Математика для Data Science. Линейная алгебра. Шпаргалка

Содержание

Первая неделя. Линейные отображения и матрицы

Векторное пространство

Вектор это упорядоченный набор действительных чисел. В машинном обучении объекты часто заменяют на векторы признаков объектов.

$\mathbb{R}^n$ – это множество всех упорядоченных наборов вида $(x_1, x_2, \dots, x_n)$, таких что $\forall i : x_i \in \mathbb{R}$. Каждый такой набор называется *вектором*. Множество $\mathbb{R}^n$ называется *векторным пространством*.

Обозначение. Мы будем обозначать вектор буквой со стрелочкой, а все его координаты – обычными буквами с индексами. Вот так: $\vec{x} = (x_1, x_2, \dots, x_n)$.

Иногда нам будет удобно представлять некоторые векторы не как строку $\vec{y} = (y_1, \dots, y_m)$, а как столбец. Вот так:

$$\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Сложение векторов из $\mathbf{R^n}$. Сложить можно любые два вектора из одного и того же векторного пространства $\mathbb{R}^n$. Результатом сложения также будет вектор из $\mathbb{R}^n$. Сумма векторов $(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ определяется так:

$(x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) = (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n) \in \mathbb{R}^n$.

Умножение вектора на число. На $\mathbb{R}^n$ определена операция умножения вектора на число. Любой вектор из $\mathbb{R}^n$ можно умножить на любое число, и результат умножения тоже будет вектором из $\mathbb{R}^n$. Число в этом случае называют *скаляром*. Умножение вектора $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ на число $c \in \mathbb{R}$ определяется так:

$c(x_1, x_2, \dots, x_n) = (cx_1, cx_2, \dots, cx_n) \in \mathbb{R}^n$.

Линейные отображения

Линейным отображением векторного пространства V в векторное пространство W называется функция $f : V \to W$ удовлетворяющая следующим двум условиям:

1. $f(\vec{x}) + f(\vec{y}) = f(\vec{x} + \vec{y})$ для любых $\vec{x}, \vec{y} \in V$,
2. $f(c\vec{x}) = cf(\vec{x})$ для любого $\vec{x} \in V$ и любого числа $c \in \mathbb{R}$.

Общий вид некоторых линейных отображений

1. Любое линейное отображение $f : \mathbb{R} \to \mathbb{R}$ задаётся одним числом $f(1)$: а именно, $f(\vec{x}) = x \cdot f(1)$.

2. Любое линейное отображение $f : \mathbb{R} \to \mathbb{R}^m$ задаётся столбцом высоты m, состоящим из чисел. Этот столбец равен $f(1)$. При этом $f(\vec{x}) = x \cdot f(1)$.

3. Любое линейное отображение $f : \mathbb{R}^n \to \mathbb{R}$ задаётся строкой длины n, состоящей из чисел.

Обозначение.

- $\vec{e}_1 := (1, 0, 0, \dots, 0)$
- $\vec{e}_2 := (0, 1, 0, \dots, 0)$
- $\vdots$
- $\vec{e}_n := (0, 0, 0, \dots, 1)$

Другими словами, у вектора $\vec{e}_i$ равна единице i-ая координата, а все остальные координаты равны нулю.
Рассмотрим любой вектор $\vec{x} = (x_1, \dots, x_n)$. По определению векторного пространства $\mathbb{R}^n$, выполнено:
$(x_1, \dots, x_n) = (x_1, 0, \dots, 0) + (0, x_2, \dots, 0) + \dots + (0, 0, \dots, x_n) = x_1\vec{e}_1 + x_2\vec{e}_2 + \dots + x_n\vec{e}_n$.
Обозначение. В таких случаях говорят, что вектор $\vec{x}$ *выражается* через векторы $\vec{e}_1, \dots, \vec{e}_n$.

