《数据挖掘技术》

⭐ CH17 潜在语义分析

➡️ Create by *Wang JingHui*

➡️ Last Revision Time: *2021.04.15*

章节目录

导读

- 潜在语义分析主要用于文本的话题分析，通过矩阵分解发现文本与单词之间的**基于话题**的语义关系。
- 词向量通常是稀疏的，词向量不考虑同义性，也不考虑多义性。
- 一个文本(Doc)一般有多个话题(Topic)。
- 潜在语义分析使用的是**非概率**的话题分析模型。
- 潜在语义分析是**构建话题向量空间的方法**(话题分析的方法)
- 单词向量转化成话题向量。文本在不同空间下的相似度用在不同空间下的向量内积表示。
- 话题向量空间T，单词-话题矩阵T，文本在话题空间的表示Y，话题-文本矩阵Y

向量空间模型

单词向量空间

每个向量对应一个文本，单词向量空间行对应单词，话题向量空间行对应话题。

单词-文本矩阵

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

元素x_{ij}代表单词w_i在文本d_j中出现的频数或者权值。

单词多，文本少，这个矩阵是稀疏矩阵。

权值通常用TFIDF

$$TFIDF_{ij} = \frac{tf_{ij}}{tf_{\cdot j}} \log \frac{\mathrm{d}f}{\mathrm{d}f_i}$$

$$i = 1, 2, \cdots, m;$$

$$j = 1, 2, \cdots, n$$

一个单词在一个文本中的TFIDF是两种重要度的乘积，表示综合重要度。

话题向量空间

每个话题由一个定义在单词集合W上的m维向量表示，称为**话题向量**。

$$t_l = \begin{bmatrix} t_{1l} & t_{2l} & \cdots & t_{ml} \end{bmatrix}^{\mathrm{T}}, l = 1, 2, \cdots, k$$

k个话题向量张成一个话题向量空间，维数为k。

$$T = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1k} \\ t_{21} & t_{22} & \cdots & t_{2k} \\ \vdots & \vdots & & \vdots \\ t_{m1} & t_{12} & \cdots & t_{mk} \end{bmatrix}$$

矩阵T可以写成$T = \begin{bmatrix} t_1 & t_2 & \cdots & t_k \end{bmatrix}$

文本在话题向量空间的表示

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{k1} & y_{k2} & \cdots & y_{kn} \end{bmatrix}$$

矩阵Y可以写做$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}$

$$x_j \approx y_{1j}t_1 + y_{2j}t_2 + \cdots + y_{kj}t_k, j = 1, 2, \cdots, n$$

从单词向量空间到话题向量空间的线性变换

- 单词-文本矩阵X可以近似的表示为单词-话题矩阵T与话题-文本矩阵Y的乘积形式，这就是潜在语义分析。

$$X \approx TY$$

基于SVD的潜在语义分析模型

单词-文本矩阵

文本集合$D = \{d_1, d_2, \cdots, d_n\}$
单词集合$W = \{w_1, w_2, \cdots, w_m\}$

表示成单词-文本矩阵$X_{m \times n}$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

截断奇异值分解

$$X \approx U_k \Sigma_k V_k^{\mathrm{T}} = \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_k \end{bmatrix} \begin{bmatrix} v_1^{\mathrm{T}} \\ v_2^{\mathrm{T}} \\ \vdots \\ v_k^{\mathrm{T}} \end{bmatrix}$$

这中间$k \le n \le m$ 这里假设了文档数量要比单词数量少，其实这个假设也不一定成立。

- U_k是$m \times k$矩阵，前k个相互正交的左奇异向量；
- Σ是k阶方阵，前k个最大奇异值；
- V_k是$n \times k$矩阵，前k个相互正交的右奇异向量；

话题空间向量

每一列u_l表示一个话题，k个话题张成一个子空间，称为话题向量空间。

$$U_k = \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix}$$

文本的话题空间向量表示

- 如果u_l表示话题向量空间，那么将文本表示成u_l的线性组合，就是文本在这个空间的表示。
- 奇异值分解得到三个矩阵，最左边的是话题向量空间，那么右边的两个矩阵的乘积，则对应了话题-文本矩阵(文本的话题空间向量表示)。

$$V^{\mathrm{T}} = \begin{bmatrix} v_{11} & v_{21} & \cdots & v_{n1} \\ v_{12} & v_{22} & \cdots & v_{n2} \\ \vdots & \vdots & & \vdots \\ v_{1k} & v_{2k} & \cdots & v_{nk} \end{bmatrix}$$

