Математические модели обработки сигналов

Тема 3: Приложения SVD

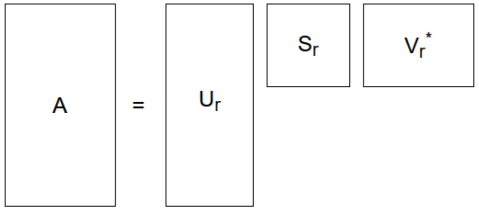
Лектор: Кривошеин А.В.

Неполное SVD

Теорема (теорема о неполном SVD). Пусть A является матрицей размера $m \times n$ с рангом $\operatorname{rank} A = r$. Тогда найдутся матрицы U_r размера $m \times r$ и V_r размера $n \times r$, удовлетворяющие равенствам

$$U_r^* U_r = I_r, \quad V_r^* V_r = I_r, \quad \text{такие что} \quad A = U_r S_r V_r^*,$$

где $S_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_1 \geq \ldots \geq \sigma_r > 0$, σ_i — сингулярные числа матрицы A.



Сокращённое SVD решает задачу наилучшего низкорангового приближения.

$$\underset{\hat{A},\ \operatorname{rank}(\hat{A})=d}{\operatorname{argmin}} \left\|A - \hat{A}\right\|_F = \underset{\hat{A},\ \operatorname{rank}(\hat{A})=d}{\operatorname{argmin}} \left\|A - \hat{A}\right\|_2 = A(d),$$

где матрица $A(d)$ имеет вид : $A(d) = U_r S_r(d) V_r^* = U_d S_d V_d^*$, $S_r(d) = \operatorname{diag}\{\sigma_1, \ldots, \sigma_d, 0, \ldots, 0\}$,

где U_d, V_d — это матрицы, составленные из первых d столбцов матриц U_r, V_r.

Эффективное вычисление SVD осуществляется алгоритмом Голуба-Кахана.

G. Golub and W. Kahan, Calculating the Singular Values and Pseudo-Inverse of a Matrix, SIAM. J. Numer. Anal., Vol. 2, 1965, pp. 202-224

SVD для решения СЛАУ

Многие задачи формулируются в терминах СЛАУ: $A\,x\ = b$,

где матрица A и вектор b известны, а вектор x неизвестен.

Если матрица A квадратна и обратима, то существует единственное решение x для каждого b.

Если матрица A вырождена или прямоугольна, то решение может быть одно, либо решений нет, либо бесконечно много решений в зависимости от вектора b и пространства столбцов матрицы A.

Проанализируем некоторые варианты.

1. Случай недоопределённой системы (англ. under-determined system). В этом случае $A \in \mathbb{C}^{m\times n}$ и $m << \mathrm{n}$, то есть A узкая-вытянутая (short-fat) матрица или число неизвестных больше числа уравнений. Наиболее вероятно, что столбцы матрицы образуют базис для $\mathbb{C}^m$ (так как количество столбцов n больше размерности пространства m, хотя конечно это и не всегда так, можно взять одинаковые столбцы). Если пространство столбцов матрицы A действительно совпадает с $\mathbb{C}^m$, то существует бесконечно много решений СЛАУ. Название under-determined system используется потому строк в матрице недостаточно, чтобы единственным образом найти решение x.

2. Случай переопределённой системы (англ. over-determined system). В этом случае $A \in \mathbb{C}^{m\times n}$ и $m >> \mathrm{n}$, то есть A высокая-тонкая (tall-skinny) матрица или число уравнений больше, чем число неизвестных. В этом случае матрица не может иметь m ЛНЗ столбцов, поэтому наиболее вероятно, что решений у СЛАУ нет. Решение (одно или бесконечно много) может быть лишь в том случае, когда b лежит в линейной оболочке столбцов матрицы A.

SVD для решения СЛАУ

Для случая недоопределённой системы, когда есть бесконечно много решений, имеет смысл найти решение x с минимальной нормой (англ. minimum-norm solution), то есть

$$\min_{x \in \mathbb{R}^n:\ A\,x = b} \|x\|_2.$$

Для случая переопределённой системы, когда решений нет, часто имеет смысл получить решение лучшее в средне-квадратичном, то есть то, которое минимизирует норму

$$\min_{x \in \mathbb{R}^n} \|A\,x - b\|_2^2.$$

И SVD позволяет получить решение указанных выше проблем. Для этого введём понятие псевдообратной матрицы.