Пусть дано линейное отображение $f : \mathbb{R}^n \to \mathbb{R}$ и для каждого i $f(\vec{e}_i) = a_i$, где $a_i \in \mathbb{R}$. Тогда f определяется набором чисел $a_1, \dots, a_n$: $f(\vec{x}) = f(x_1\vec{e}_1 + \dots + x_n\vec{e}_n) = f(x_1\vec{e}_1) + \dots + f(x_n\vec{e}_n) = x_1 f(\vec{e}_1) + \dots + x_n f(\vec{e}_n) = x_1a_1 + \dots + x_na_n = a_1x_1 + \dots + a_nx_n$.

Вычисление значения функции f мы будем записывать как *действие* строки $(a_1, \dots, a_n)$ на столбец $\vec{x}$:

$$f(\vec{x}) = (a_1, \dots, a_n)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} := a_1x_1 + \dots + a_nx_n = \sum_{i=1}^{n} a_ix_i.$$

Можно говорить, что мы *умножаем* строку $(a_1, \dots, a_n)$ на столбец $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$.

Матрицы

Любое линейное отображение $f : \mathbb{R}^n \to \mathbb{R}^m$ задаётся матрицей с m строками и n столбцами.

$$A := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

где $a_{ij} \in \mathbb{R}$ для всех i, j. Числа a_{ij} называют *элементами* матрицы.

Построим по матрице A линейное отображение:
i-ой строке матрицы $(a_{i1}, a_{i2}, \dots, a_{in})$ соответствует линейное отображение $f_i : \mathbb{R}^n \to \mathbb{R}$:

$$f_i(\vec{x}) = (a_{i1}, a_{i2} \dots, a_{in})\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} := a_{i1}x_1 + a_{i2}x_2 + \dots + a_{in}x_n = \sum_{j=1}^{n} a_{ij}x_j \in \mathbb{R}.$$

Пользуясь отображениями $f_1, \dots, f_m$, мы так определяем действие отображения f на векторе $\vec{x}$:

$$f(\vec{x}) = \begin{pmatrix} f_1(\vec{x}) \\ f_2(\vec{x}) \\ \vdots \\ f_m(\vec{x}) \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{pmatrix} \in \mathbb{R}^m.$$

Обозначение. Полученный вектор $f(\vec{x})$ называется результатом *действия* матрицы A на векторе $\vec{x}$, и обозначается $A\vec{x}$.

Записывают действие матрицы A на векторе $\vec{x}$ так:

$$A\vec{x} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \sum_{j=1}^{n} a_{2j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{pmatrix} \in \mathbb{R}^m$$

Также можно говорить, что мы *умножили* A на $\vec{x}$ и получили $A\vec{x}$.

Как воспринимать столбцы матрицы: для любого j вектор $f(\vec{e}_j) = A\vec{e}_j$ совпадает с j-ым столбцом матрицы A. Таким образом, столбцы матрицы — это те векторы, в которые f переводит единичные векторы, направленные вдоль осей координат. $A = \begin{pmatrix} f(\vec{e}_1) & f(\vec{e}_2) & \dots & f(\vec{e}_n) \end{pmatrix}$

Итак, любая матрица задаёт линейное отображение. При этом любое линейное отображение задаётся какой-то матрицей. Кроме того, разные матрицы задают разные линейные отображения. Значит, есть *взаимно однозначное соответствие* между матрицами и линейными отображениями. То есть каждой матрице соответствует ровно одно линейное отображение, и каждому линейному отображению поставлена в соответствие ровно одна матрица.

Умножение матриц

Пусть определены два линейных отображения: $f : \mathbb{R}^n \to \mathbb{R}^m$ и $g : \mathbb{R}^m \to \mathbb{R}^k$.

Их *композицией* называется отображение из $\mathbb{R}^n$ в $\mathbb{R}^k$ и обозначается $g \circ f$. Действует $g \circ f$ на любом векторе $\vec{x} \in \mathbb{R}^n$ так:

1. применяя f, получаем вектор $f(\vec{x}) \in \mathbb{R}^m$
2. применяя к полученному вектору g, получаем вектор $g(f(\vec{x})) \in \mathbb{R}^k$

Вектор $g(f(\vec{x})) \in \mathbb{R}^k$ и называется результатом действия $g \circ f$ на $\vec{x}$.

