

Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心



Deep Gaussian Mixture Models

Deep Gaussian Mixture Models

硕士研究生 王殿元

2019年10月27日

- 背景简介
- 基本概念
 - 高斯混合模型回顾
 - 深度学习
- 算法原理
- 优劣分析
- 应用总结
- 参考文献



- 预期收获
 - 1. 回顾高斯分布、混合模型等基本概念
 - 2. 了解深度学习在聚类上的应用
 - 3. 了解深度高斯混合模型思想和原理



背景简介

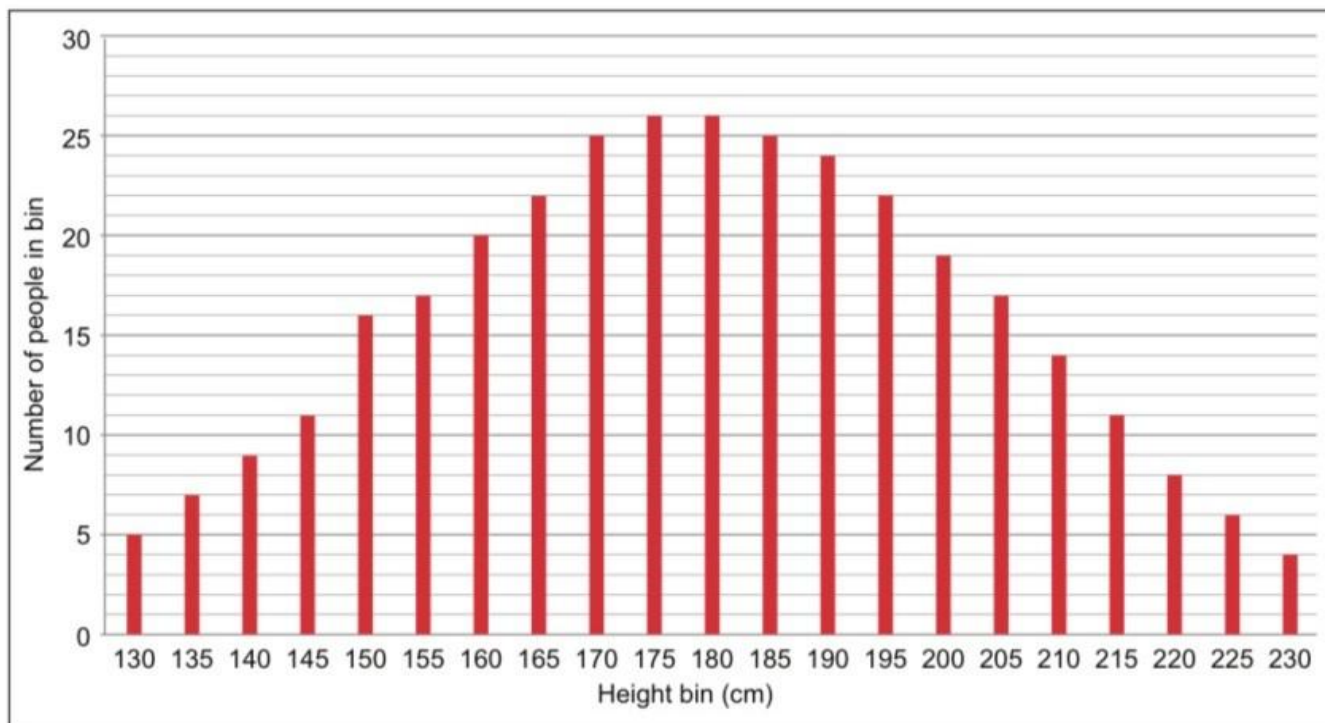
- 算法提出的原因
 - 深度学习在监督分类方面取得了巨大的成功
 - DeepFace (9层神经网络, 1.2亿个连接权值, 识别准确率97.35%)
 - 语音识别、目标检测、模式识别、图像处理等多领域监督分类任务
 - 深度学习在无监督聚类方面的研究有限



基本概念

- 高斯分布

高斯分布 (Gaussian distribution) 也称正态分布 (normal distribution), 是一种在自然界大量的存在的、最为常见的分布形式。



334个成年人的身高统计

分布形式呈高斯分布

- 混合模型

几种不同模型组合而成的一种模型。**线性混合模型**、**高斯混合模型**等...

