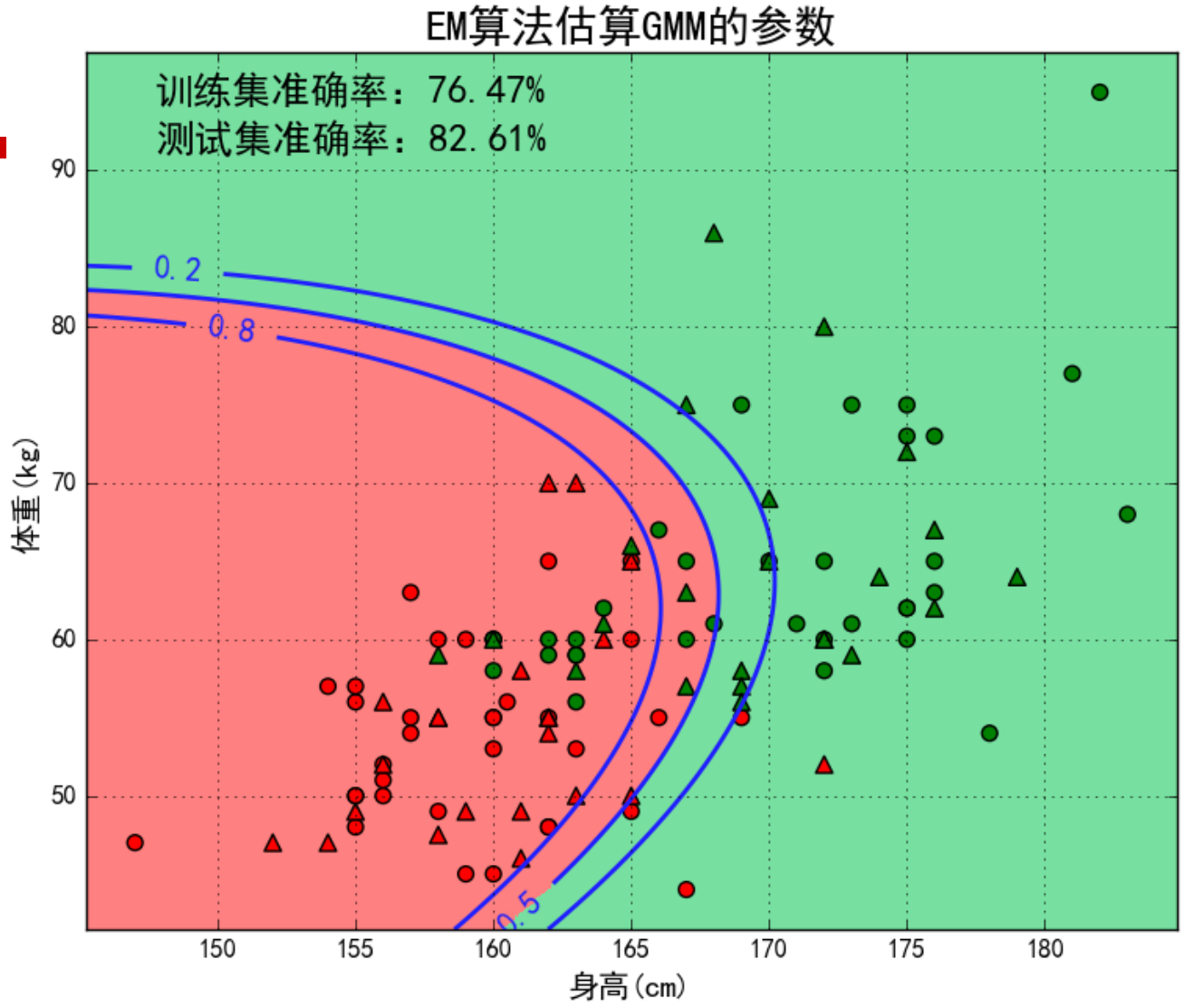
多维GMM聚类：EM算法





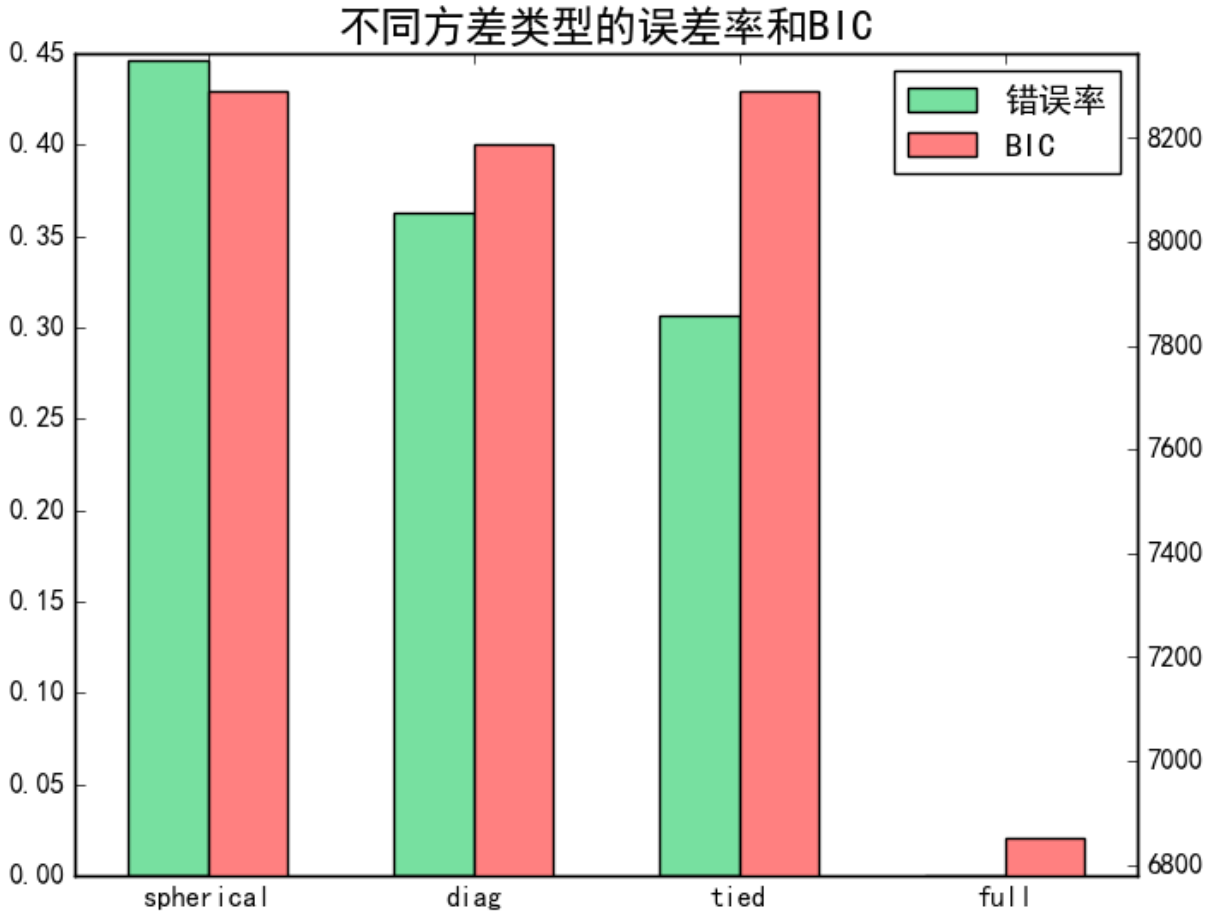
EM算法

- 副产品
  - 等值线


EM算法估算GMM的参数
训练集准确率：76.47%
测试集准确率：82.61%
0.2
0.8
0.5
体重(kg)
身高(cm)
90
80
70
60
50
150
155
160
165
170
175
180