Псевдообратная матрица

Определение. Пусть $A \in \mathbb{C}^{m\times n}$, $\operatorname{rank} A = r$ и неполное SVD матрицы A имеет вид $A = U_r\, S_r\, V_r^*$, где матрицы U_r размера $m\times r$ и V_r размера $n\times r$, удовлетворяющие равенствам

$$U_r^*\, U_r = I_r, \quad V_r^*\, V_r = I_r, \quad S_r := \operatorname{diag}\{\sigma_1, \ \ldots, \sigma_r\}, \ \sigma_1 \geq \ldots \geq \sigma_r > 0$$

σ_j — ненулевые сингулярные числа матрицы A.

Тогда матрицу вида $A^\dagger = V_r\, S_r^{-1}\, U_r{}^*$ называют псевдообратной матрицей Мура-Пенроуза.

Отметим сразу результат произведения матриц A и $A^\dagger$:

$$A\, A^\dagger = (U_r\, S_r\, V_r^*)\left(V_r\, S_r^{-1}\, U_r{}^*\right) = U_r\, U_r{}^*,$$

$$A^\dagger\, A = \left(V_r\, S_r^{-1}\, U_r{}^*\right)(U_r\, S_r\, V_r^*) = V_r\, V_r{}^*.$$

Это квадратные матрицы размера $m\times m$ и $n\times n$, соответственно.

Если A является квадратной невырожденной матрицей, то S_r является квадратной диагональной матрицей с ненулевыми значениями на диагонали и псевдообратная матрица является обратной, поскольку матрицы U_r и V_r будут унитарны.

Псевдообратная матрица и СЛАУ

Вернёмся к СЛАУ $A\,x = b$. Построим вектор $x^{\square} = A^{\dagger}\,b$. Если этот вектор подставить обратно в СЛАУ, получим

$A\,x^{\square} = A\,A^{\dagger}\,b = U_r\,U_r^{*}\,b$

Здесь матрица $U_r\,U_r^{*}$ не обязательно единичная, однако, вектор $x^{\square}$ обладает следующими свойствами.

Теорема (о решении СЛАУ с помощью псевдообратной матрицы). Пусть $A \in \mathbb{C}^{m\times n}$, $b \in \mathbb{C}^{m}$, $x^{\square} = A^{\dagger}\,b \in \mathbb{C}^{n}$.

1. Для всех $x \in \mathbb{C}^{n}$: $\|A\,x^{\square} - b\| \le \|A\,x - b\|$.

2. Если СЛАУ имеет хотя бы одно решение, то $x^{\square}$ является решением с наименьшей нормой среди всех любых других решений, то есть $\|x^{\square}\| \le \|x\|$, где $A\,x = b$.

Таким образом, вектор $x^{\square} = A^{\dagger}\,b$ даёт

1. решение с минимальной нормой, если решений много
2. наилучшее в средне-квадратичном решение, если точного решения нет.

Линейная регрессия

Рассмотрим классическую задачу линейной регрессии: есть набор из m штук однотипных объектов и каждый объект характеризуется набором из n численных признаков. Цель заключается в том, чтобы по известным признакам уметь вычислять некоторый целевой признак. Иными словами, надо построить модель, которая по набору из n численных признаков предсказывает целевой признак.

Решение задачи методом линейной регрессии означает, что модель является линейным отображением вида

$$M(x) = \langle w, x\rangle + b = \sum_{j=1}^{n} w_j x_j + b, \quad \text{где } x \text{ вектор признаков из } \mathbb{R}^n,$$

где значения вектора w называют весами (англ. weight) модели, а число b — смещением (англ. bias). Набор значений из вектора w и числа b также называют набором параметров модели.

Чтобы найти наиболее подходящие параметры модели необходимы:

A — матрица размера $m \times n$, содержащая данные с признаками,

y — вектор размера $m \times n$ с известными значениями целевого признака.

Обозначим i-ю строку матрицы A за a_i, это строка содержит набор из n численных признаков i-го объекта,

y_i — это значение целевого признака i-го объекта.

