概率图模型理论与方法

课程大作业说明

郭文博

gwb17@mails.tsinghua.edu.cn

清华大学 自动化系

2021年11月

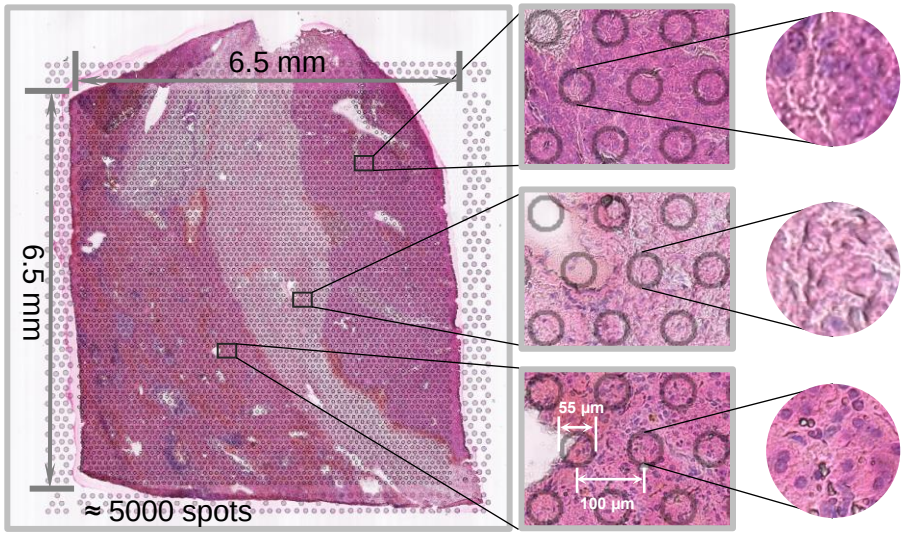
数据 & 任务

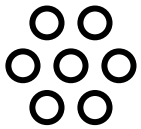
数据：

- 对于一个组织切片，我们有两种模态的数据：
  1) 病理图像（Hematoxylin and Eosin (H&E) Staining Image）
  2) 空间位点-基因的表达矩阵（几千个基因在每个空间位点的这一小块组织上的基因表达量）

任务:

- 任务 1：基于贝叶斯网络对空间位点进行聚类
- 任务 2：基于马尔科夫网络对病理图像进行切割





如左图所示：

1）紫色的图案为病理图像；

2）图像上每个小圆圈为一个空间位点，里面包含了十多个细胞，空间位点按六边形排列。

任务 1：基于贝叶斯网络对空间位点进行聚类

数据文件：

1. **基因表达矩阵** (*Brain-expr_matrix.txt*)

- 行是基因（2000个），列是空间位点(spot)（2264个），数值为该基因在该位点的表达量（已完成数据标准化等预处理，并取整），可将spot视作文档，基因视作单词，表达量视作单词出现的频次，主题则为spot的类别标签

2. **空间位点的位置信息** (*Brain-spot_position.txt*)

- 三列：空间位点名 (barcode)；横坐标 (x)；纵坐标 (y)

3. **小数据** (*Brain-expr_matrix-smallData.txt, Brain-spot_position-smallData.txt*)

- 6个基因，20个spot，文件含义同上，仅用于实践PyMC3时使用

Brain-expr_matrix.txt

1	TCGGCTAACTTCCCTT-1	ATATCGTGCCAGACCC-1	ACGCTTAGTGTCTCTC-1											
2	Pmch	0	0	0	0	0	0	0	0	0	0	1	0	
3	Ttr	2	2	1	3	2	2	2	2	2	2	1	2	1	1	
4	Hbb-bs	2	2	3	2	3	6	6	2	0	5	0	2	0	
5	Mbp	3	2	3	3	3	3	3	3	3	2	2	3	2	2	
6	Plp1	2	2	2	3	3	2	2	4	3	2	2	3	3	2	
7	Prkcd	2	2	0	2	0	1	2	2	0	0	0	1	1	
8	Hba-a2	1	2	2	2	3	5	5	0	0	4	1	1	2	
9	Mobp	2	1	2	2	2	2	2	2	2	2	1	2	2	
10	Ptgds	6	6	4	3	5	6	6	5	3	6	2	5	1	
11	Hcrt	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	Penk	2	1	2	2	1	2	2	2	2	1	1	1	2	
13	Ppp1r1b	2	1	2	2	2	1	2	1	2	1	2	2	2	
14	Sst	0	2	2	2	2	2	3	1	2	2	2	3	2	3	
15	Pcp4	2	2	3	1	3	2	2	2	2	2	2	0	2	1	
16	Pvalb	0	1	0	0	0	0	0	0	2	0	2	0	2	2	
17	Tac2	0	0	0	1	1	0	2	1	0	0	1	0	0	
18	Hbb-bt	1	0	0	0	0	4	3	0	0	0	0	0	0	
19	Camk2n1	5	5	6	5	5	5	5	5	6	5	6	6	6	
20	C1ql2	0	0	0	0	0	0	0	0	0	0	0	0	0	