Композиция двух линейных преобразований тоже будет линейным преобразованием.

Если отображение $f : \mathbb{R}^n \to \mathbb{R}^m$ задано матрицей A, а отображение $g : \mathbb{R}^m \to \mathbb{R}^k$ — матрицей B, то $(g \circ f)(\vec{x}) = g(f(\vec{x})) = g(A(\vec{x})) = B(A(\vec{x}))$. Матрицу преобразования $g \circ f$ обозначают BA и называют *произведением* матрицы B и матрицы A.

Правило умножения матриц

На пересечении i-ой строки и j-ого столбца произведения двух матриц стоит произведение i-ой строки первой матрицы и j-ого столбца второй матрицы.

При этом произведение BA определено, только если число столбцов матрицы B совпадает с числом строк матрицы A.

Умножение называется *коммутативным*, если $B \cdot A = A \cdot B$ для любых A и B. Умножение матриц не является коммутативным: то есть, бывают матрицы A и B размера n на n, что $AB \neq BA$.

Примеры некоторых типов матриц

1. Матрица, на *главной диагонали* которой стоят единицы, а на всех остальных местах стоят нули, называется *единичной матрицей*. Чаще всего её обозначают буквами E или I. Если важно подчеркнуть, что матрица именно размера n на n, то добавляют нижний индекс: E_n, I_n.

$$E_n = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

2. Иногда матрицы удобнее записывать в виде нескольких блоков, каждый из которых соответствует меньшей матрице. Например, пусть

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{12} \end{pmatrix}, Y = \begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{12} \end{pmatrix}.$$

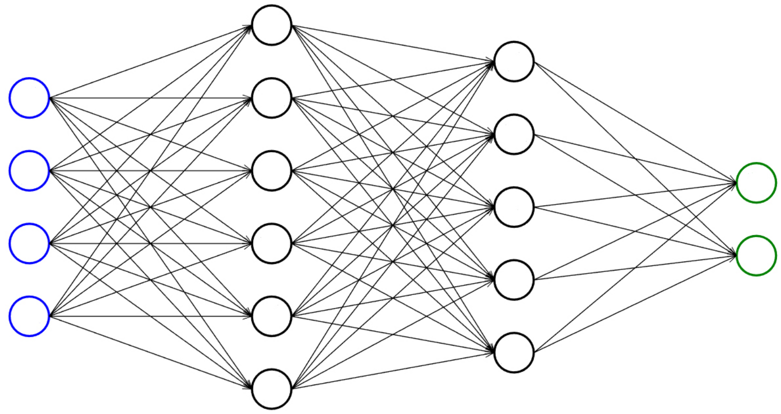
Тогда мы можем так определить *блочную* матрицу $\begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix}$:

$$\begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} := \begin{pmatrix} x_{11} & x_{12} & 0 & 0 \\ x_{21} & x_{12} & 0 & 0 \\ 0 & 0 & y_{11} & y_{12} \\ 0 & 0 & y_{21} & y_{12} \end{pmatrix}.$$

При этом два нуля в этой блочной матрице обозначают заполненные нулями матрицы размера 2 на 2.

Вторая неделя. Пространства, базисы и ранг

Нейронные сети



Кружки на картинках — это *нейроны*. Слой нейронов, который выделен синим на картинке, называют *входным, распределительным* или *нулевым*. Числа, принимаемые на вход, будут называться *состояниями* нейронов нулевого слоя.