GMM调参

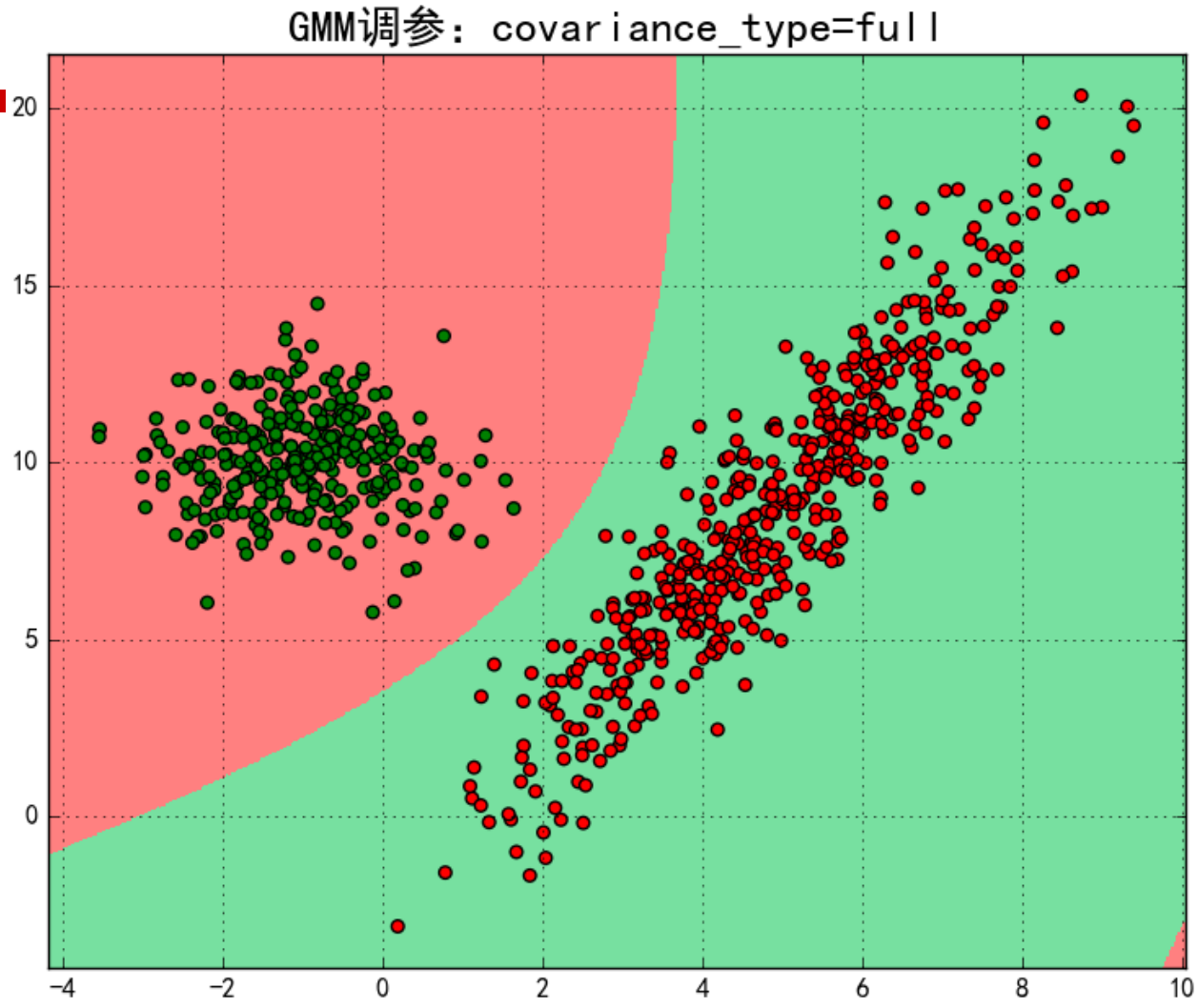
- 副产品
  - 双y轴



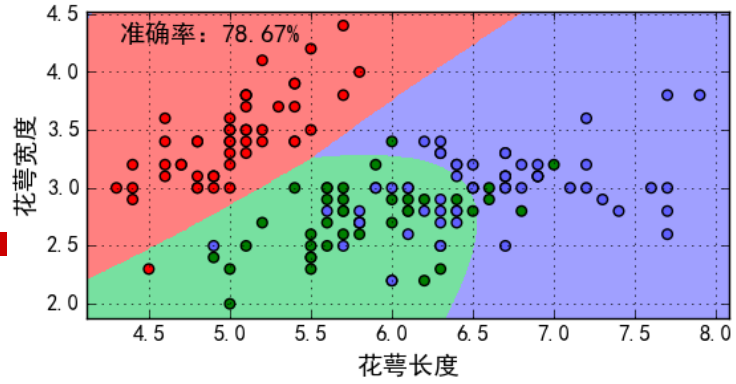
模型选择的准则

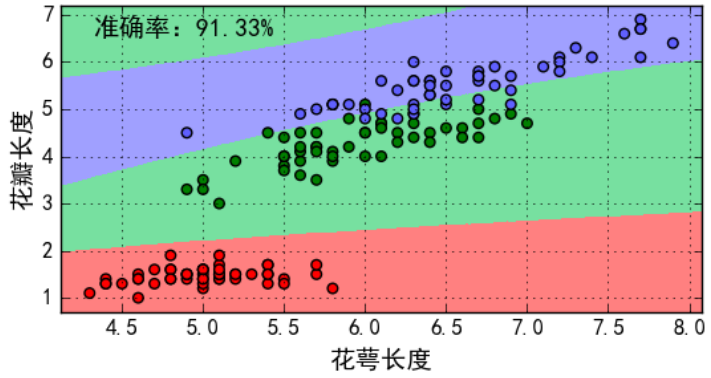
- 记：L为某模型下样本的似然函数值，k为模型中未知参数的个数(维度)，n为样本个数，则：
- $AIC = -2\ln L + 2k$
  - akaike information criterion
  - 日本统计学家赤池弘次(Akaike)于1973年提出
- $BIC = -2\ln L + (\ln n)k$
  - Bayesian Information Criterion/Schwarz criterion
  - Akaike于1976年通过改进AIC得到
  - Gideon E. Schwarz在1978年根据Bayesian理论重新发现