矩阵X的第j列向量x_j满足：

$$
\begin{aligned}
x_j &\approx U_k(\Sigma_k V_k^{\mathrm{T}})_j \\
&= \begin{bmatrix} u_1 & u_2 & \cdots & u_k \end{bmatrix} \begin{bmatrix} \sigma_1 v_{j1} \\ \sigma_2 v_{j2} \\ \vdots \\ \sigma_k v_{jk} \end{bmatrix} \\
&= \sum_{l=1}^{k} \sigma_l v_{jl} u_l, j = 1, 2, \cdots, n
\end{aligned}
$$

上式是文本d_j的近似表达式，由k个话题向量u_l的线性组合构成。
矩阵$(\Sigma_k V_k^{\mathrm{T}})$的每一个列向量是一个文本在话题向量空间的表示。

基于NMF的潜在语义分析模型

NMF

- X是非负矩阵则表示为$X \geq 0$
- $X \approx WH, W \geq 0, H \geq 0$称为非负矩阵分解
- 非负矩阵分解旨在通过较少的基向量、系数向量来表达较大的数据矩阵。

模型定义

$m \times n$的非负矩阵$X \geq 0$。

假设文本集合包含k个话题，对X进行非负矩阵分解。即求$m \times k$的非负矩阵和$k \times n$的非负矩阵满足

$$X \approx WH$$

其中

- $W = \begin{bmatrix} w_1 & w_2 & \cdots & w_k \end{bmatrix}$表示话题向量空间,
- $w_1, w_2, \cdots, w_k$表示文本集合的k个话题
- $H = \begin{bmatrix} h_1 & h_2 & \cdots & h_k \end{bmatrix}$表示文本在话题向量空间的表示
- $h_1, h_2, \cdots, h_k$表示文本集合的n个文本

以上是基于非负矩阵分解的潜在语义分析模型。

非负矩阵分解有很直观的解释，话题向量和文本向量都非负，对应着“伪概率分布”，向量的线性组合表示**局部构成总体**。这个其实和DL里面的意思是一样的。

形式化为最优化问题求解-损失函数

- 平方损失

  两个非负矩阵$A = [a_{ij}]_{m\times n}$和$B = [b_{ij}]_{m\times n}$的平方损失定义为

$$\|A - B\|^2 = \sum_{i,j}(a_{ij} - b_{ij})^2$$

- 散度

$$D(A\|B) = \sum_{i,j}\left(a_{ij}\log\frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij}\right)$$

A和B不对称。

当$\sum_{i,j} a_{ij} = \sum_{i,j} b_{ij} = 1$时散度损失函数退化为Kullback-Leiber散度或相对熵，这时A和B是概率分布。

问题定义

针对不同的损失函数有不同的问题定义

1. 平方损失

$$\min_{W,H} \|X - WH\|^2$$
$$s.t. W, H \geq 0$$

2. 散度损失

$$\min_{W,H} D(X\|WH)$$
$$s.t. W, H \geq 0$$

算法

目标函数只是对W和H之一的凸函数，而不是同时两个变量的凸函数，所以通过数值优化求解局部最优解。

1. 平方损失

$$H_{lj} \leftarrow H_{lj} \frac{(W^{\mathrm{T}} X)_{lj}}{(W^{\mathrm{T}} W H)_{lj}}$$
$$W_{il} \leftarrow W_{il} \frac{(X H^{\mathrm{T}})_{il}}{(W H H^{\mathrm{T}})_{il}}$$

2. 散度损失

$$H_{lj} \leftarrow H_{lj} \frac{\sum_i [W_{il} X_{ij} / (WH)_{ij}]}{\sum_i W_{il}}$$

$$W_{il} \leftarrow W_{il} \frac{\sum_j [H_{lj} X_{ij} / (WH)_{ij}]}{\sum_j H_{lj}}$$

算法

输入：单词-文本矩阵$X \geq 0$，文本集合的话题个数k，最大迭代次数t；

输出：话题矩阵W，文本表示矩阵H

1. 初始化
   $W \geq 0$，**并对W的每一列数据归一化**
   $H \geq 0$
2. 迭代
   对迭代次数从1到t执行下列步骤：
   a. 更新W的元素，每次迭代对W的列向量归一化，**使基向量为单位向量。**
   b. 更新H的元素

习题

习题18.3

参考

https://zhuanlan.zhihu.com/p/144367432

Enjoy your machine learning!

https://github.com/wjssx/Statistical-Learning-Slides-Code

E-mail: csr_dsp@sina.com