

主题模型

解决一词多义/多词一义的不足

①应用：语义分析，文档分类/聚类，文章摘要，社区挖掘

基于内容的图像聚类，目标识别，生物信息数据应用，

用主题模型做特征降维：eg N 个词的特征空间 $\xrightarrow{\text{主题模型}}$ 20个主题的特征空间

带权重的主题模型，当样本通常偏向于1-2个少数主题时，认为是一个soft聚类

(文档在20个主题上的权重情况 \rightarrow one-hot 编码)

②输入输出 (20个主题为例)

输入：从 N 篇文档提出的 N 个样本，每个样本都是一个词袋 (用 TF-IDF 等预处理得到)

输出：每篇文档在 20 个主题上的分布 $\xrightarrow{\text{主题模型}} \leftarrow \text{文档 } i$

每种文档类型在 20 个主题上的分布 $\xrightarrow{\text{主题模型}} \leftarrow \text{武侠}$

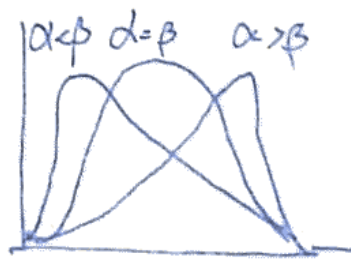
20 个主题，以及代表每个主题的关键词列表

③用到的数学知识：

* Γ 函数 (阶乘在实数域上的推广)：见概率

* Beta 分布：
$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & x \in [0, 1] \\ 0 & \text{其它} \end{cases}$$
 是两点/二项分布的共轭先验分布。

$$E(x) = \frac{\alpha}{\alpha+\beta}$$



* 共轭先验分布，先验概率，后验概率，“证据”，似然概率 (贝叶斯概率论)

参数 θ 的样本 x 的集合 $\{x_1, x_2, \dots, x_N\}$

$$P(\theta|x) = \frac{P(x|\theta) P(\theta)}{P(x)}$$

后验概率：从样本计算参数 θ 的概率
似然概率：给定参数 θ 取值，计算样本发生的概率
先验概率：参数 θ 取值不受 θ 影响时，样本发生的概率
“证据”归一化因子 (与 θ 无关的常数)

当后验概率与先验概率分布相同 $P(\theta|x) = P(\theta)$ “同分布”

<a> $P(\theta|x)$ 与 $P(\theta)$ 是共轭分布
 $P(\theta)$ 先验分布被叫作 $P(x|\theta)$ 的共轭先验分布

用“二项分布”建模，似然概率就是二项分布
用“高斯分布”建模，似然概率就是高斯分布

④ 共轭先验分布应用于二分类问题

与 θ 无关
↑
将 $P(x)$ 当作常数消去

后验 似然 先验 证据 正比于

$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)} \propto P(x|\theta) P(\theta)$$

二分类问题用二项分布建模

$$P(x|\theta) = C_n^k \theta^k (1-\theta)^{n-k}$$

用 Beta 分布来建模

$$P(\theta|\alpha, \beta) = \begin{cases} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} & \theta \in [0, 1] \\ 0 & \text{其它} \end{cases}$$

后验分布也是 Beta 分布，推导如下

$$\begin{aligned}
 P(\theta|x) &= P(x|\theta) \cdot P(\theta) / P(x) \propto P(x|\theta) P(\theta) \\
 &= (C_n^k \theta^k (1-\theta)^{n-k}) \left(\frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \right) \\
 &= \left[\frac{C_n^k}{B(\alpha, \beta)} \right] \theta^{(k+\alpha)-1} (1-\theta)^{(n-k+\beta)-1} \\
 &\quad \left[\frac{C_n^k}{B(\alpha, \beta)} \right] \text{ 成正比, 因为是两个常数} \\
 &\propto \left[\frac{1}{B(k+\alpha, n-k+\beta)} \right] \theta^{(k+\alpha)-1} (1-\theta)^{(n-k+\beta)-1}
 \end{aligned}$$

共轭分布

- * 先验分布：服从 Beta 分布，数学期望是 $\frac{\alpha}{\alpha+\beta}$ → 是二项分布的共轭先验分布
- * 后验分布：服从 Beta 分布，数学期望是 $\frac{k+\alpha}{n+\alpha+\beta}$