- 数据异质性

数据之间存在的差异，包括可观测的和不可观测的。

总体的数据可能由多个不同的子总体混合而成，比如男性和女性、城镇居民与农村居民。

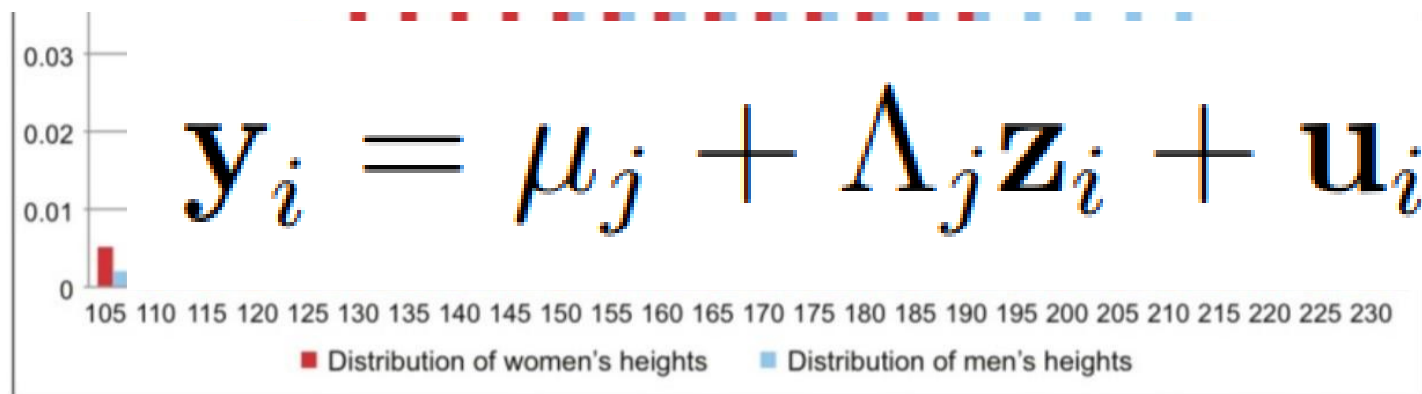
- 隐变量 (Hidden Variables)

