Сингулярное разложение (SVD)

SVD (singular value decomposition) – разложение, использующееся для приближения матриц матрицами заданного ранга.

Введём матрицу $M \in R^{m\times n}$ и зададим несколько определений (заметим, что SVD разложение можно вводить и для $M \in C^{m\times n}$ - вещественный случай рассмотрен для простоты).

Определение:

Матрица $A \in R^{n\times n}$ называется **ортогональной**, если $AA^T = A^TA = I$.

Определение:

Неотрицательное вещественное число σ называется **сингулярным числом** матрицы M, тогда и только тогда, когда существуют два вектора единичной длины $u \in R^m$ и $v \in K^n$ такие, что: $Mv = \sigma u$ и $MT\ u = \sigma v$.

Такие векторы u и v называются, соответственно, **левым сингулярным вектором** и **правым сингулярным вектором**, соответствующим сингулярному числу σ.

Определение:

Сингулярным разложением матрицы M является разложение следующего вида $M = U\Sigma V^T$, где Σ — матрица размера $m\times n$, у которой элементы, лежащие на главной диагонали — это сингулярные числа (а все элементы, не лежащие на главной диагонали, являются нулевыми), а матрицы $U \in R^{m\times m}$ и $V \in R^{n\times n}$ — это две ортогональные матрицы, состоящие из левых и правых сингулярных векторов соответственно.

Метод главных компонент (PCA)

Сингулярное разложение обладает и другим интересным свойством. Оно позволяет "приближать" матрицу M некоторой матрицей M_k меньшего ранга ($rank(M_k) = k$; $k \leq rank(M)$).

Теорема Эккарта Янга:

Для данной матрицы M существует её аппроксимация M_k ($rank(M_k = k \leq rank(M)$) такая, что: $\forall B \in (R^{m\times n} \cap rank(B) = k) : ||M - M_k||_F \leq ||M - B||_F$, где $||\cdot||_F$ - норма Фробениуса.

При этом, если элементы на диагонали матрицы Σ упорядочены по невозрастанию, то M_k = $U_k \Sigma k V^T{}_k$, где U_k, V_k получаются из U, V сингулярного разложения обрезанием до k первых столбцов, а Σ_k получается из Σ - обрезанием до первых k столбцов и строк.

Рассмотрим x – одно наблюдение за m - мерным пространством и x' - его ортогональную проекцию на некоторое подпространство m - мерного пространсва. Исходя из того, то проекция ортогональная: $||x||^2 = ||x'||^2 + ||x - x'||^2$. Теперь составим матрицы X и X' из n таких наблюдений и их ортогональных проеций и получим: $||X||^2{}_F = ||X'||^2{}_F + ||X - X'||^2{}_F$. Также заметим, что: $argmax\ ||X'||_F = argmin||X - X'||_F$ (в этом равенстве $X' \in X(k)$, X - фиксирована, а $X(k)$ - множество матриц ранга k со столбцами - ортогональными проекциями столбцов X на некоторое k - мерное подпространство).

Получено выражение из теоремы Эккарта-Янга. Значит для данной матрицы M на её аппроксимация из этой теоремы Mk достигается не только минимум $||M - M_k||_F$, но и максимум $||M_k||_F$ среди всех матриц $M(k)$.

Этот факт интерпретировать так: сингулярное разложение позволяет снижать размерность признакового пространства посредством проецирования данных на некоторое подпространство, по направлениям которого сохраняется максимальный ”разброс” признаков начального пространства. В этом и состоит метод главных компонент (principal component analysis).