Brain-spot_position.txt

1	barcode	x	y
2	TCGGCTAACTTCCCTT-1	98.726896031426	50
3	ATATCGTGCCAGACCC-1	98.726896031426	48
4	ACGCTTAGTGTCTCTC-1	98.726896031426	46
5	TAGTGGAACTCATACA-1	98.726896031426	44
6	GTTATTAAATACGACC-1	98.726896031426	42
7	AACTGGGTTCGAGCCG-1	98.726896031426	38
8	CATGCACGTGTTACTG-1	98.726896031426	36
9	CATTCACTGACAGCTA-1	95.2627944162882	60
10	CATGAACCTCTTATCA-1	95.2627944162882	58
11	TTAATGCGAGGTAACT-1	96.9948452238571	57
12	AATAAGTCCTCGAGAC-1	95.2627944162882	56
13	ACCAGCCCGGTCTTTG-1	96.9948452238571	55
14	CTACGAACTAGGTCGA-1	95.2627944162882	54
15	ACATCTCAACGCGTAA-1	96.9948452238571	53
16	CACTACTCAGTTCTGT-1	95.2627944162882	52
17	CCGACTCGCATAGTCT-1	96.9948452238571	51
18	CATTTATCGTTCAAGA-1	95.2627944162882	50
19	CAAACGTGGTCTTGCG-1	96.9948452238571	49
20	TAGAAACCACTAAGTA-1	95.2627944162882	48

small-data.txt

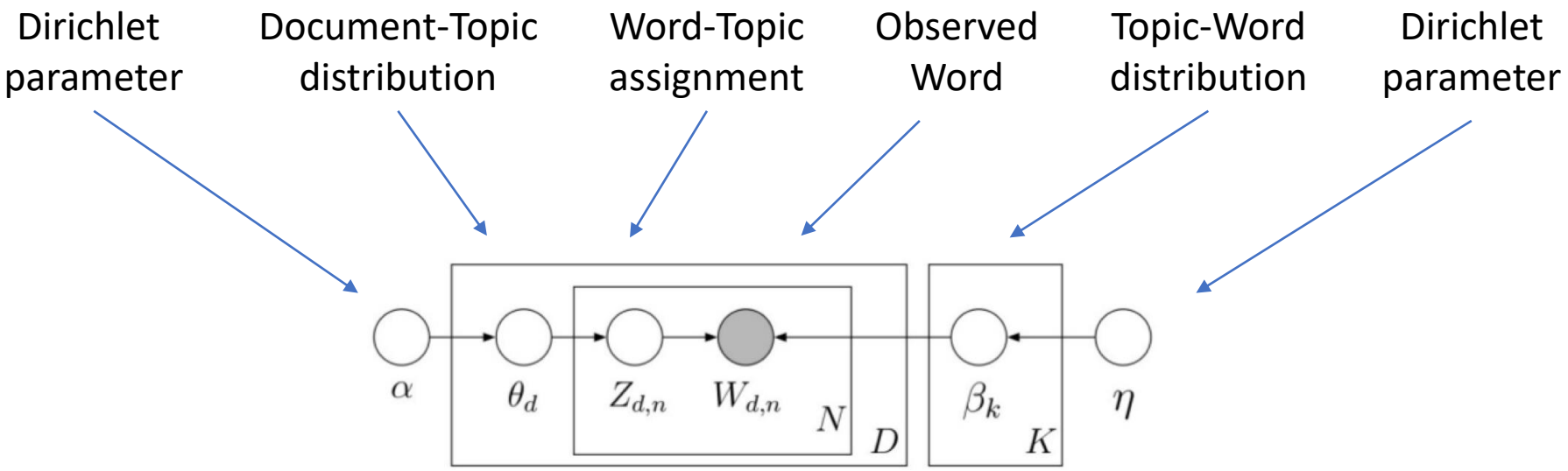
1		AACGTCAGACTAGTGG-1	AAGGCGCGTAAAGCTT-1						
2	Hs3st2	0	0	2	0	0	0	1	0	1
3	Pcp4	3	2	4	4	3	2	3	1	3
4	Nefm	2	2	3	3	2	0	3	3	3
5	Negr1	3	3	2	2	3	3	3	3	3
6	Cux2	2	2	0	0	2	3	2	2	1
7	Rorb	2	2	1	1	3	2	2	3	2