对应**可观察变量**，是不直接观察但是通过观察到的其他变量**推断**的变量。

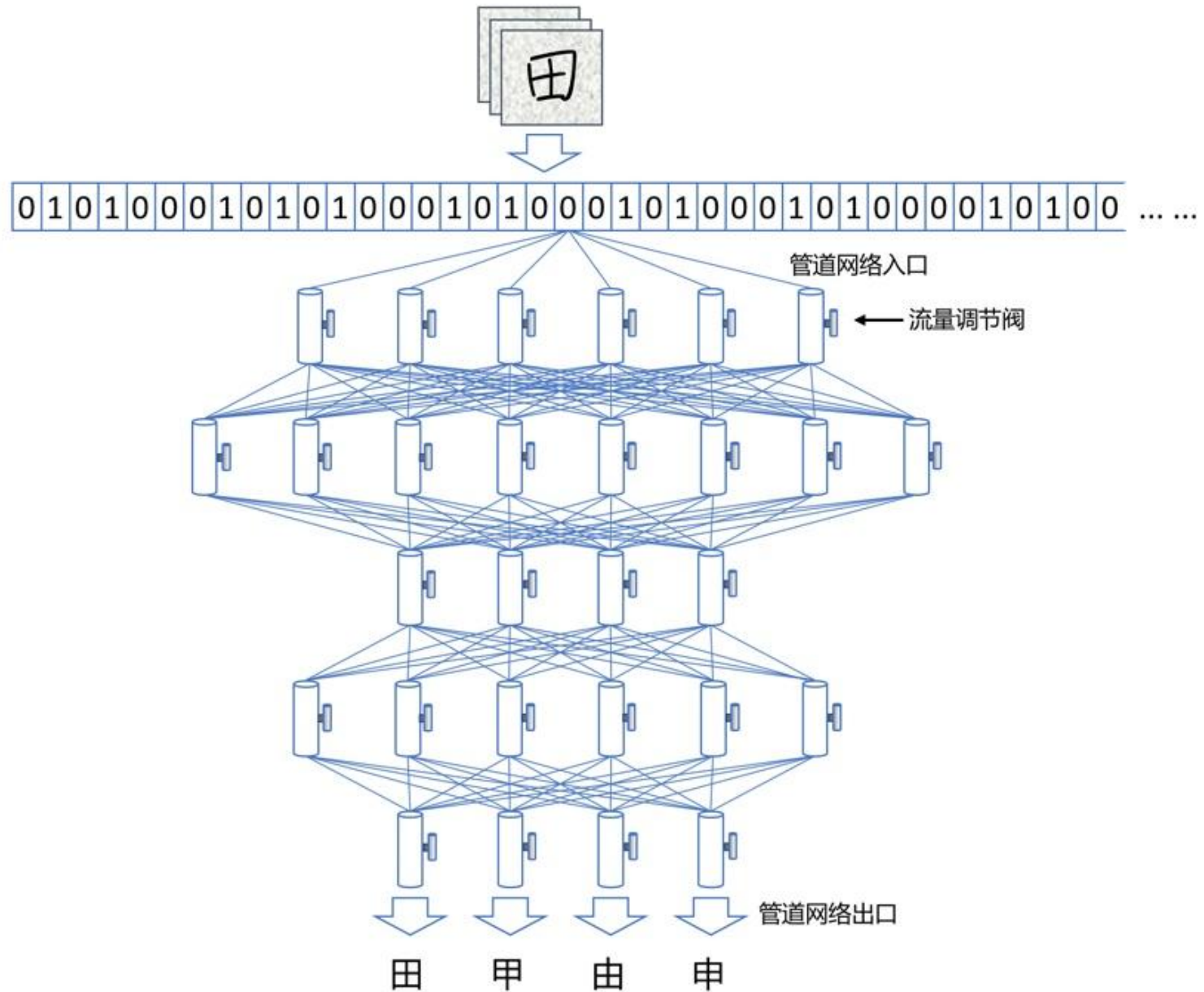
- 高斯混合模型

高斯混合模型 (GMM) 是对高斯分布进行简单的扩展，使用多个高斯分布的组合来刻画数据分布。