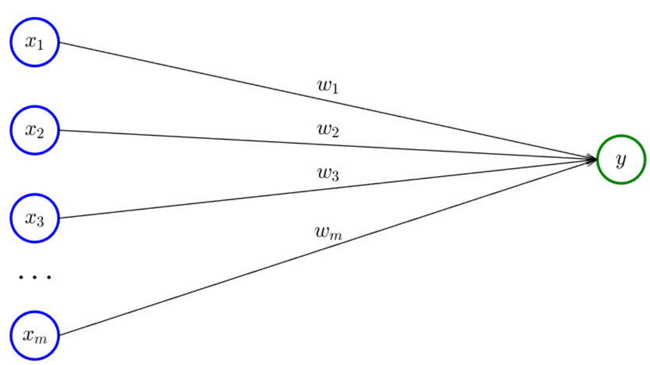
Каждой *связи*, то есть каждому ребру между нейронами, соответствует некоторое преобразование (функция) — из числа в число. Слои в сети нумеруются слева направо и по состоянию нейронов предыдущего слоя с помощью связей мы вычисляем состояние нейронов следующего слоя. В результате, каждому нейрону будет присвоено *состояние* (число). Состояние нейронов последнего слоя — это *выход* нейронной сети.

Каждая нейросеть — это пара:

1. **Архитектура** — количество слоев, число нейронов на каждом слое, описание типов связей между нейронами. На картинке сверху нарисована именно архитектура нейронной сети (правда, без описания типов связей)

2. **Веса.** Каждой связи соответствует функция из числа в число. Архитектура задает только тип таких функций. То есть некоторый класс функций, которые могут быть использованы: например, линейные функции. У этих функций есть параметры, также называемые весами.

Рассмотрим простейшую нейросеть: входной слой $x_1, x_2, \dots, x_m$ и один нейрон y на выходном слое.

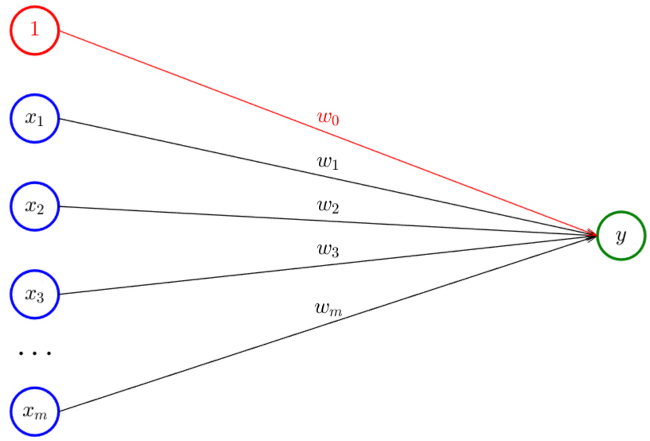


Нейросеть с *линейной связью* работает следующим образом: $y = w_1x_1 + w_2x_2 + \dots + w_mx_m = \sum_{i=1}^{m} w_ix_i$. Или, то же самое через умножение матриц:

$$y = \begin{pmatrix} w_1 & w_2 & \dots & w_m \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

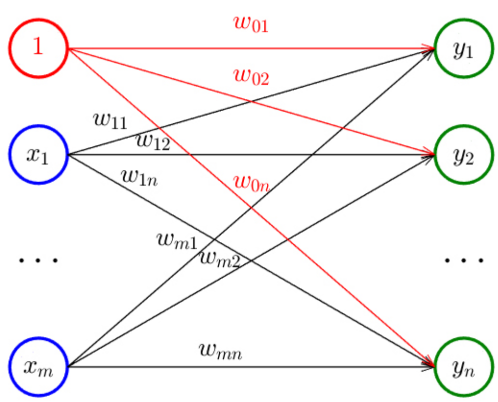
Веса в данном случае — это значения коэффициентов $w_1, w_2, \ldots, w_m \in \mathbb{R}$.

Для более точных предсказаний при вычислении состояния каждого нейрона мы будем добавлять ещё *сдвиг* (*bias*) — нейрон с постоянным единичным состоянием.



