Singular Value Decomposition (SVD)
SVD разложение

Регуляризация
1. Многокритериальная оптимизация
2. Предотвращение чрезмерной подгонки
3. **Разумные значения параметров модели**
4. Борьба с чрезмерной подгонкой

1. Линейная алгебра

Определение 1:
Пусть A вещественная матрица размера $n \times n$. Тогда собственным числом матрицы называется любое решение уравнения $det(A - \lambda I) = 0$, где I единичная матрица размера $n \times n$. Если выполнено $Ab = \lambda b$, то n-мерный вектор b единичной длины называется собственным вектором, соответствующим собственному числу λ.

Теорема 1:
Если A симметричная неотрицательно определенная матрица, то все собственные числа вещественные и неотрицательные.

Теорема 2 (Спектральное разложение матрицы):
Пусть A вещественная симметричная матрица размера $n \times n$. Тогда A может быть представлена в виде

$$A = U \times S \times U^T \tag{1}$$

При этом,
1. матрица U ортогональная размера $n \times n$;
2. матрица S диагональная, на главной диагонали стоят собственные числа $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;
3. i-й столбец матрицы U - собственный вектор, соответствующий i-ому по величине собственному числу;
4. если все собственные числа различные, то представление матрицы A единственное.

2. Метод главных компонент. Вероятностный подход.

Замечание. Метод главных компонент часто называют факторным анализом.

Рассмотрим случайный вектор $X^T = (X_1, X_2, \ldots, X_k)$
Обозначим R матрицу корреляций случайного вектора X.

Теорема 3:
Ковариационная (корреляционная) матрица симметричная и неотрицательно определенная.

Ищем "хорошие"линейные комбинации элементов вектора.

Задача 1.
Найти линейную комбинацию $Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k$, у которой максимальная дисперсия $D(Y_1)$.

Дополнительное условие 1: $a_1^T \cdot a_1 = 1$, где $a_1^T = (a_{11}, a_{12}, \ldots, a_{1k})$

Задача 2.
Найти линейную комбинацию $Y_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2k}X_k$, у которой максимальная дисперсия $D(Y_2)$.
Дополнительное условие 2.1: $a_2^T \cdot a_2 = 1$, где $a_2^T = (a_{21}, a_{22}, \ldots, a_{2k})$
Дополнительное условие 2.2: $corr(Y_2, Y_1) = 0$

Задача 3.
Найти линейную комбинацию $Y_3 = a_{31}X_1 + a_{32}X_2 + ... + a_{3k}X_k$, у которой максимальная дисперсия $D(Y_3)$.
Дополнительное условие 3.1: $a_3^T \cdot a_3 = 1$, где $a_3^T = (a_{31}, a_{32}, \ldots, a_{3k})$
Дополнительное условие 3.2: $corr(Y_3, Y_1) = 0$
Дополнительное условие 3.3: $corr(Y_3, Y_2) = 0$

Задача k.
Найти линейную комбинацию $Y_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kk}X_k$, у которой максимальная дисперсия $D(Y_k)$.
Дополнительное условие k.1: $a_k^T \cdot a_k = 1$, где $a_k^T = (a_{k1}, a_{k2}, \ldots, a_{kk})$
Дополнительное условие k.2: $corr(Y_k, Y_1) = 0$
Дополнительное условие k.3: $corr(Y_k, Y_2) = 0$
...
Дополнительное условие k.k: $corr(Y_k, Y_{k-1}) = 0$

Решение задач 1–k
Вектор a_i - собственный вектор матрицы корреляций, соответствующий i-ому по величине собственному числу:

$$Ra_i = \lambda_i a_i \quad i = 1 \ldots k \tag{2}$$

При этом $D(Y_i) = \lambda_i$

Обозначение
Обозначим G матрицу размера $k \times k$, составленную из векторов столбцов a_i.

Следствие. Матрица G ортогональная.

Перепишем уравнение (2)

$$RG = SG \quad i = 1 \ldots k \tag{3}$$

где матрица S диагональная, на главной диагонали стоят собственные числа матрицы R
$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$

3. Метод главных компонент. Статистический подход.