Brain-spot_position-smallData.txt

1	barcode	x	y
2	AACGTCAGACTAGTGG-1	62.3538290724796	87
3	AAGGCGCGTAAAGCTT-1	62.3538290724796	89
4	AATGACTGTCAGCCGG-1	64.0858798800485	80
5	ACACCTTAAGTAGGGC-1	60.6217782649107	84
6	ACTACCAGCTCTCTGG-1	60.6217782649107	86
7	AGTCGGCCCAAACGAC-1	60.6217782649107	88
8	ATCTTGACCTGCAACG-1	58.8897274573418	83
9	ATTCCTCCGCCAGTGC-1	64.0858798800485	88
10	CAGATACTAACATAGT-1	58.8897274573418	81

任务 1：基于贝叶斯网络对空间位点进行聚类

Latent Dirichlet Allocation (LDA)



Multinomial distribution

$$P(x_1, x_2, \ldots, x_k; n, p_1, p_2, \ldots, p_k) = \frac{n!}{x_1! \ldots x_k!} {p_1}^{x_1} \ldots {p_k}^{x_k}$$

Dirichlet distribution

$$f(x_1, x_2, \ldots, x_k; \alpha_1, \alpha_2, \ldots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} {x_i}^{\alpha^i - 1}$$

任务 1：基于贝叶斯网络对空间位点进行聚类

参考资料（可对照下页要求）：

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[2] Yalamanchili H B, Kho S J, Raymer M L. Latent dirichlet allocation for classification using gene expression data. 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, 2017: 39-44.

[3] Kho S J, Yalamanchili H B, Raymer M L, et al. A novel approach for classifying gene expression data using topic modeling. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2017: 388-393.

[4] Yuan B, Gao X, Niu Z, et al. Discovering Latent Topics by Gaussian Latent Dirichlet Allocation and Spectral Clustering. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2019, 15(1): 1-18.

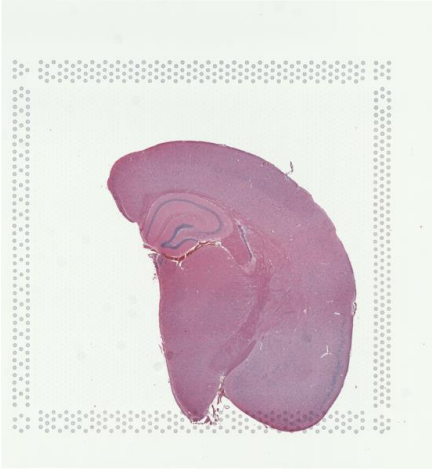
[5] PyMC3 Documentation: https://docs.pymc.io/en/v3/

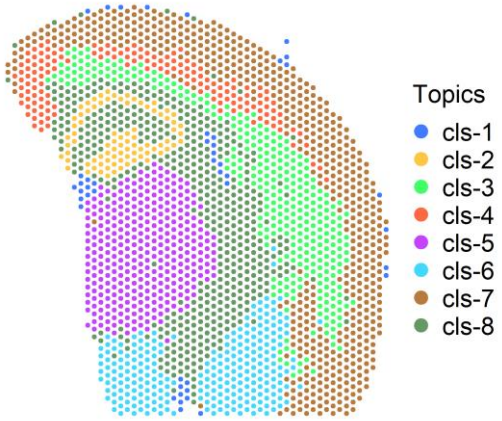
任务 1：基于贝叶斯网络对空间位点进行聚类

要求：

1. 按照概率图模型三段论（variables, relations, local CPDs）的形式给出LDA的表示，概述它的推断、学习算法。（参考资料[1]）
2. 使用已有LDA工具包，对空间位点进行聚类。

