法律声明

- 课程详情请咨询
  - 微信公众号：北风教育
  - 官方网址：*http://www.ibeifeng.com/*

人工智能之机器学习

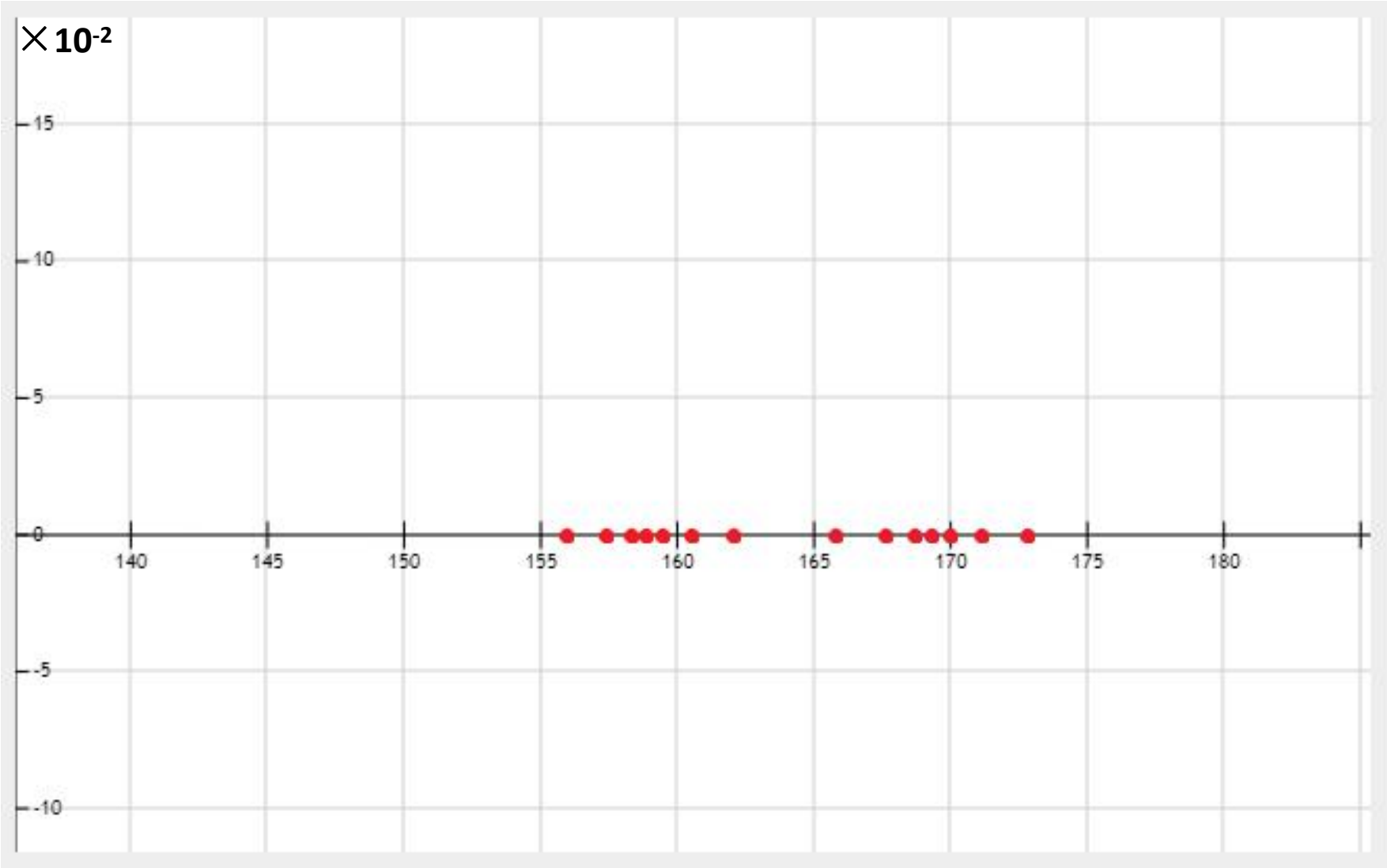
高斯混合模型（Gaussian Mixture Model）

主讲人：赵翌臣

上海育创网络科技有限公司

高斯混合模型直观理解

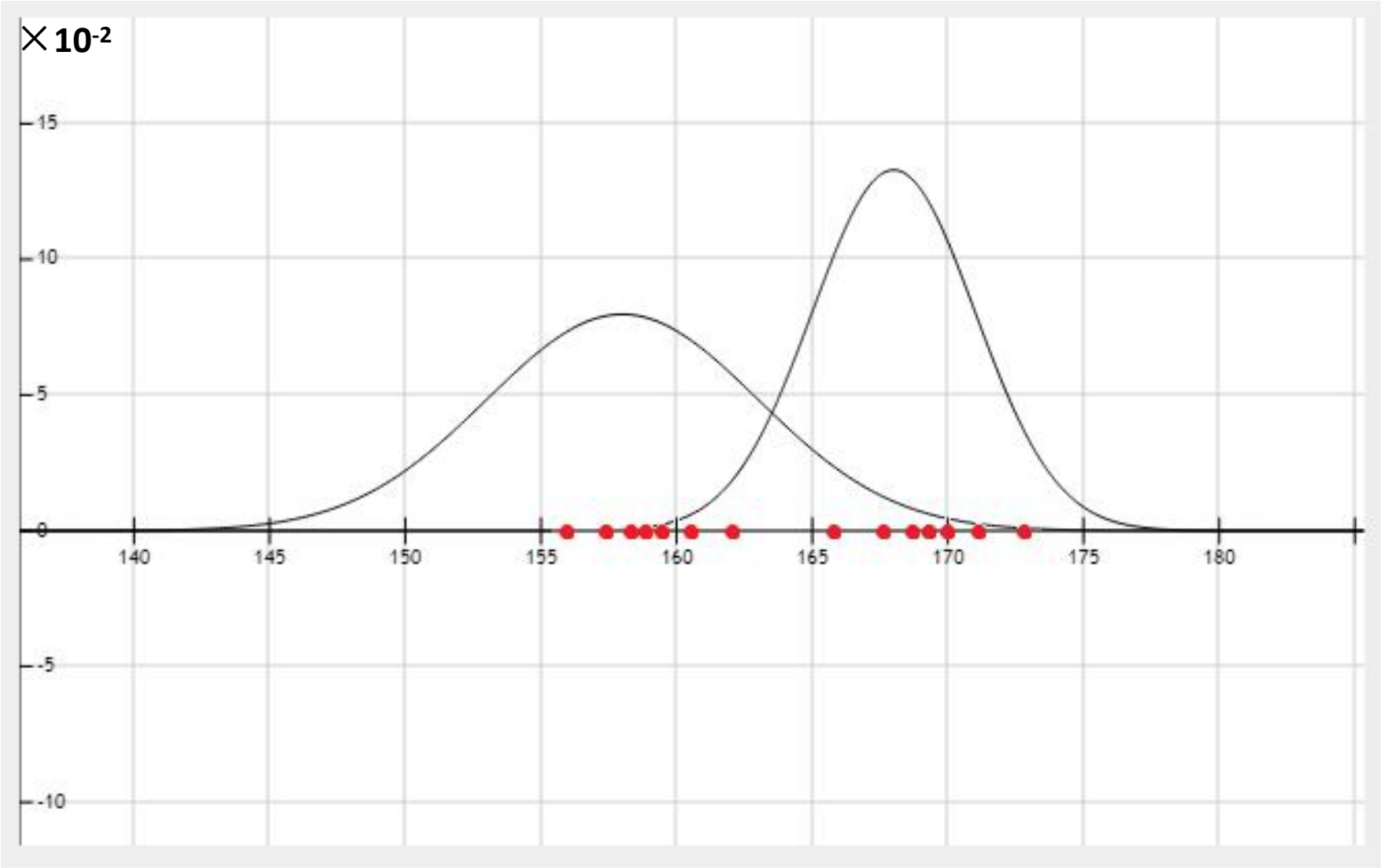
- 假设有如下数据集，试建立一个高斯模型



合理吗

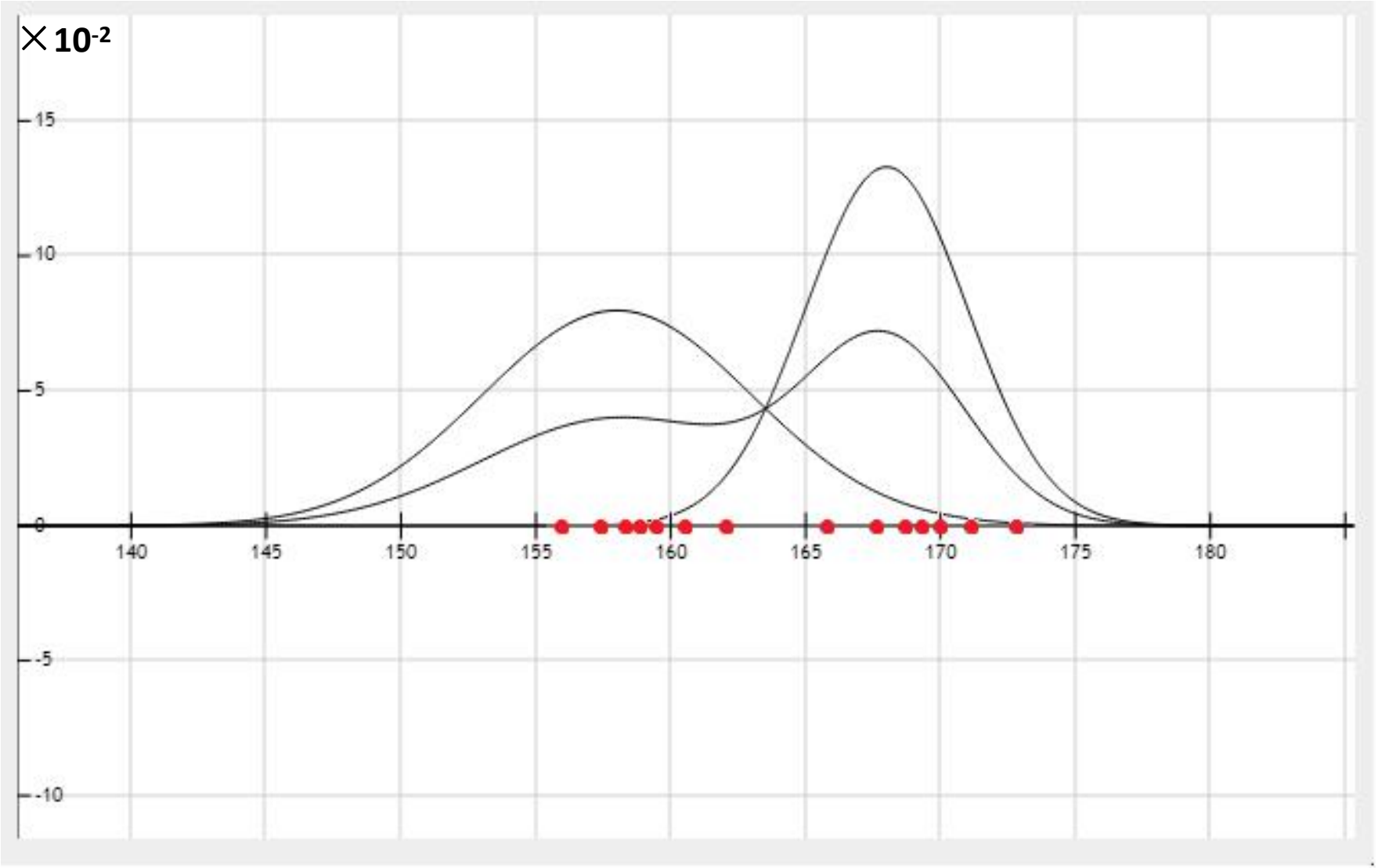
高斯混合模型直观理解

- 使用两个高斯模型呢?



高斯混合模型直观理解

- 数据的分布可能是多个分布的叠加



高斯混合模型（Gaussian Mixture Model）

- 介绍
  - 在统计学中，混合模型是用于表示总体群体中亚群体的存在的概率模型。
  - 高斯混合模型（Gaussian Mixture Model）为单一高斯概率密度函数的延伸，用多个高斯概率密度函数（正态分布曲线）精确地量化变量分布，是将变量分布分解为若干基于高斯概率密度函数（正态分布曲线）分布的统计模型。
- 应用
  - 聚类：样本受到哪个高斯分布的作用大，就认为样本属于哪个高斯分布

高斯混合模型（Gaussian Mixture Model）

- 高斯混合模型是指具有如下形式的概率分布模型：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$$

- 其中，αk是系数，αk ≥0，$\sum_{k=1}^{K}\alpha_k=1$；$\phi(y|\theta_k)$ 是高斯分布，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right)$$