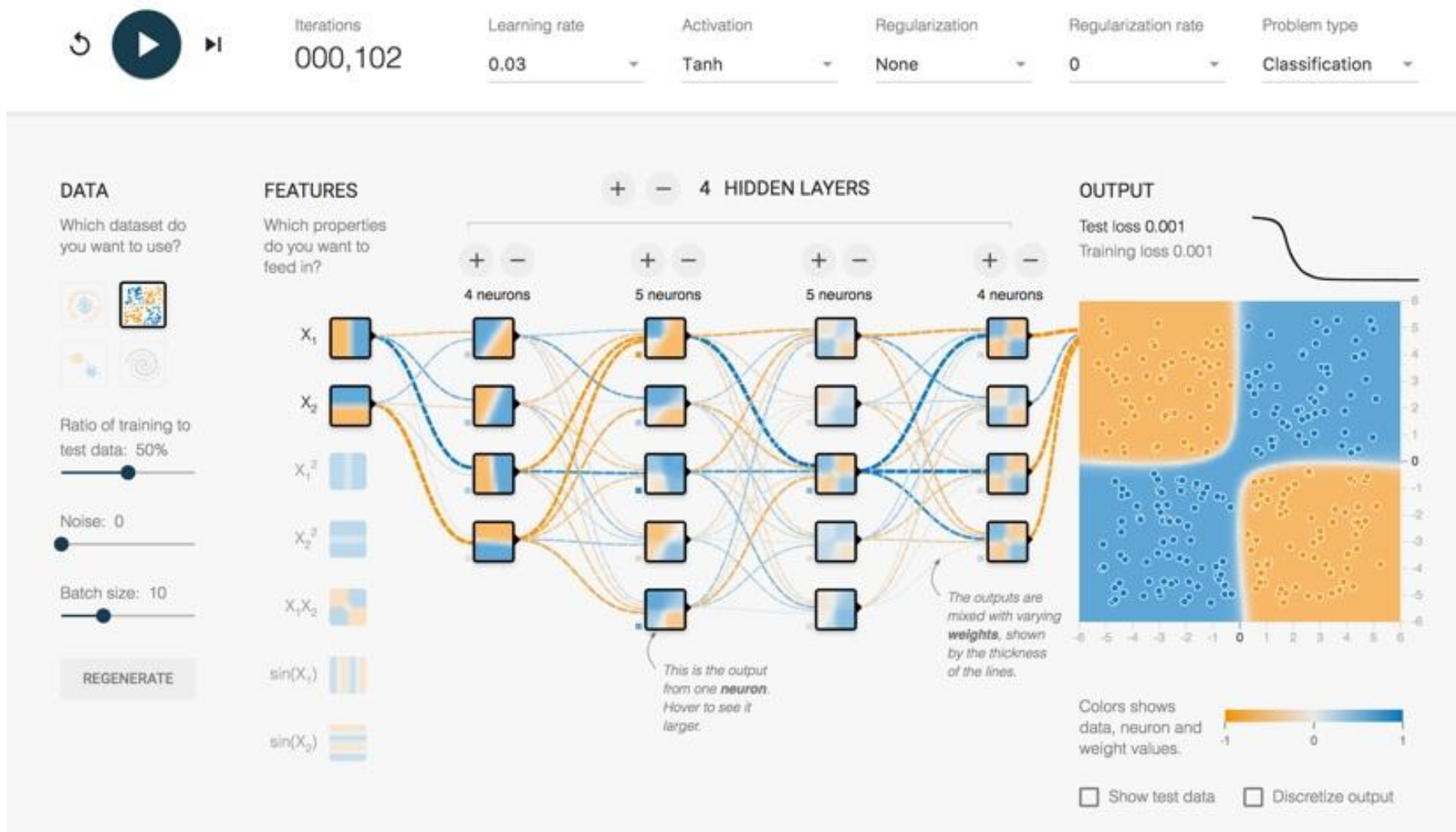
$$f(\mathbf{y} | \Sigma_j = \Lambda_j \Lambda_j^\top + \Psi_j, \Sigma_j) = \prod_{j=1}^k$$