- **内容：**1）基于空间位点的基因表达量矩阵（*Brain-expr_matrix.txt*），利用LDA模型对此问题进行建模，实现对空间位点的聚类；2）使用空间位点的位置信息（*Brain-spot_position.txt*）展示聚类结果；3）测试不同参数设置，比较结果。
- 可使用已有的工具（如Python包Gensim），可阅读参考资料[2-4]。
- **样例结果**：



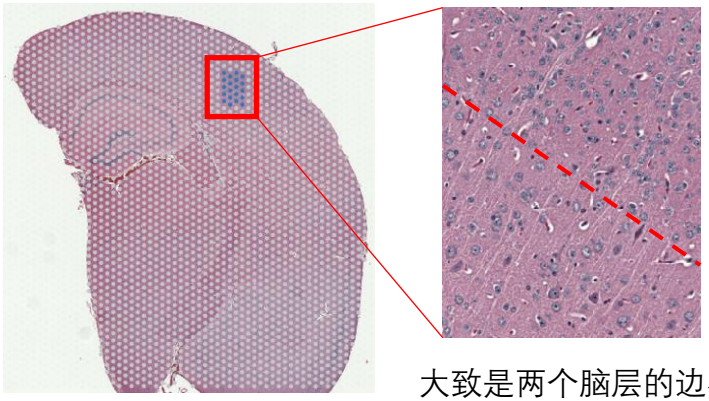


任务 1：基于贝叶斯网络对空间位点进行聚类

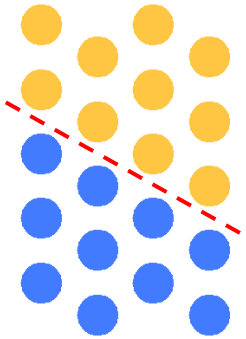
要求：

3. 使用Python包PyMC3实现LDA模型（可学习参考资料[5]）

- **内容**：1）利用PyMC3包搭建LDA模型，并在一个小数据（*Brain-expr_matrix-smallData.txt*）上应用测试；2）使用空间位点的位置信息（*Brain-spot_position-smallData.txt*）展示聚类结果；3）如有余力，可尝试修改LDA模型并应用。
- 已提供一个参考代码框架（*LDA code framework - PyMC3.py*），仅供参考。
- 小数据来源&样例结果：



大致是两个脑层的边界



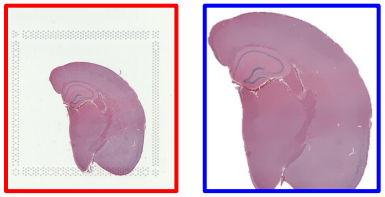
注：

- 对问题背景和内容有疑问，请及时联系助教；
- 经测试，当前数据规模基本可在PC上完成，如确实有计算资源不足等方面的问题，可联系助教。

任务2：基于马尔科夫随机场进行图像分割

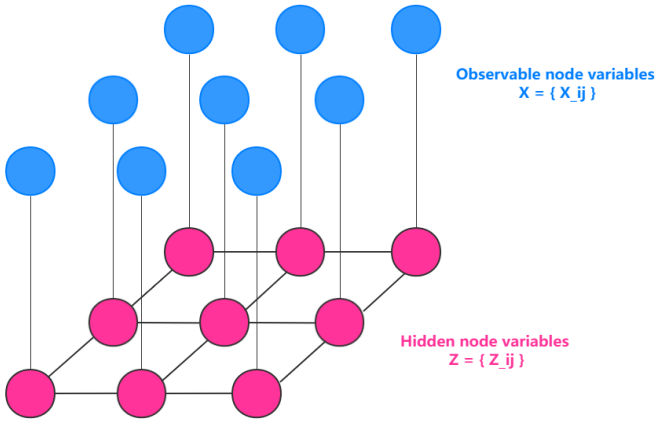
数据

- 组织切片的苏木精-伊红染色图（*raw-HE_image.png*）
- 预处理后的图片（*processed_image.png*，**后续任务使用此图**）



介绍

利用MRF等概率图有关的模型对病理图片进行区域分割



- 观测节点X (灰度信息、RGB信息…)
- 隐节点Z (每个像素的类别标号)
- 建模先验$P(Z)$、似然$P(X|Z)$；求后验$P(Z|X)$

$$P(\boldsymbol{Z}|\boldsymbol{X}) \propto P(\boldsymbol{Z})P(\boldsymbol{X}|\boldsymbol{Z}), \quad \boldsymbol{Z}^* = \arg\max_{Z} P(Z)P(\boldsymbol{X}|\boldsymbol{Z})$$