称为第k个分模型

用EM算法估计高斯混合模型参数

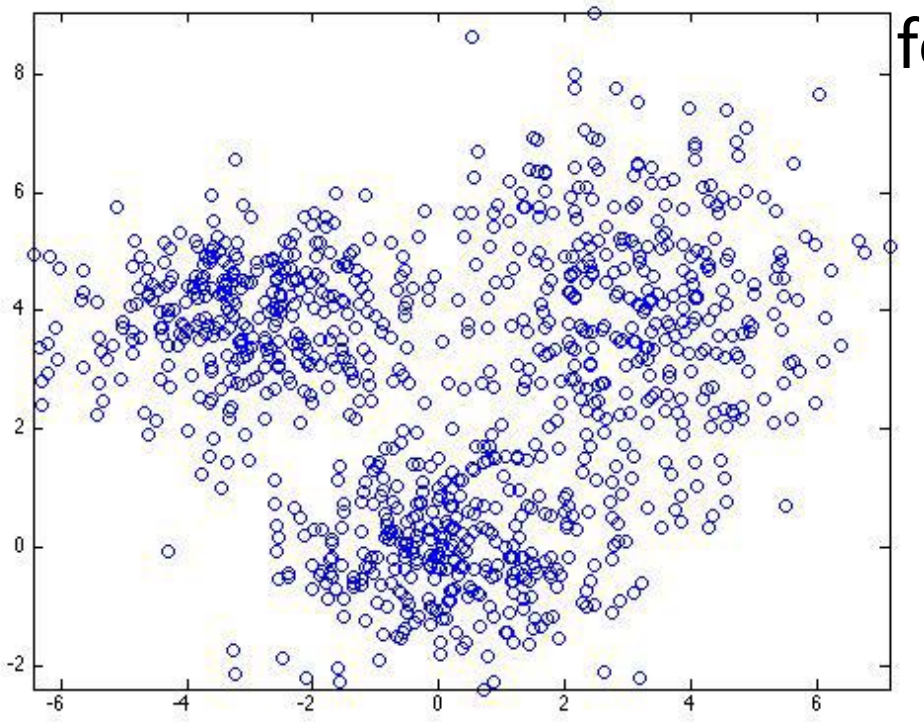
- 假设观测数据$y_1, y_2, ..., y_N$由高斯混合模型生成，

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k)$$

- 其中，$\theta=(\alpha_1,\alpha_2,...,\alpha_K;\theta_1,\theta_2,...,\theta_K)$。我们用EM算法估计高斯混合模型的参数θ
- 确定隐变量：可以设想观测数据y是这样产生的，首先依概率αk选择第k个高斯分布分模型；然后依该分模型生成观测数据y。这时，观测数据y是已知的，但是他出自那个分布是未知的，这就可以用EM算法进行迭代求解，因为EM算法就是含有隐变量的概率模型参数的极大似然估计法

用EM算法估计高斯混合模型参数（可视化）

- 假设你有如下数据集，你说他们来自某一个高斯分布吗？显然不能，这是一个典型的混合高斯分布，使用EM算法进行求解



for:

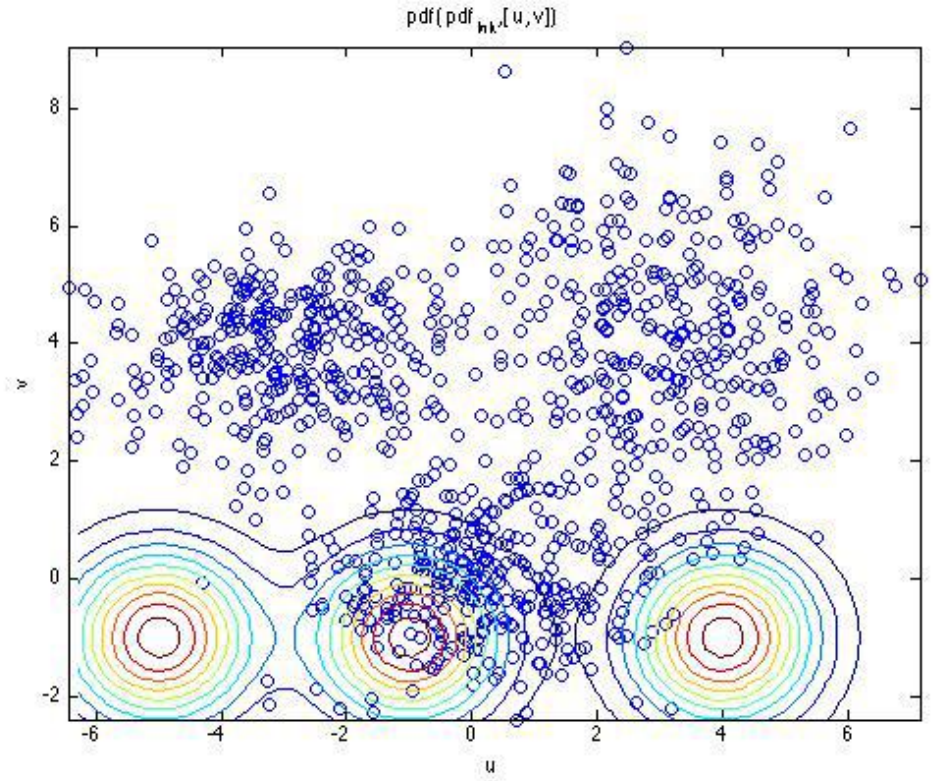
E步：

确定$Q(\theta,\theta^{(i)})$

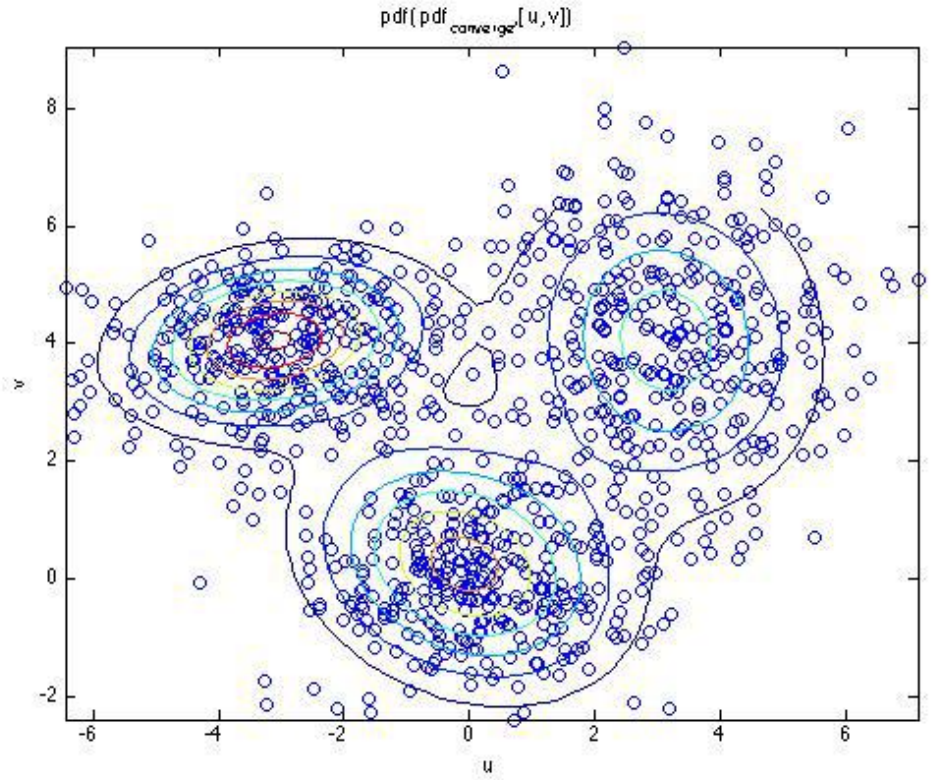
M步：

$$\theta^{(i+1)}=\arg\max_{\theta}Q(\theta,\theta^{(i)})$$

用EM算法估计高斯混合模型参数（可视化）



用EM算法估计高斯混合模型参数（可视化）



Sklearn之GMM实战——根据身高体重进行人群聚类

Sex	Height(cm	Weight(kg)
0	156	50
0	160	60
0	162	54
0	162	55
0	160.5	56
0	160	53
0	158	55
0	164	60
0	165	50

高斯混合模型和Kmeans对比

- Kmeans是简单的，因为它是基于假设一个样本仅以1或0的概率属于某一簇，这两者之间的取值并没有考虑，他无法考虑中间的取值，即一个点仅以某个概率属于某个类别是不能计算的。
- 高斯混合模型不是简单的考虑欧式距离的问题，它是使用高斯概率密度函数（正态分布曲线）精确地量化事物，它是一个将事物分解为若干的基于高斯概率密度函数（正态分布曲线）形成的模型。 高斯模型就是用高斯概率密度函数（正态分布曲线）精确地量化事物

THANK YOU

上海育创网络科技有限公司