- 深度学习模型



- 可视化软件——Tensorflow — Neural Network Playground



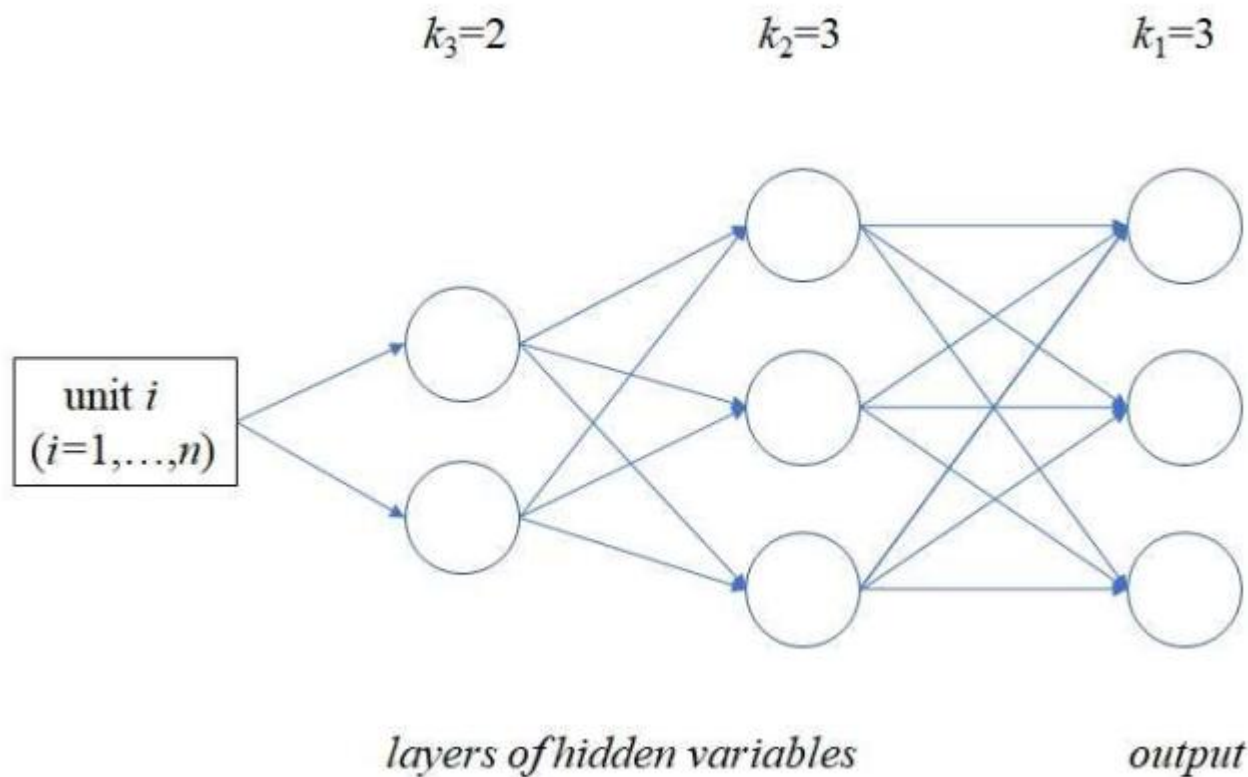


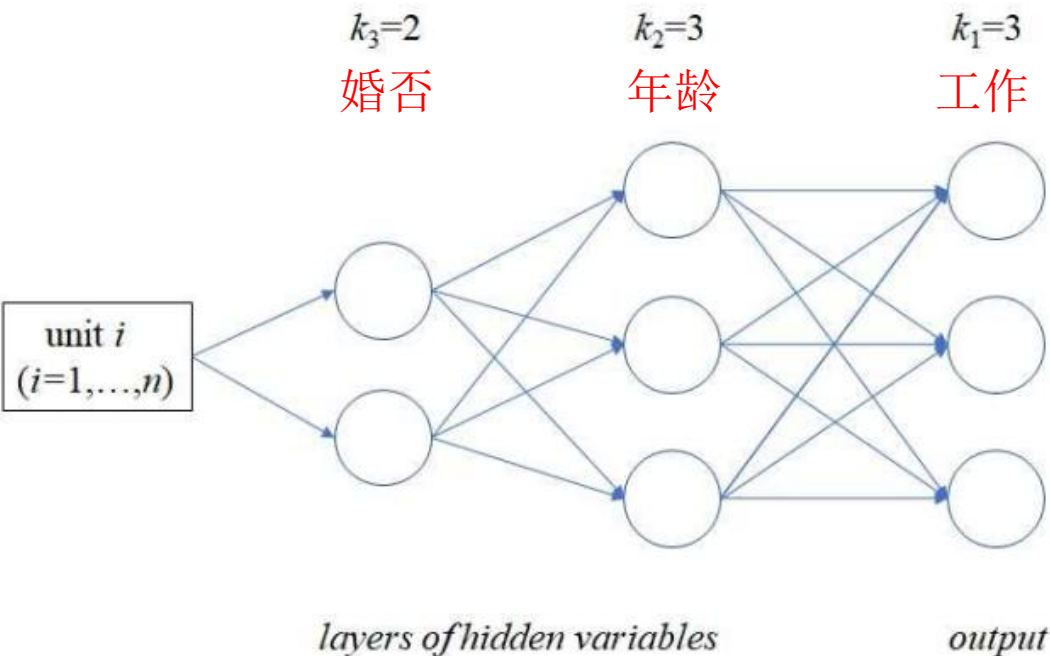
算法原理

T	降低多重隐变量对异质数据聚类效果的影响
I	异质数据
P	Deep Gaussian Cluster
O	高性能深度聚类模型

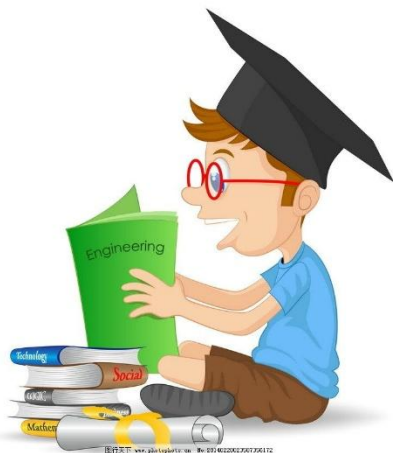
P	如何划分最优的深度模型层数以及每层的变量数
C	数据基本服从高斯分布，存在2个以上隐变量
D	多重隐变量的可解释性
L	2019NIPS

- 深度高斯混合模型
 - 由多层隐变量构成的网络，每一层的变量都服从高斯分布
 - 由一组嵌套混合的线性模型组成，这些模型全局地提供了一个非线性模型





姓名	性别	...	籍贯	购买力
小赵	男	...	BJ	1000
小钱	女	...	BC	50
小孙	女	...	SY	222
小李	男	...	SH	8000
...	...	...	...	...