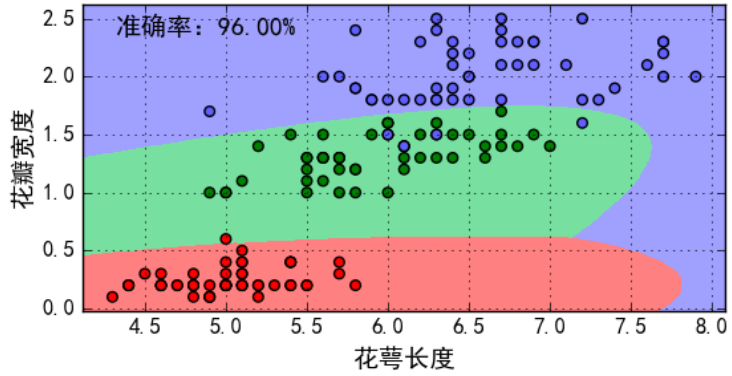
GMM调参

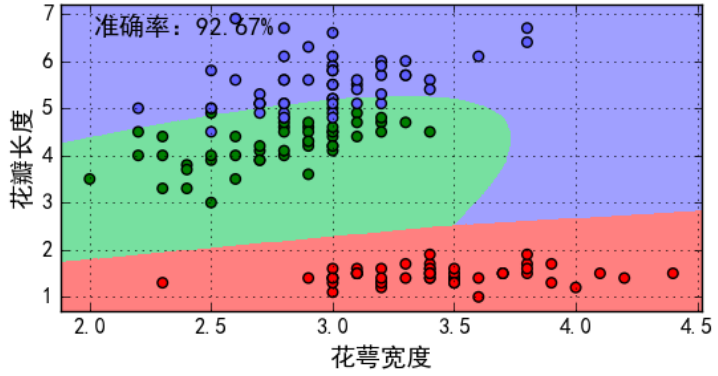


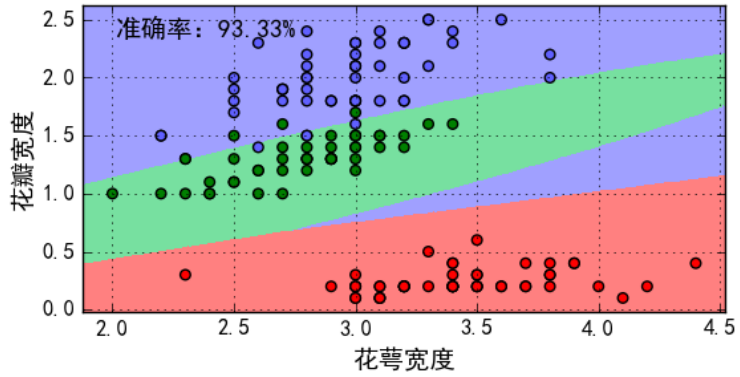
EM算法无监督分类鸢尾花数据

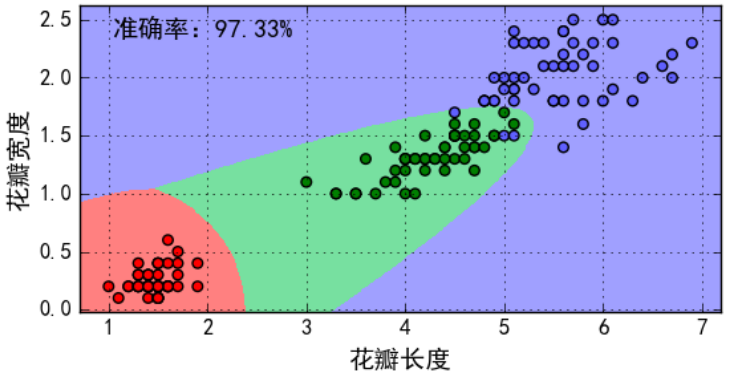






