Школа лингвистики, 2020-21 уч. год
Линейная алгебра и математический анализ
Метод наименьших квадратов. Латентный семантический анализ и сингулярное разложение. Мера tf-idf. (07.12.2020/08.12.2020)
Ю. Г. Кудряшов, И. В. Щуров, А. М. Изосимов, Д. А. Филимонов,Р. Я. Будылин

1 Приближение данных (фит)

Обычно в математических моделях различных процессов содержится множество параметров, которые необходимо отыскать, пользуясь конкретными данными. Для отыскания таких параметров используется метод наименьших квадратов, который позволяет аппроксимировать данные заданной функцией.

Пусть имеется массив данных $\{x_i; y_i\}$, $i = 1..k$. Из теории предполагается, что данные должны лежать на кривой $f(x; a_1, ..., a_n)$, где $a_1, ..., a_n$ — некие параметры. Тогда эти параметры ищутся так, чтобы минимизировать выражение

$$\sum_{i=1}^{k} (y_i - f(x_i))^2 \to \min$$

Задача 1. Имеются следующие данные:

x_i	0	1	2
y_i	0	1	3

Найти методом наименьших квадратов параметры при приближении функцией:

(a) $f(x) = ax$;
(b) $f(x) = ax + b$;
(c) $f(x) = ax^2 + bx + c$.

Во всех случаях посчитать качество приближения (сумму квадратов отклонений от данных).

2 Латентный семантический анализ и сингулярное разложение

Задача 2. Для следующих наборов предложений постройте матрицы термы-документы в метрике tf-idf.

(a)
- Я сел на стул.
- Я сел за стол.
- Он сел на стол.

(b)
- Быть, или не быть?
- Может, не надо?
- Быть того не может!

Задача 3. Найти сингулярное разложение матрицы $X = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}$

Задача 4. Для матрицы $X = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 2 & 0 \end{pmatrix}$

(a) Найдите сингулярное разложение.

(b) Спроецируйте столбцы (документы) и строки (термы) на 2-мерное подпространство главных компонент. Как найти базисные векторы нового пространства в старых координатах?

(c) Найдите проекцию вектора $d = (0, 0, 1, 1)$ на найденное 2-мерное семантическое подпространство.

(d) Если считать, что матрица X — это матрица термы-документы, а вектор v — это поисковый запрос, то в каком порядке стоит ранжировать документы в поисковой выдаче?