Тогда критерий подбора параметров модели заключается в обеспечении наименьшей средне-квадратичной ошибки на данных, то есть

$$\sum_{i=1}^{m} |M(a_i) - y_i|^2 = \sum_{i=1}^{m} |\langle w, a_i\rangle + b - y_i|^2 \to \min.$$

Линейная регрессия

Для удобства значение смещения можно включить в вектор весов, если в матрицу данных добавить ещё один столбец с фиктивным признаком, значение этого признака для всех объектов равно 1.

Тогда можно считать, что модель имеет вид

$M(x) = \langle w, x \rangle$, а значение ошибки

$$\sum_{i=1}^{m} |M(a_i) - y_i|^2 = \sum_{i=1}^{m} |\langle w, a_i \rangle - y_i|^2 = \|A\,w - y\|^2.$$

Тем самым, для решения задачи линейной регрессии надо решить переопределённую СЛАУ с наилучшим в средне-квадратичном решением. Для этого можно использовать SVD, чтобы построить псевдообратную матрицу к матрице X.

PCA

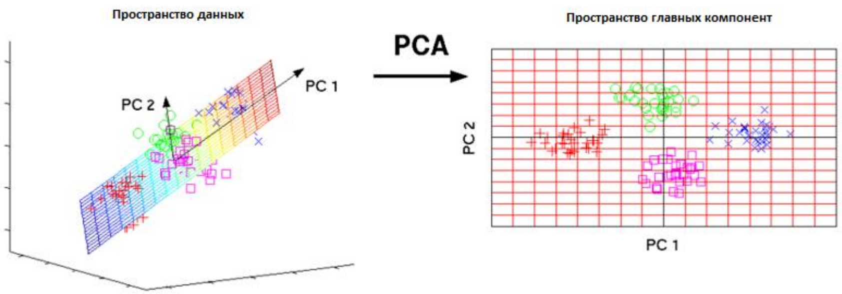
Метод главных компонент (англ. Principal Component Analysis, далее PCA) — это один из инструментов анализа данных. PCA строим новую координатную систему для более “экономного” представления данных, причём с иерархической структурой (то есть PCA позволяет выделить более существенные координаты от менее важных). PCA основан на SVD.

Рассмотрим матрицу данных вида $A = \begin{pmatrix} a_1 \\ \dots \\ a_m \end{pmatrix}$ размера $m \times n$, здесь a_i — это вектор-строка, содержащая набор из n признаков i-го объекта. Элементы матрицы обозначим $A = \{a_{i,j}\}_{i=1,\dots,m,\ j=1,\dots,n}$.

Цель PCA в том, чтобы получить матрицу A' размера $m \times d$, где $d < n$, которая представляет тот же набор данных, но с помощью меньшего числа признаков.

Математически это значит, что надо спроектировать исходные n-мерные вектора a_i на d-мерное аффинное подпространство в $\mathbb{R}^n$, так чтобы сохранить максимум информации о структуре данных.

Аффинное подпространство в $\mathbb{R}^n$ имеет вид $x + L$, где L обычное подпространство в $\mathbb{R}^n$, а x — некоторый вектор из $\mathbb{R}^n$.



PCA

Метод главных компонент (англ. Principal Component Analysis, далее PCA) — это один из инструментов анализа данных. PCA строим новую координатную систему для более “экономного” представления данных, причём с иерархической структурой (то есть PCA позволяет выделить более существенные координаты от менее важных). PCA основан на SVD.

Рассмотрим матрицу данных вида $A = \begin{pmatrix} a_1 \\ \dots \\ a_m \end{pmatrix}$ размера $m \times n$, здесь a_i — это вектор-строка, содержащая набор из n признаков i-го объекта. Элементы матрицы обозначим $A = \{a_{i,j}\}_{i=1,\dots,m,\ j=1,\dots,n}$.

Цель PCA в том, чтобы получить матрицу A' размера $m \times d$, где $d < n$, которая представляет тот же набор данных, но с помощью меньшего числа признаков.

Математически это значит, что надо спроектировать исходные n-мерные вектора a_i на d-мерное аффинное подпространство в $\mathbb{R}^n$, так чтобы сохранить максимум информации о структуре данных.

Аффинное подпространство в $\mathbb{R}^n$ имеет вид $x + L$, где L обычное подпространство в $\mathbb{R}^n$, а x — некоторый вектор из $\mathbb{R}^n$.

