Toy Data的取法和结果

1. 话题和词汇

Toy Data有两个view。第一个view上有4个，第二个view上有6个topic。每个topic的词汇数量都是20，且概率相同。另外有Wunrelated个词，在所有topic上都有分布，且概率都是单个词汇的1/10。

有一个Ground Truth的A矩阵Agt，其值为：

0.1250 0 0 0 0.1250 0

0 0.1250 0 0 0.1250 0

0 0 0.1250 0 0 0.1250

0 0 0 0.1250 0 0.1250

1. 构造文档

每个object有两个文档，分别用两个view下选定的topic中的词汇，按多项式分布产生，词的数量都是200。Topic选择的联合概率由Agt确定。每篇文档按Agt确定一对topic，作为ground truth。一共有100个objects。

1. 测试

初始化方式对最终结果的影响较大。

1. Wunrelated=1时，最多迭代10次，15次运行的错误率如下。

平均值：

|  |  |  |  |
| --- | --- | --- | --- |
| PLSA初始化的我们的方法 | KMEANS初始化的我们的方法 | PLSA结果 | KMEANS结果 |
| **0.0927** | 0.3650 | 0.1140 | 0.2440 |

每次的值：

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 0.2700 | 0.1100 | 0.0700 |
| 0.0700 | 0.2000 | 0.0900 | 0.4100 |
| 0.1500 | 0.4500 | 0.1200 | 0.2500 |
| 0.1800 | 0.4500 | 0.1100 | 0.3000 |
| 0 | 0.3300 | 0.1100 | 0.1600 |
| 0 | 0.7900 | 0.2400 | 0.2000 |
| 0 | 0.1900 | 0 | 0.3000 |
| 0.0200 | 0.8200 | 0.1300 | 0.1300 |
| 0.2700 | 0.3200 | 0.1100 | 0.2800 |
| 0.1100 | 0.2600 | 0.1300 | 0.4300 |
| 0.2300 | 0.4000 | 0.2600 | 0.4600 |
| 0 | 0.2300 | 0.2200 | 0.1300 |
| 0.2800 | 0.2400 | 0.0800 | 0.1500 |
| 0.0800 | 0.2100 | 0 | 0.2800 |
| 0 | 0.1800 | 0 | 0.1100 |

1. Wunrelated=1时，最多迭代100次，14次运行的错误率如下。

平均值：

|  |  |  |  |
| --- | --- | --- | --- |
| PLSA初始化的我们的方法 | KMEANS初始化的我们的方法 | PLSA结果 | KMEANS结果 |
| **0.1329** | 0.3614 | 0.1557 | 0.2793 | |

PLSA初始化－PLSA相关系数：-0.001504863

KMEANS初始化－KMEANS相关系数：0.197233745

每次的值：

|  |  |  |  |
| --- | --- | --- | --- |
| 0.1000 | 0.1800 | 0.2000 | 0.2000 |
| 0.2500 | 0.2900 | 0 | 0.1300 |
| 0.1000 | 0.3200 | 0.1200 | 0.2300 |
| 0.2000 | 0.8300 | 0.1100 | 0.4300 |
| 0.1000 | 0.8300 | 0.1000 | 0.1600 |
| 0.1100 | 0.2300 | 0.3200 | 0.1100 |
| 0.3500 | 0.5100 | 0.2400 | 0.4100 |
| 0 | 0.4100 | 0 | 0.6300 |
| 0 | 0.1800 | 0.1300 | 0.4200 |
| 0.2100 | 0.2000 | 0 | 0.3200 |
| 0.2100 | 0.1800 | 0.1300 | 0.2800 |
| 0.0700 | 0.2100 | 0.6000 | 0.2300 |
| 0.1600 | 0.2500 | 0.2300 | 0.1500 |
| 0 | 0.4400 | 0 | 0.2100 |

1. Wunrelated=100时的feature selection结果。

以平均值为阈值，正确找出了附加的100个与topic无关的词，但多出了9个词。

1. 与LDA的比较
2. Single view

|  |  |  |
| --- | --- | --- |
| PLSA-MF | PLSA | LDA |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0.29 |
| 0 | 0 | 0.2 |
| 0 | 0 | 0.35 |
| 0 | 0 | 0 |
| 0 | 0 | 0.28 |
| 0 | 0.29 | 0 |
| 0 | 0 | 0.35 |
| 0 | 0 | 0.31 |
| 0 | 0 | 0.18 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0.25 |
| 0 | 0 | 0.24 |
| 0 | 0 | 0.3 |
| 0 | 0 | 0 |
| 0 | 0 | 0.3 |
| 0 | 0 | 0 |
| *0* | *0.0145* | *0.1525* |

1. Double views

|  |  |  |  |
| --- | --- | --- | --- |
| PLSA-MF | LDA-MF | PLSA | LDA |
| 0 | 0.24 | 0 | 0.12 |
| 0 | 0.13 | 0 | 0.11 |
| 0 | 0 | 0 | 0.12 |
| 0.16 | 0.15 | 0 | 0.35 |
| 0 | 0.28 | 0 | 0.16 |
| 0.22 | 0 | 0 | 0.28 |
| 0.26 | 0.21 | 0.49 | 0.21 |
| 0.17 | 0.82 | 0.3 | 0.43 |
| 0.21 | 0 | 0.18 | 0.2 |
| 0.2 | 0.1 | 0.01 | 0.36 |
| 0.01 | 0.01 | 0 | 0.33 |
| 0 | 0 | 0 | 0.08 |
| 0.21 | 0.84 | 0 | 0.44 |
| 0.05 | 0.82 | 0.12 | 0.4 |
| 0 | 0.16 | 0 | 0.33 |
| 0 | 0.26 | 0 | 0.18 |
| 0.24 | 0 | 0.14 | 0.05 |
| 0 | 0.83 | 0 | 0.68 |
| 0.3 | 0.07 | 0.24 | 0.26 |
| 0.26 | 0.27 | 0.4 | 0.43 |
| *0.1145* | *0.2595* | *0.094* | *0.276* |

1. PLSA-MF与PLSA在干扰项（Wunrelated=1000）时的表现

|  |  |
| --- | --- |
| PLSA-MF | PLSA |
| 0.06 | 0.66 |
| 0.18 | 0.15 |
| 0.13 | 0.05 |
| 0.09 | 0.11 |
| 0.06 | 0.03 |
| 0.07 | 0.59 |
| 0.04 | 0.61 |
| 0.02 | 0.13 |
| 0 | 0 |
| 0.09 | 0.62 |
| 0 | 0 |
| 0 | 0.04 |
| 0.11 | 0.62 |
| 0.07 | 0.67 |
| 0.15 | 0.25 |
| 0 | 0.02 |
| 0.06 | 0.64 |
| 0.08 | 0.8 |
| 0.11 | 0.1 |
| 0 | 0.05 |
| *0.066* | *0.307* |