- $P(\boldsymbol{Z})$是标号场$\boldsymbol{Z}$的先验概率，在此认为其符合MRF模型；
- $P(\boldsymbol{X}|\boldsymbol{Z})$是似然函数，是特征场$\boldsymbol{X}$对于类别标号$Z$的条件概率；
- $P(\boldsymbol{Z}|\boldsymbol{X})$是标号场的后验概率，概率值越大，表示对每个像素点给定的标号越适宜。

任务2：基于马尔科夫随机场进行图像分割

表示

1）根据HC定理，标号场$\boldsymbol{Z}$的概率可以用吉布斯分布表示：

$$P(\boldsymbol{Z}) = \frac{1}{Z_p}\exp(-Q_1(\boldsymbol{Z}))$$

- 能量函数$Q_1(\boldsymbol{Z})$由每个Clique的能量U_c组成，即$Q_1(\boldsymbol{Z}) = \sum_c U_c(\boldsymbol{Z_c})$
- 病理图上的近邻趋向同质，可以设计标号场的先验吉布斯分布$\boldsymbol{P(Z)}$

$$U_c(\boldsymbol{Z_c}) = U_{\{i,j \in E\}}(Z_i, Z_j) = \begin{cases} -\beta, & Z_i = Z_j \\ +\beta, & Z_i \neq Z_j \end{cases}$$

2）观测特征场的条件分布$\boldsymbol{P(X|Z)}$可使用混合高斯模型

$$Q_2(\boldsymbol{X}|\boldsymbol{Z}) = -log\,[P(\boldsymbol{X}|\boldsymbol{Z})] = -\log[N(X|\mu_k, \Sigma_k)]$$

推断　为简化模型，推断步骤可使用 (maximum a posteriori, MAP)

$$\boldsymbol{Z}^* = \underset{Z}{\operatorname{argmax}}\ P(\boldsymbol{Z})P(\boldsymbol{X}|\boldsymbol{Z}) \qquad \boldsymbol{Z}^* = \underset{Z}{\operatorname{argmin}}\ Q_1(\boldsymbol{Z}) + Q_2(\boldsymbol{X}|\boldsymbol{Z})$$

学习　参数学习方法，可参考高斯混合模型 GMM

任务2：基于马尔科夫随机场进行图像分割

内容

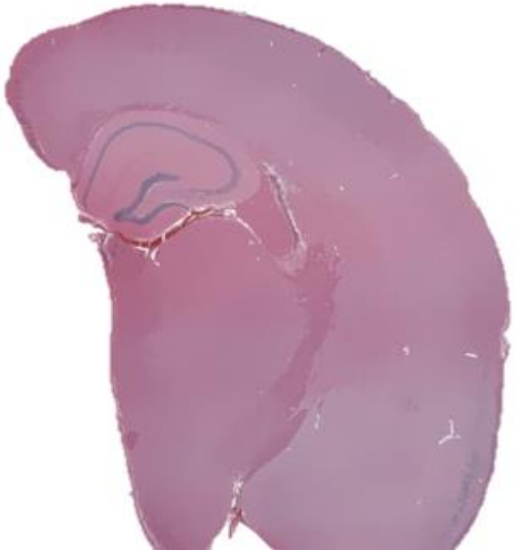
1. 完成概率图模型的表示； 设计类别标号场$\boldsymbol{Z}$的先验概率$P(\boldsymbol{Z})$，以及观测节点的局部条件概率分布$P(\boldsymbol{x}|z)$，并写出标号场的后验概率$P(\boldsymbol{Z}|\boldsymbol{X})$；
2. 根据所学的推断与学习的方法，实现模型的求解，完成病理图片的区域分割（注：为简化模型，推断步骤可使用MAP；参数学习方法，可参考GMM）；
3. 在合理范围内设置不同的类别数K、不同的耦合系数β，观察病理图区域分割结果的变化。

要求

1. 隐马尔科夫随机场的概率图模型表示；
2. 说明所使用的推断方法与参数学习方法；
3. 图片区域分割结果的展示与简单分析；
4. 思考区域分割结果是否受初始化影响，如果更好地初始化。

任务2：基于马尔科夫随机场进行图像分割

样例结果



参考文献

1. 基于马尔科夫随机场的纹理图像分割方法研究
2. 基于MRF-ICM(iteration condition model)的图像分割算法
3. 高斯混合模型，马尔科夫场与吉布斯分布