Слово "спроектировать" значит, что надо найти подходящее d-мерное аффинное подпространство в $\mathbb{R}^n$, которое описывается своим ОНБ, далее надо разложить каждый вектор a_i по этому базису, полученные d штук коэффициентов разложения для каждого вектора a_i и будут составлять новую матрицу A'.

Сохранение максимума информации о структуре данных означает, что требуется сохранить информацию об изменчивости данных и топологию данных, то есть взаимное расположение точек a_i относительно друг друга.

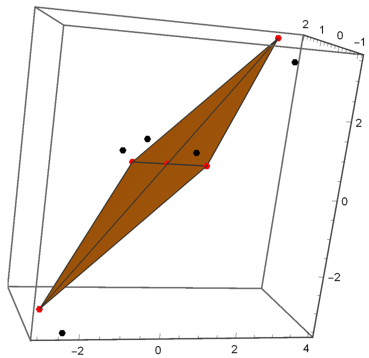
PCA

Представим иллюстрацию результата работы метода PCA.

Рассмотрим пример с 5-ю точками из $\mathbb{R}^3$.

Цель в том, чтобы снизить размерность представления до 2.

Ниже чёрными точками отмечены исходные точки в $\mathbb{R}^3$. Оранжевая плоскость — это подходящее двумерное аффинное подпространство в $\mathbb{R}^3$, красные точки — это проекции чёрных точек на выбранное двумерное аффинное подпространство.



Чтобы упростить задачу и вместо проекции на аффинное подпространство рассматривать обычные подпространства, надо центрировать данные. Для этого все вектора a_i, $i = 1, \ldots, m$, надо сдвинуть на усреднённый вектор, то есть на вектор

$$a^{\text{av}} = \frac{1}{m}\sum_{j=1}^{m} a_j. \quad A_{\text{cent}} = \begin{pmatrix} a_1 - a^{\text{av}} \\ \ldots \\ a_m - a^{\text{av}} \end{pmatrix}.$$

Тогда центральная точка сдвинутых данных будет находится в начале координат. Далее, будем считать, что в матрице A данные уже отцентрированы.

PCA

Как описывать d-мерные подпространства в $\mathbb{R}^n$?

Используем понятие прямоугольных ортогональных матриц. Пусть $1 \le d \le n$.

Обозначим за $O_{n,d}$ множество всех $n \times d$ матриц $W = (w_1, .., w_d)$, где столбцы этой матрицы $w_i \in \mathbb{R}^n$ попарно ортогональны $\langle w_j, w_k \rangle = \delta_{j-k}$ или $W^T W = I_d$.

Для каждой матрицы $W = (w_1, ..., w_d) \in O_{n,d}$ её столбцы являются ОНБ для своей линейной оболочки $\operatorname{span}\{w_1, ..., w_d\}$, которая является d-мерным подпространством в $\mathbb{R}^n$.

Таким образом, **d-мерные подпространства в $\mathbb{R}^n$ можно характеризовать матрицами из** $O_{n,d}$.

Для каждого вектора $w \in \operatorname{span} W$ существует единственное представление

$$w = \sum_{j=1}^{d} c_j w_j = W \begin{pmatrix} c_1 \\ ... \\ c_d \end{pmatrix} = W c, \quad c_j \in \mathbb{R}.$$

Компоненты вектора $c = (c_1, ..., c_d)$ называют координатами вектора w в координатной системе с базисом из векторов $w_1, .., w_d$. При этом получить коэффициенты $c_1, ..., c_d$ для вектора $w \in \operatorname{span} W$ можно так:

$W^T w = W^T W c = c$ или $w^T W = c^T$.

Вектор w из коэффициентов восстанавливается по формулам : $w = W c$ или $w^T = c^T W^T$.

Иными словами, вектор-строка w^T длины n теперь характеризуется вектором коэффициентов c^T длины d.

PCA

Обсудим применение SVD для метода PCA. Пусть A матрица данных размера $m\times n$. Пусть $\operatorname{rank} A = r$. Её неполное SVD имеет вид $A = U_r S_r V_r^T$, где матрицы U_r размера $m\times r$ и V_r размера $n\times r$, удовлетворяющие равенствам

где $U_r^T U_r = I_r$, $V_r^T V_r = I_r$, $S_r := \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}$, $\sigma_1 \geq \ldots \geq \sigma_r > 0$,

σ_j, $j = 1, \ldots, r$, — это ненулевые сингулярные числа матрицы A.