* 如何理解上面的概率

先验分布：例如男生占总人数比例为 50%。即 $\alpha=500, \beta=500$ $\frac{\alpha}{\alpha+\beta} = 50\%$

↑
事先给定的超参数，是伪计数

似然概率：作了 n 次独立实验，共观察到 k 个男生，例如 $n=200, k=150$

后验分布：数学期望就是 $\frac{k+\alpha}{n+\alpha+\beta} = \frac{150+500}{200+500+500}$

↑
来自样本

* 在似然概率中，样本越多，~~结果~~后验分布越接近 n 次试验得到的概率
样本越少，越接近来自经验的先验概率。

⑤ 共轭先验分布用于多分类问题

$$P(\theta|x) \stackrel{\text{后验}}{=} \stackrel{\text{似然}}{P(x|\theta)} \cdot \stackrel{\text{先验}}{P(\theta)} \stackrel{\text{证据}}{=} \frac{P(x|\theta) \cdot P(\theta)}{P(x)} \propto P(x|\theta) \cdot P(\theta)$$

多分类问题用多项分布建模

$$P(x=n_1, \dots, x_k=n_k) = \begin{cases} n! \prod_{i=1}^k \frac{p_i^{n_i}}{n_i!}, & \sum_{i=1}^k n_i = n \\ 0, & \text{otherwise} \end{cases}$$

将样本喂给似然函数, 得到
在k个类别上样本出现的次数

$$\vec{C} = \{C_1, C_2, \dots, C_k\}$$

Dirichlet分布是多项分布的共轭先验分布
因此后验分布也服从Dirichlet, 略去推
导过程, 得到后验分布

$$P(\theta|x) \text{ 服从 } \text{Dir}(K, \vec{\alpha} + \vec{C})$$

数学期望

$$E(p_i) = (\alpha_i + C_i) / \sum_{i=1}^k (\alpha_k + C_k)$$

$$\text{Dir}(K, \vec{\alpha} + \vec{C})$$

$$= \text{Dir}(K, C_1 + \alpha_1, \dots, C_k + \alpha_k)$$

⑥ 二项分布 \rightarrow 多项分布

Beta分布 \rightarrow Dirichlet分布

是从二分类到多分类的扩展

用Dirichlet分布来建模

$$\text{Dir}(\vec{p}|\vec{\alpha}) = \begin{cases} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^K p_k^{\alpha_k-1}, & p_k \in [0,1] \\ 0, & \text{其它} \end{cases}$$

$$\text{其中 } \Delta(\vec{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$$

数学期望

$$E(p_i) = \alpha_i / \sum_{i=1}^K \alpha_k$$

记为 $\text{Dir}(K, \vec{\alpha})$

$\vec{\alpha}$ 调节 (见PPT)

将 $\vec{\alpha}$ 简化为对称Dirichlet分布的
情况, 即 $\alpha_1 = \alpha_2 = \dots = \alpha_k$ 作
函数图, 得到:

① α_i 值越小, 主题越分明 (对于1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354, 355, 356, 357, 358, 359, 360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386, 387, 388, 389, 390, 391, 392, 393, 394, 395, 396, 397, 398, 399, 400, 401, 402, 403, 404, 405, 406, 407, 408, 409, 410, 411, 412, 413, 414, 415, 416, 417, 418, 419, 420, 421, 422, 423, 424, 425, 426, 427, 428, 429, 430, 431, 432, 433, 434, 435, 436, 437, 438, 439, 440, 441, 442, 443, 444, 445, 446, 447, 448, 449, 450, 451, 452, 453, 454, 455, 456, 457, 458, 459, 460, 461, 462, 463, 464, 465, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 520, 521, 522, 523, 524, 525, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565, 566, 567, 568, 569, 570, 571, 572, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 624, 625, 626, 627, 628, 629, 630, 631, 632, 633, 634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 659, 660, 661, 662, 663, 664, 665, 666, 667, 668, 669, 670, 671, 672, 673, 674, 675, 676, 677, 678, 679, 680, 681, 682, 683, 684, 685, 686, 687, 688, 689, 690, 691, 692, 693, 694, 695, 696, 697, 698, 699, 700, 701, 702, 703, 704, 705, 706, 707, 708, 709, 710, 711, 712, 713, 714, 715, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 730, 731, 732, 733, 734, 735, 736, 737, 738, 739, 740, 741, 742, 743, 744, 745, 746, 747, 748, 749, 750, 751, 752, 753, 754, 755, 756, 757, 758, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800, 801, 802, 803, 804, 805, 806, 807, 808, 809, 810, 811, 812, 813, 814, 815, 816, 817, 818, 819, 820, 821, 822, 823, 824, 825, 826, 827, 828, 829, 830, 831, 832, 833, 834, 835, 836, 837, 838, 839, 840, 841, 842, 843, 844, 845, 846, 847, 848, 849, 850, 851, 852, 853, 854, 855, 856, 857, 858, 859, 860, 861, 862, 863, 864, 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 875, 876, 877, 878, 879, 880, 881, 882, 883, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 900, 901, 902, 903, 904, 905, 906, 907, 908, 909, 910, 911, 912, 913, 914, 915, 916, 917, 918, 919, 920, 921, 922, 923, 924, 925, 926, 927, 928, 929, 930, 931, 932, 933, 934, 935, 936, 937, 938, 939, 940, 941, 942, 943, 944, 945, 946, 947, 948, 949, 950, 951, 952, 953, 954, 955, 956, 957, 958, 959, 960, 961, 962, 963, 964, 965, 966, 967, 968, 969, 970, 971, 972, 973, 974, 975, 976, 977, 978, 979, 980, 981, 982, 983, 984, 985, 986, 987, 988, 989, 990, 991, 992, 993, 994, 995, 996, 997, 998, 999, 1000)