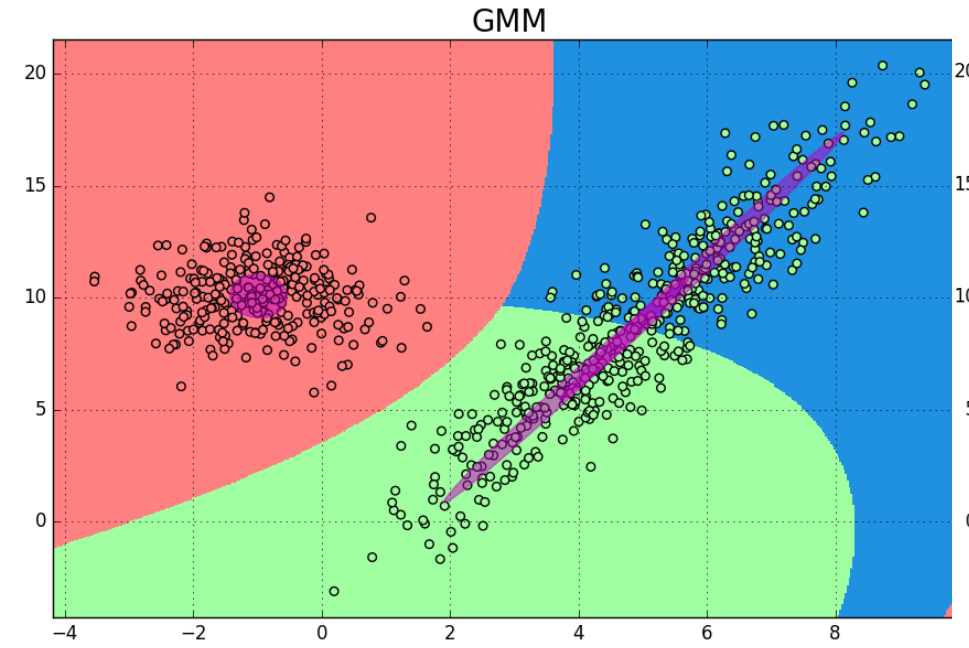


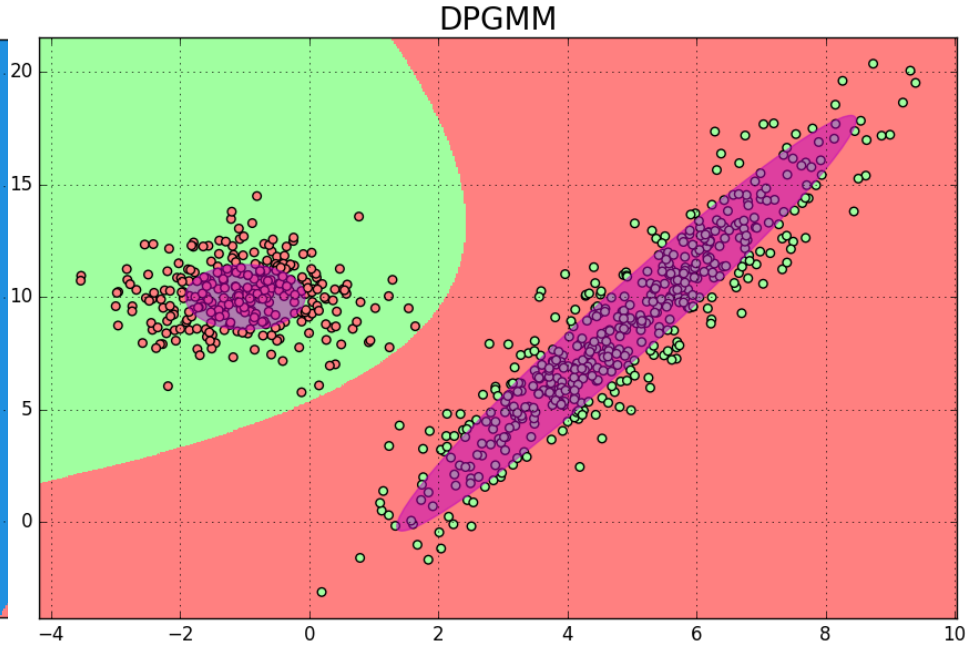




DPGMM

- Dirichlet Process Gaussian Mixture Model
  - 先验分布





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：