Пусть $1 \leq d < r$. Рассмотрим сокращённое SVD.

$A_{\text{cut}} = U_d S_d V_d^T$, $U_d = (u_1, \ldots, u_d)$, $V_d = (v_1, \ldots, v_d)$, $S_d = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$.

Отметим, что $V_d \in O_{n,d}$, а значит эта матрица характеризует некоторое d-мерное подпространство в $\mathbb{R}^n$. Спроектируем вектора-строки матрицы A на это подпространство. Проекции этих векторов будут вектор-строками матрицы $A V_d$. Обозначим

$Y_d := A V_d$

PCA

Определение. Пусть $1 \le d < r$. Вектора-строки $y_1, \ \ldots, \ y_m$, составляющие матрицу $Y_d = \begin{pmatrix} y_1 \\ \ldots \\ y_m \end{pmatrix}$, называют **представлением данных** матрицы A **в пространстве меньшей размерности**.

Нахождение этой матрицы Y_d по матрице A и составляет суть PCA.

Чтобы восстановить исходные данные по их представлению в пространстве меньшей размерности вектора, надо использовать строки матрицы Y_d как коэффициенты при векторах, составляющих столбцы матрицы V_d:

$A_{\text{approx}} := Y_d V_d^T = A V_d V_d^T$.

Конечно, строки матрицы A_{approx} после проделанного снижения размерности будут лишь приближением строк исходной матрицы A, так как при снижении размерности неизбежно теряется информация. Тем не менее, это приближение в некотором смысле оптимально.

PCA

Теорема (об оптимальном снижении размерности данных). Пусть $A \in \mathbb{R}^{m\times n}$ матрица центрированных данных с неполным SVD вида $A = U_r S_r V_r^T$.

Тогда представление данных матрицы A в меньшей размерности с помощью матрицы

$Y_d = \begin{pmatrix} y_1 \\ \dots \\ y_m \end{pmatrix} = A\, V_d$ удовлетворяет неравенству $\sum_{i=1}^{m} \| y_i V_d^T - a_i \|^2 \le \sum_{i=1}^{m} \| q_i W_d^T - a_i \|^2$

для всех матриц $W_d \in O_{n,d}$ и всех векторов-строк $q_1, \dots, q_m \in \mathbb{R}^d$.

Перефразируя эту теорему можно сказать, что

любое другое d-мерное подпространство в $\mathbb{R}^n$ с ОНБ в виде столбцов матрицы $W_d \in O_{n,d}$ и

любой набор из m штук векторов в этом подпространстве

не даёт приближение лучше, чем представление данных матрицы A в пространстве меньшей размерности с помощью матрицы Y_d.

Итак, d правых сингулярных векторов $v_1, \dots, v_d$ (столбцов матрицы V_d) задают лучшую координатную систему для оптимального снижения размерности n-мерных строк матрицы A до d-мерных, $d < n$. Эти вектора $v_1, \dots, v_d$ и называют **главными компонентами** матрицы данных A.

В частности, для снижения к 1-мерному подпространству используется 1-ая главная компонента v_1, порождающая подпространство для представления данных с помощью одной размерности.

Для снижения к 2-мерному представлению используются первые две главных компоненты v_1, v_2 и так далее.

PCA

Когда данные матрицы A не центрированы, то их необходимо центрировать, чтобы затем применять процесс снижения размерности данных к центрированной матрице A_{cent}.

Если мы получили приближения векторов матрицы A_{cent} векторами $Y_d\, V_d^T$, то приближения для векторов исходной матрицы получаются в виде

$$y_1\, V_d^T + a^{\text{av}},\ \ \ldots,\ \ y_m\, V_d^T + a^{\text{av}}.$$

Итоговая ошибка приближения данных исходной матрицы имеет вид

$$\left\| A_{\text{cent}}\, V_d\, V_d^* + \begin{pmatrix} a^{\text{av}} \\ \ldots \\ a^{\text{av}} \end{pmatrix} - A \right\|_F .$$

Выше метод PCA применялся для того, чтобы модифицировать пространство признаков. Но аналогичным образом можно решать задачу уменьшения данных, как снижение числа записей об объектах, то есть уменьшения числа строк матрицы m. Для этого надо лишь транспонировать эту матрицу и применить метод PCA к этой матрице.