Обозначим X таблицу данных размера $n \times k$, где X_{ij} - i-ое наблюдение j-ой случайной величины.

Все столбцы стандартизованы, то есть для каждой характеристики выборочное среднее равно 0, а выборочная дисперсия равна 1.

Обозначим Y таблицу данных размера $n \times k$, где Y_{ij} - i-ое наблюдение j-ой главной компоненты.

Обозначим $\hat{R}_X$ матрицу выборочных корреляций переменных.

Обозначим $\hat{R}_Y$ матрицу выборочных корреляций главных компонент.

Тогда

$$\hat{R}_X = \frac{1}{n-1} X^T X \tag{4}$$

$$\hat{R}_Y = \frac{1}{n-1} Y^T Y = S \tag{5}$$

матрица S диагональная, на главной диагонали стоят собственные числа $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;

Для дальнейшего важно, что матрица корреляций пропорциональна произведению матриц данных

4. Singular Value Decomposition (SVD)

Теорема 3. ('thin' SVD)
Вещественная матрица X ранга r и размера $n \times m$ может быть представлена в виде

$$X = U \times S \times V^T \tag{6}$$

При этом,
1. матрицы U, S и V вещественные;
2. размер матрицы U равен $n \times r$;
3. размер матрицы S равен $r \times r$;
4. размер матрицы V^T равен $r \times m$;
5. столбцы матрицы U ортонормированы, то есть $U^T \times U = I$;
6. столбцы матрицы V ортонормированы, то есть $V^T \times V = I$;
7. матрица S диагональная, на главной диагонали стоят сингулярные значения $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$;
8. если все сингулярные значения различные, то представление матрицы X единственное, с точностью до одновременного умножения на -1 столбца матрицы U и соответствующей строки матрицы V^T .

Комментарий 1.
На практике почти всегда $r = m$.

Комментарий 2.
В теореме используется V^T, а не V, для единообразия в приложениях.
Каждому наблюдению сопоставляется строка матрицы U.
Каждой переменной сопоставляется строка матрицы V.

5. SVD и метод главных компонент

Одновременный факторный анализ столбцов и строк.

Собственные числа и сингулярные значения
Rebranding факторного анализа

Latent semantic analysis. Latent Semantic Indexing (LSI)
Запатентовано

строки - слова
столбцы - тексты

6. SVD и сокращение размерности данных

Оставить "малое"число факторов,

Норма Фробениуса,
справедливо

Пусть матрица $\tilde{X}$ получена с помощью сокращенного

$$\tilde{X} = argmin_{B:rank(B)=k} \sum_{i,j} (X_{ij} - B_{ij})^2$$

6. SVD матрицы данных, содержащей пропуски. Метод Simon Funk'a

Ищем ~ в виде

$$\tilde{x_{ij}} = \vec{u_i}\vec{v_j}^T$$

Комментарий: нет диагональной матрицы!

Погрешность подгонки

$$e_{ij} = x_{ij} - \tilde{x_{ij}}$$

Минимизируем, но суммируем только по элементам матрицы, не являющимися пропусками

$$\tilde{X} = min(\sum_{i,j} (e_{ij})^2 + \gamma(\sum_{i,j} (u_{ij})^2 + \sum_{i,j} (u_{ij})^2))$$

Инициализируем, затем итеративно подправляем элементы факторов
Элементы векторов $\vec{u_i}$ и $\vec{v_j}$ модифицируются по правилу

$$v_{ik} = v_{ik} + \lambda(e_{ui}u_{uk} - \gamma v_{ik})$$

$$u_{uk} = u_{uk} + \lambda(e_{ui}v_{ik} - \gamma u_{uk})$$

где λ - скорость обучения,

γ - вес регуляризационного слагаемого

7. Использование SVD матрицы данных, содержащей пропуски

funkSVD: Funk SVD for Matrices with Missing Data

SVD++ the method that became popular from the Netflix prize contest has nothing to do with the SVD besides the name.

irlba