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

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

二项分布与先验举例

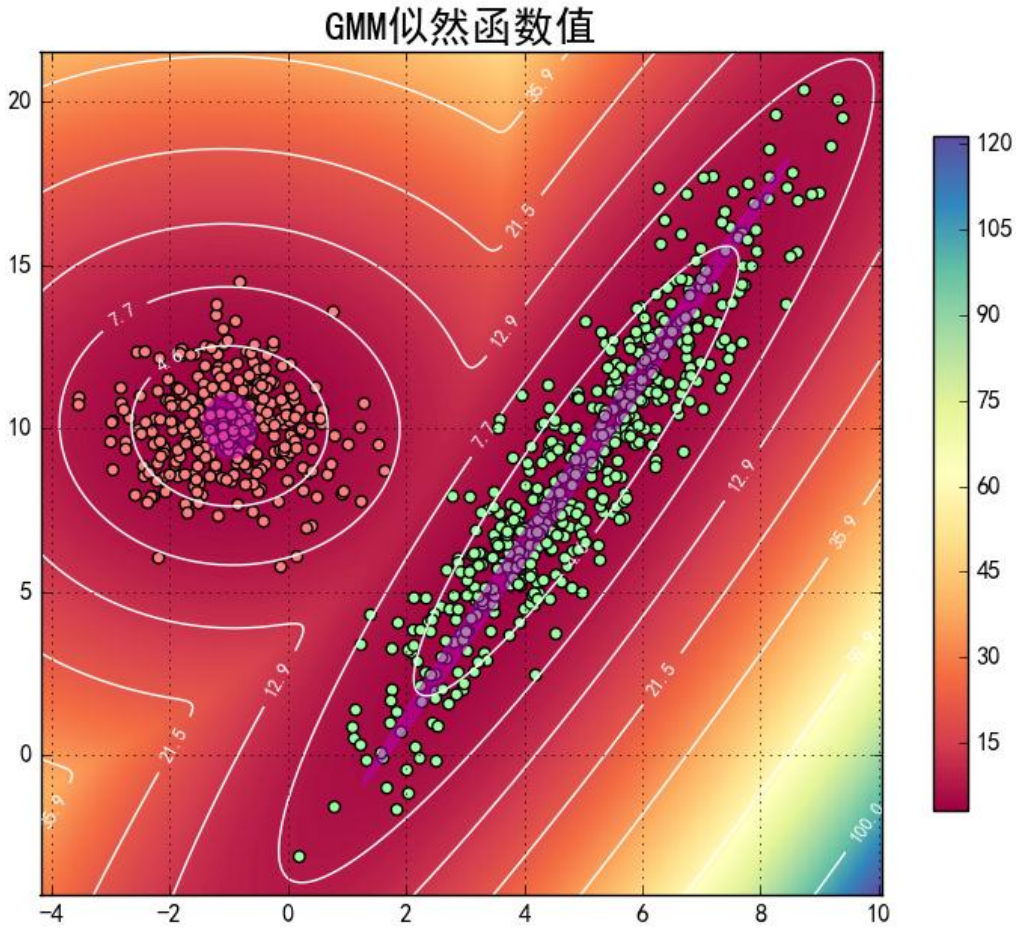
- 在校门口统计一定时间段内出入的男女生数目分别为N_B和N_G，估算该校男女生比例。