② $\vec{\alpha}$ 越不均一 越容易在某些主题
聚集

③ $\vec{\alpha}$ 越大, 先验分布越分散

④ $\vec{\alpha}$ 越小, 越聚焦

⑦ LDA 的算法理解

每一篇文章都有一个主题分布 $D_i \sim (\theta_{i1}, \theta_{i2}, \dots, \theta_{ik})$: 超参数记为 α 设为 α

↓

根据⑥, 后验都是 Dirichlet 分布

↑

每个主题都有一个词分布 $T_j \sim (\varphi_{j1}, \varphi_{j2}, \dots, \varphi_{jv})$ 超参数记为 β 设为 β

先验概率: 建模为 Dirichlet 分布, $\vec{\alpha} = \{\alpha_1, \dots, \alpha_k\}$ 方法1: 事先给定参数 方法2: 计算(见后面)

似然概率: 建模为多项分布

先验概率: 建模为 Dirichlet 分布, $\vec{\beta} = \{\beta_1, \dots, \beta_v\}$ 方法1: 事先给定参数 方法2: 计算(见后面)

似然概率: 建模为多项分布

偏例 词位 $i: 0 \sim \max\{\text{length}(D_j)\}$

偏例 文章序号: $j: 0 \sim m$

拿到二元组 $\langle j, i \rangle$, 即第 j 篇文章第 i 个词的位置标记, 向这个位置填一个词

如何填这个词?

文章 $D_j \xrightarrow{\text{根据 } D_j \text{ 的主题分布采样一个主题}} \text{主题 } T_k \xrightarrow{\text{根据 } T_k \text{ 的词分布采样一个词}} \text{词 } w_{ji}$

$\theta_{D_j} = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jk})$ $\varphi_{T_k} = (\varphi_{k1}, \varphi_{k2}, \dots, \varphi_{kv})$

反过来理解, 假定我们知道 $\vec{\theta}$ 和 $\vec{\varphi}$, 如何生成 M 篇文章的词 $\{w_{ij}\}$

而实际中, 词 w_{mn} 是能从样本观察到的已知变量, $\vec{\alpha}, \vec{\beta}$ 是事先给定的先验参数, $\vec{\theta}, \vec{\varphi}$ 是未知的隐变量是我们要求的。

① 如何从已知词 $w_{m,n}$ (文档 m , 词 n) 背后主题是 z_k 的概率? (算主题的词)

用 Gibbs 采样:

- <0> 初始给每个 $w_{m,n}$ 随机指定 z_k
- <1> 每一轮迭代:
 - a. 统计每个主题 z_k 出词 t 的数量
 - b. 统计每个文档 m 下出现主题 z_k 的数量
 - c. 用计算 $p(z_k | z_{-i}, d, w)$: 即排除当前词的主题分布
 - d. 用新的分布更新 z_k
- <2> 直到收敛或达到最大迭代次数