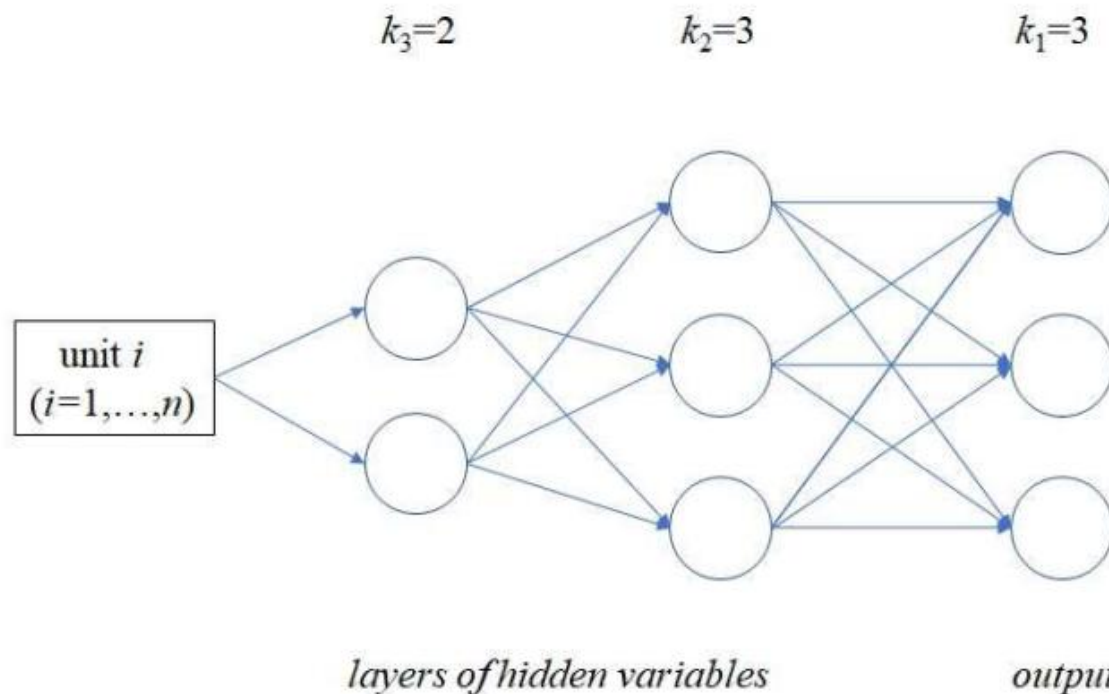


未婚
青年
在校学生

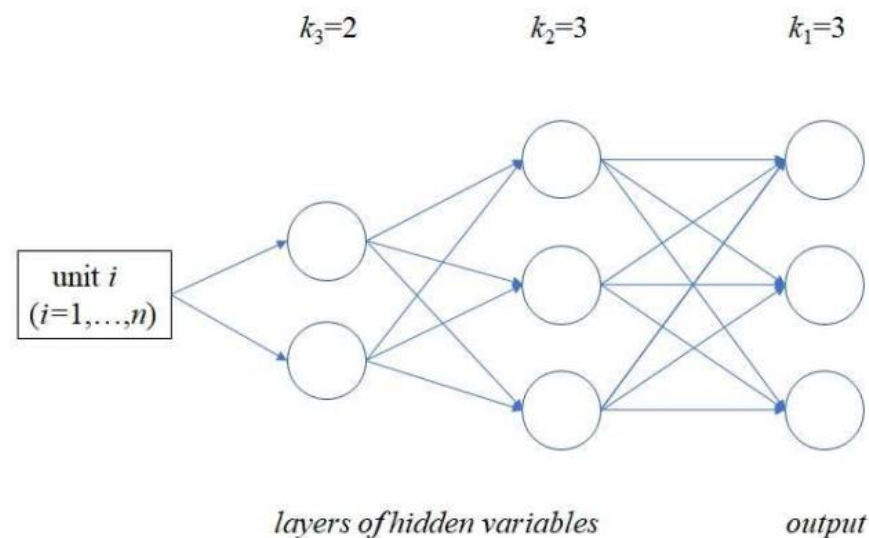


已婚
老年
退休干部





$$\begin{aligned}
 \mathbf{y}_i &= \eta_{s_1}^{(1)} + \Lambda_{s_1}^{(1)} \mathbf{z}_i^{(1)} + \mathbf{u}_i^{(1)} \text{ with prob. } \pi_{s_1}^{(1)}, \quad s_1 = 1, \dots, k_1, \\
 \mathbf{z}_i^{(1)} &= \eta_{s_2}^{(2)} + \Lambda_{s_2}^{(2)} \mathbf{z}_i^{(2)} + \mathbf{u}_i^{(2)} \text{ with prob. } \pi_{s_2}^{(2)}, \quad s_2 = 1, \dots, k_2, \\
 &\dots \\
 \mathbf{z}_i^{(h-1)} &= \eta_{s_h}^{(h)} + \Lambda_{s_h}^{(h)} \mathbf{z}_i^{(h)} + \mathbf{u}_i^{(h)} \text{ with prob. } \pi_{s_h}^{(h)}, \quad t = 1, \dots, k_h
 \end{aligned}$$



$$\sum_{s \in \Omega} \pi_s = \sum_{s_1, \dots, s_h} \pi_{(s_1, \dots, s_h)} = 1$$