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

- 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：

$$\begin{cases} P_B = \dfrac{N_B + 5}{N_B + N_G + 10} \\ P_G = \dfrac{N_G + 5}{N_B + N_G + 10} \end{cases} \Rightarrow \begin{cases} P_B = \dfrac{1+5}{1+4+10} = 40\% \\ P_G = \dfrac{4+5}{1+4+10} = 60\% \end{cases}$$

似然函数值：复习Matplotlib的绘图



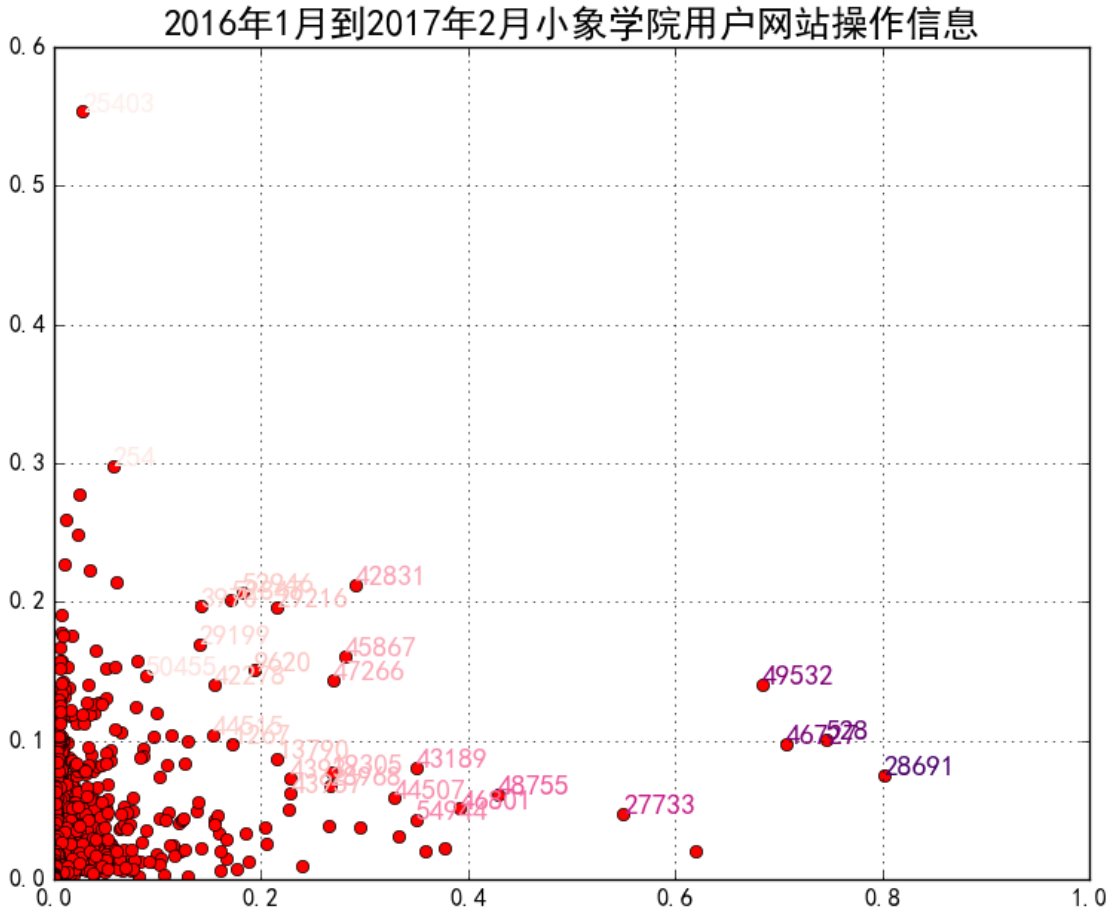
附：小象学院可疑账号检测

- 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，拥有十余年互联网和在线教育经验，注册账号数十万，截止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

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
  - 小象学院
  - 大数据分析挖掘

感谢大家！

恳请大家批评指正！