理解下面推导不用懂

词 w 是主题 k 的概率

$p(z_i = k | z_{-i}, w) = \dots \propto \frac{n_{k-i}^{(t)} + \beta_i}{\sum_{t=1}^V n_{k-i}^{(t)} + \beta_i} \leftarrow \text{词 } w \text{ 是主题 } k \text{ 的根概率 (文档)}$

$\propto \frac{(n_{m-i}^{(k)} + \alpha_k)}{\sum_{k=1}^K (n_{m-i}^{(k)} + \alpha_k)} \leftarrow \text{这是主题 } k \text{ 的概率}$

$\propto \text{其它词是主题 } k \text{ 的概率 (文档)}$

⑥ 结论:

主题的词分布: $\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^V n_k^{(t)} + \beta_t} = \dots$ (见 PPT)

文章的主题分布: $\psi_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^K n_m^{(k)} + \alpha_k} = \dots$ (见 PPT)

⑦ 代码: 见 PPT.

⑧ 主题个数 (超参数 K) 的确定

方法 1: 主题间相似度.

K 个主题词分布向量, 计算向量间的两两相似度求均值.



↑ 选中间这一段的取值.

方法 2: 使用概率分布困惑度 (Perplexity)

① 原理: 困惑度定义 $\text{Perplexity} = e^{-\frac{1}{N} \sum_{i=1}^N \ln q(x_i)}$

\uparrow 样本数 \uparrow 样本 i 的概率分布

也可定义为 $a^{-\frac{1}{N} \sum_{i=1}^N \log_a q(x_i)}$

\uparrow 任意整数

$\rightarrow = e^{H(q)}$ 信息熵. 越大越接近均匀分布, 越小越退化成一种确定性的分布. (越好)

如果已知样本真实分布 P , 这个指数部分也可换成交叉熵.

② 计算 LDA 模型的困惑度

算出 $q(x)$ 代入 Perplexity 公式 就可以

$q(x) = P(w | \text{Model}) = \prod_{i=1}^V \underbrace{p(w_i | \text{Model})}_{z_i} = e^{-\frac{1}{V} \sum_{i=1}^V \log p(w_i | \text{Model})}$

$H(P, q) = -\sum_x P(x) \ln q(x)$

③ 使用: 为不同 K 训练出的 LDA 模型也计算困惑度,

选困惑度小的 (实践效果并没有理论那么好)

$$p(w_i | \text{Model}) = \prod_{n=1}^{N_m} \sum_{k=1}^K p(z=k | d=m) p(w_i | z=k)$$

$$= \prod_{t=1}^V \left(\sum_{k=1}^K \psi_{m,k} \varphi_{k,t} \right)$$

⑨ α, β 超参数确定.

α : 越小越鲜明 (文章主题的超参数)



β : (主题的词分布超参数) 越小主题越突出, 越不受无关词的影响

eg. β 小 \rightarrow 可能只有大理、段誉这样的词影响武侠这个主题

β 大 \rightarrow 扬州、洛阳这样的词也会关联到武侠上.

<a> 直接设置 (不具普遍适用性) \leftarrow 但实践中这么设就够了

$$\alpha = 50/k$$

$$\beta = 200/W$$

<d> 用 Di-gamma (双 Gamma) 函数迭代求解

$$\alpha_k = \frac{\left[\sum_{m=1}^M \psi(n_m^{(k)} + \alpha_k) \right] - M \psi(\alpha_k)}{\left[\sum_{m=1}^M \psi \left[n_m + \sum_{j=1}^k \alpha_j \right] \right] - M \cdot \psi \left(\sum_{j=1}^k \alpha_j \right)} \cdot \alpha_k$$

\uparrow
先指定
然后一轮迭代

k : 主题编号

$n_m^{(k)}$: 主题 k 在文档 m 中出现多少次. (?)

⑩ Text Rank

模仿 PageRank 的算法, 从文章中提取代表句子.

⑪ word2vec \rightarrow 用来发现近义词 (给定一个词张度)

把文章中的每个词, 映射到一个 V 维向量 (V 维空间一个点)

维度减小时 (如减到 200) \rightarrow 对应 200 维空间的一个点上. (词嵌入)

* 如果 [词A] 与 [词B] 所在的语境相近, 就会映射到接近的两个向量

(相邻词相近)
词A与词B就是“近义词”

* 文中使用方法 (见 PPT) (实践课末尾)