Домашнее задание №3

Информационный поиск. 6 курс. Осенний семестр.

Горбунов Егор Алексеевич

11 ноября 2016 г.

Задание №1 Перед вами матрица смежности «термин-документ», описывающая некую коллекцию (строки - термы, столбцы - документы):

$$C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

(a) Вычислите матрицу совместной встречаемости CC^T. Что собой представляют диагональные элементы этой матрицы?

(b) Убедитесь, что сингулярное разложение матрицы C выглядит следующим образом:

$$U = \begin{pmatrix} -0.816 & 0.000 \\ -0.408 & -0.707 \\ -0.408 & 0.707 \end{pmatrix}, \Sigma = \begin{pmatrix} 1.732 & 0.000 \\ 0.000 & 1.000 \end{pmatrix}, V^T = \begin{pmatrix} -0.707 & -0.707 \\ 0.707 & -0.707 \end{pmatrix}$$

(c) Что собой представляют элементы матрицы $C^T C$?

Решение:

(a) Обозначим вектора слов:

$$w_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, w_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, w_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Элемент вектора $w_i[j]$ обозначает, встретилось ли слово w_i в документе D_j.

Тогда матрица совместной встречаемости ($\cdot$ - скалярное произведене, dot product):

$$CC^T = \begin{pmatrix} w_1^T \\ w_2^T \\ w_3^T \end{pmatrix} \begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix}^T = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} \begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix} = \begin{pmatrix} w_1 \cdot w_1 & w_1 \cdot w_2 & w_1 \cdot w_3 \\ w_2 \cdot w_1 & w_2 \cdot w_2 & w_2 \cdot w_3 \\ w_3 \cdot w_1 & w_3 \cdot w_2 & w_3 \cdot w_3 \end{pmatrix}$$

Откуда:

$$CC^T = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

В силу того, что w_i — вектор над $\{1, 0\}$, видно, что диагональный элемент матрицы $CC^T[i, i]$ равен числу документов коллекции, в которых встретилось слово w_i. Вообще: $CC^T[i, j]$ - это скалярное произведение $w_i \cdot w_j$, которое представляет из себя сумму элементов вектора, в котором 1 стоят на таких позициях k, что $w_i[k] = w_j[k] = 1$. Таким образом сумма элементов данного вектора будет равна число документов, в которых одновременно встречается как слово w_i так и слово w_j.

(b) Перемножив данные матрицы получим:

$$U\Sigma V^T = \begin{pmatrix} 0.9992 & 0.9992 \\ -0.0002 & 0.9994 \\ 0.9994 & -0.0002 \end{pmatrix}$$

Вообщем-то похоже, но с погрешность. Как минимум по-этому SVD стоит пересчитать руками =) Также, по определению сингулярного разложения исходная матрица раскладывается в произведение унитарной, диагональной (из ненулевых сингулярных чисел) и ещё одной унитарной матрицы. Матрица U, приведённая в задании не является унитарной, как минимум потому, что не является квадратной. Таким образом приведённое разложение нельзя называть сингулярным (как я понимаю), хотя это и не означает, что такое разложение неверно. Поэтому найдём сингулярное разложение матрицы C своими силами. Будем искать матрицы U, V и Σ, что U имеет размер 3×3, Σ 3×2, а V 2×2. В силу унитарности матриц U и V можем записать следующее:

$$U^T U = UU^T = V^T V = VV^T = I$$
$$CC^T = U\Sigma V^T V \Sigma^T U^T = U\Sigma\Sigma^T U^T$$
$$C^T C = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T \Sigma V^T$$

Тут матрица $\Sigma^T\Sigma = \Sigma\Sigma^T = \Lambda$ — это квадратная диагональная матрица 3×3 (на диагонали могут быть нули).

Откуда мы получаем, домножая справа обе части на нужные U в одном уравнении и на V в другом:

$$(CC^T)U = U\Lambda$$
$$(C^T C)V = V\Lambda$$

Можно переписать это так:

$$(CC^T)\vec{u_i} = \vec{u_i}\lambda_i, \text{ для всех столбцов } u_i \text{ матрицы } U$$
$$(C^T C)\vec{v_i} = \vec{v_i}\lambda_i, \text{ для всех столбцов } v_i \text{ матрицы } V$$
$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \ \Sigma[i][i] = \sqrt{\lambda_i}$$