$$\begin{aligned} \mu_s &= \eta_{s_1}^{(1)} + \Lambda_{s_1}^{(1)} (\eta_{s_2}^{(2)} + \Lambda_{s_2}^{(2)} (\dots (\eta_{s_{h-1}}^{(h-1)} + \Lambda_{s_{h-1}}^{(h-1)} \eta_h^{(h)}))) \\ &= \eta_{s_1}^{(1)} + \sum_{l=2}^h \left(\prod_{m=1}^{l-1} \Lambda_{s_m}^{(m)} \right) \eta_{s_l}^{(l)} \end{aligned}$$

$$f(\mathbf{y}; \Theta) = \sum_{s \in \Omega} \pi_s N(\mathbf{y}; \mu_s, \Sigma_s)$$

$$\begin{aligned} \Sigma_s &= \Psi_{s_1}^{(1)} + \Lambda_{s_1}^{(1)} (\Lambda_{s_2}^{(2)} (\dots (\Lambda_{s_h}^{(h)} \Lambda_{s_h}^{(h)\top} + \Psi_{s_h}^{(h)}) \dots) \Lambda_{s_2}^{(2)\top}) \Lambda_{s_1}^{(1)\top} \\ &= \Psi_{s_1}^{(1)} + \sum_{l=2}^h \left(\prod_{m=1}^{l-1} \Lambda_{s_m}^{(m)} \right) \Psi_{s_l}^{(l)} \left(\prod_{m=1}^{l-1} \Lambda_{s_m}^{(m)} \right)^\top. \end{aligned}$$

- 评价指标
 - ARI (Adjusted Rand Index) :