Видим, что собственные числа (ненулевые!) матриц $C^T C$ и CC^T совпадают и равны квадратам искомых сингулярных чисел, составляющих Σ, а столбцы матрицы V — собственные вектора $C^T C$ и столбцы U — собственные вектора CC^T. Таким образом нам нужно отыскать собственные числа и вектора матриц:

$$CC^T = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},\ C^TC = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Начнём с собственных чисел. Ищем их через характеристические многочлены (приравнивая их к 0, почему так делается я не поясняю, т.к. это за рамками курса):

$$\det\begin{pmatrix} 2-\lambda & 1 & 1 \\ 1 & 1-\lambda & 0 \\ 1 & 0 & 1-\lambda \end{pmatrix} = (2-\lambda)(1-\lambda)^2 - (1-\lambda) - (1-\lambda) = \lambda(1-\lambda)(\lambda-3)$$

$$\Lambda = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Чтобы получить U и V теперь нужно найти собственные векторы соответствующие собственным числам 3 и 1 матриц CC^T и C^TC. Теперь уже совсем очевидно, что столбец $\vec{u_3}$ может быть произвольным собственным вектором (соответствует с.ч. 0) и роли он в разложении играть не будет. Для векторов получаем следующие системы (матрица системы та же, что под $\det$ выше, но с уже подставленными λ) (u_1, v_1 соответствует $\lambda = 3$, u_2, v_2 соотв. $\lambda = 1$). Т.к. итоговые матрицы U и V должны быть унитарными, то вектора необходимо нормализовать.

$$\begin{cases} -u_{11} + u_{12} + u_{13} = 0 \\ u_{11} - 2u_{12} = 0 \\ u_{11} - 2u_{13} = 0 \end{cases} \text{, откуда легко подобрать } \begin{pmatrix} -2 \\ -1 \\ -1 \end{pmatrix} \Rightarrow u_1 = \frac{1}{\sqrt{6}}\begin{pmatrix} -2 \\ -1 \\ -1 \end{pmatrix}$$

$$\begin{cases} u_{21} + u_{22} + u_{23} = 0 \\ u_{21} = 0 \\ u_{21} = 0 \end{cases} \text{, откуда легко подобрать } \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \Rightarrow u_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$$

Аналогично можно подобрать собственный вектор для $\lambda_3 = 0$: $u_3 = \frac{1}{\sqrt{3}}\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$. Для v_i:

$$\begin{cases} -v_{11} + v_{12} = 0 \\ v_{11} - v_{12} = 0 \end{cases} \text{, откуда легко подобрать } \begin{pmatrix} -1 \\ -1 \end{pmatrix} \Rightarrow v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ -1 \end{pmatrix}$$

$$\begin{cases} v_{21} + v_{22} = 0 \\ v_{21} + v_{22} = 0 \end{cases} \text{, откуда легко подобрать } \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Итого мы получили искомое разложение:

$$U = \begin{pmatrix} u_1 & u_2 & u_3 \end{pmatrix} = \begin{pmatrix} \frac{-2}{\sqrt{6}} & 0 & \frac{-1}{\sqrt{3}} \\ \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{pmatrix}$$

$$V^T = \begin{pmatrix} v_1^T \\ v_2^T \end{pmatrix} = \begin{pmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Конечно, при вычислении собственных векторов я выбирал те знаки, которые бы соответствовали тому, что дано в задании =) Теперь, если перевести все значения в десятичные дроби легко увидеть, что полученные матрицы совпадают (если округлить до нужного числа знаков), кроме того, что вычисленная явно U является честно унитарной, хоть на дальнейшие выкладки (например, при вычислении новых векторов для документов (LSA)) это не влияет.

(c) Аналогично пункту (a) введём вектора документов:

$$D_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} D_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Тут если $D_i[k] = 1$, то терм w_k встретился в документе D_i.
Тогда получим:

$$C^T C = \begin{pmatrix} D_1 \cdot D_1 & D_1 \cdot D_2 \\ D_2 \cdot D_1 & D_2 \cdot D_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Откуда легко видеть, опять же аналогично первому пункту задания, что $C^T C[i, j]$ — это число термов, которые *одновременно* встречаются в документах D_i и D_j, т.е. это мощность пересечения мешков слов для документов D_i и D_j. Ещё можно написать, что $C^T C[i, j] = sim(D_i, D_j)||D_i||||D_j||$.

Задание №2 Для чего используются распределения Дирихле $Dir(\vec{\alpha})$ и $Dir(\vec{\beta})$ в тематических моделях? Что контролируют параметры $\vec{\alpha}$ и $\vec{\beta}$? Какие значения этих параметром имеет смысл использовать и почему?

Решение: Начнём с того, что распределение Дирихле - это *сопряжённое априорное распределение* для мультиномиального распределения. Функция плотности вероятности распределения Дирихле:

$$Dir(\vec{p}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{m} p_i^{\alpha_i - 1}, \text{ где } \vec{p} = \{p_1, \ldots, p_m\}, \vec{\alpha} = \{\alpha_1, \ldots, \alpha_m\}$$

$$B(\vec{\alpha}) = \frac{\prod_{i=1}^m \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^m \alpha_i)} = [\text{при натуральных } \alpha_i] = \frac{\prod_{i=1}^m (\alpha_i - 1)!}{(\sum_{i=1}^m \alpha_i - 1)!}$$

$B(\vec{\alpha})$ — бета-функция, при раскрытии которой выше было использовано, что $\Gamma(x)$ (гамма-функция) — это обобщение факториала.