$$RI = \frac{a+b}{C_2^{n_{\text{samples}}}}$$

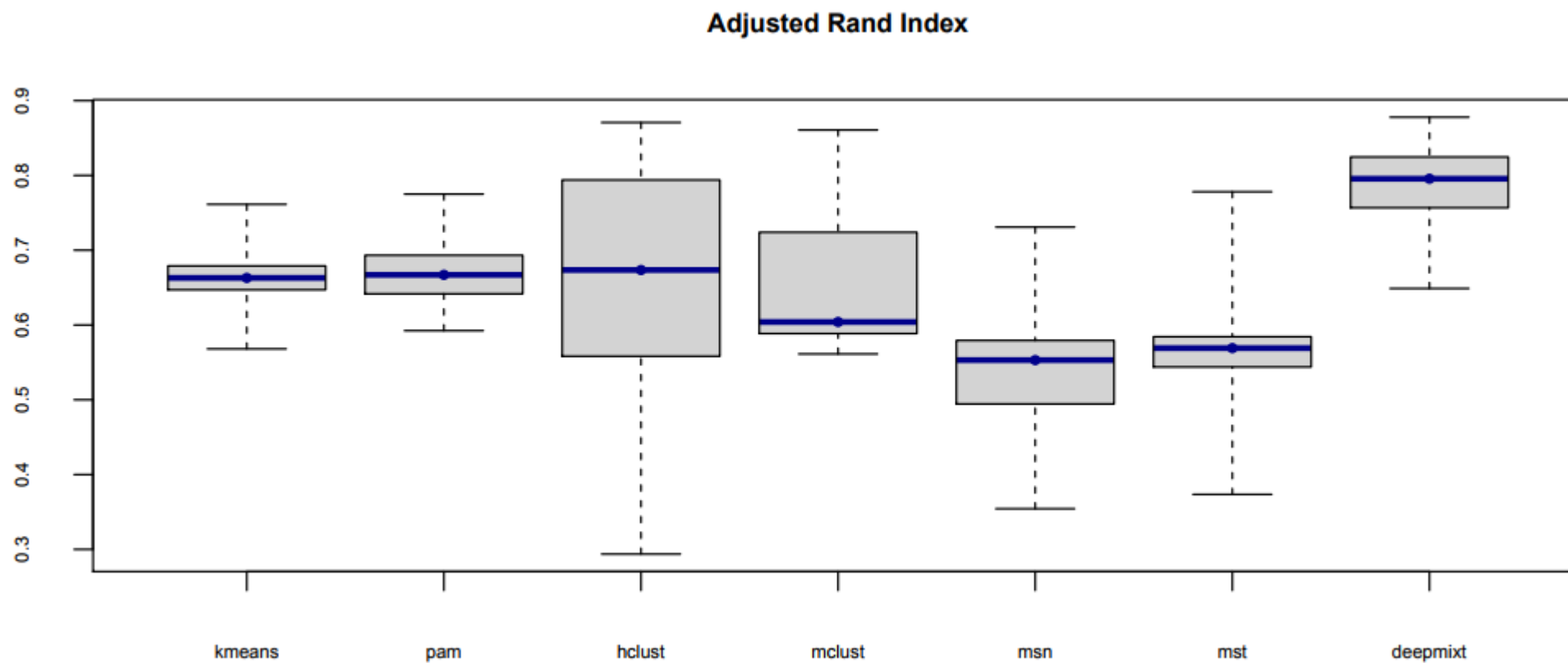
a是真实在同一类、预测也在同一类的样本数；

b是真实在不同类、预测也在不同类的样本数；

为了实现“在聚类结果随机产生的情况下，指标应该接近零”

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

- 评价指标
 - ARI (Adjusted Rand Index) :



- 实验结果
 - 对于较小的数据集（Wine, Olive, Ecoli, Vehicle, Satellite），构建 $h=2$ 的深度模型即可，通过BIC可以确定每一层的最佳簇数。

<i>Dataset</i>	<i>Wine</i>		<i>Olive</i>		<i>Ecoli</i>		<i>Vehicle</i>		<i>Satellite</i>	
	ARI	m.r.	ARI	m.r.	ARI	m.r.	ARI	m.r.	ARI	m.r.
<i>k</i> -means	0.930	0.022	0.448	0.234	0.548	0.298	0.071	0.629	0.529	0.277
PAM	0.863	0.045	0.725	0.107	0.507	0.330	0.073	0.619	0.531	0.292
Hclust	0.865	0.045	0.493	0.215	0.518	0.330	0.092	0.623	0.446	0.337
GMM	0.917	0.028	0.535	0.195	0.395	0.414	0.089	0.621	0.461	0.374
SNmm	0.964	0.011	0.816	0.168	-	-	0.125	0.566	0.440	0.390
STmm	0.085	0.511	0.811	0.171	-	-	0.171	0.587	0.463	0.390
FMA	0.361	0.303	0.706	0.213	0.222	0.586	0.093	0.595	0.367	0.426
MFA	0.983	0.006	0.914	0.052	0.525	0.330	0.090	0.626	0.589	0.243
DGMM	0.983	0.006	0.997	0.002	0.749	0.187	0.191	0.481	0.604	0.249

- 横向对比
 - 优：同传统的GMM相比需要调参的数量减少，同k-means等模型相比准确率更高
 - 劣：可解释性存疑，隐变量的优先级没有体现在计算中
- 纵向对比
 - 优：是深度学习应用于无监督聚类的一次成功尝试
 - 劣：更多的数据+有监督学习 > 无监督学习

- 算法的应用领域
 - 框架+思想
- 未来的发展
 - 深度学习+无监督聚类
 - 对于多层隐变量模型的可解释性以及隐变量重要性定义可深入研究

- [1]Deep Gaussian Mixture Models. Cinzia Viroli , Geoffrey J. McLachlan
- [2]Deep Learning. Yann LeCun , Yoshua Bengio , Geoffrey Hinton



图虫创意 stock.tuchong.com



上善若水。水善利万物
而不争，处众人之所恶，
故几於道。居善地，心
善渊与善仁，言善信，
正善治，事善能，动善
时。夫唯不争，故无尤。

谢谢！