Получим $y = \begin{pmatrix} w_0 & w_1 & w_2 & \ldots & w_m \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = w_0 + \sum\limits_{i=1}^{m} w_i x_i$

Однослойная нейронная сеть



Пусть $x_1, x_2, \ldots, x_m$ — вход нейронной $y_1, y_2, \ldots, y_n$ — состояния первого слоя (линейного слоя), он же выход, w_{ij} — вес на связи между i-ым нейроном входа и j-ым нейроном выхода. Тогда $\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} w_{01} + \sum\limits_{i=1}^{m} w_{i1} x_i \\ w_{02} + \sum\limits_{i=1}^{m} w_{i2} x_i \\ \vdots \\ w_{0n} + \sum\limits_{i=1}^{m} w_{in} x_i \end{pmatrix}$

Пусть A — матрица $n \times m$. *Транспонированной* матрицей для A будет матрица B размеров $m \times n$ такая, что $A_{ij} = B_{ji}$ для всех $1 \leqslant i \leqslant n$ и $1 \leqslant j \leqslant m$. Транспонированная матрица обозначается A^T.

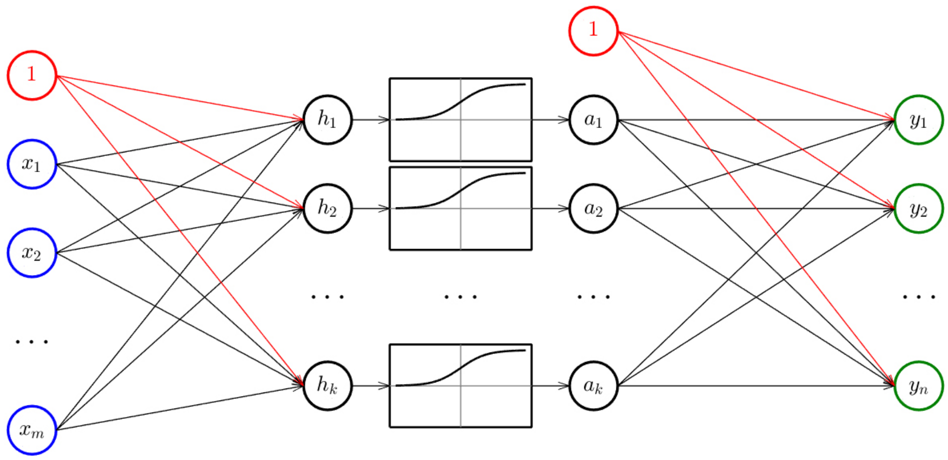
Таким образом, обозначив $W := \begin{pmatrix} w_{01} & w_{02} & \ldots & w_{0n} \\ w_{11} & w_{12} & \ldots & w_{1n} \\ w_{21} & w_{22} & \ldots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \ldots & w_{mn} \end{pmatrix}$, получаем формулу $\vec{y} = W^T \vec{x}$, где $\vec{x} = (x_1, x_2, \ldots, x_m)^T$

и $\vec{y} = (y_1, y_2, \ldots, y_m)^T$.

Нейросети называются *эквивалентными* в случае, если для любого входного вектора их выход одинаков.

Утверждение. N подряд идущих линейных слоев эквивалентны одному линейному слою.

Поэтому между линейными слоями применяют нелинейную функцию к состояниям нейронов — *функцию активации*.



Обозначим $\vec{x}' = (1, x_1, x_2, \ldots, x_m)^T$ — входной вектор расширенный нейроном сдвига,

$\vec{h} = (h_1, h_2, \ldots, h_k)^T$ — вектор состояний скрытого слоя

$\vec{a} = (a_1, a_2, \ldots, a_k)^T$ — вектор состояний скрытого слоя после применения активации (способ нахождения a_i описан ниже)

$\vec{a}' = (1, a_1, a_2, \ldots, a_k)^T$ — вектор $\vec{a}$ расширенный нейроном сдвига

$\vec{y} = (y_1, y_2, \ldots, y_n)^T$ — выход нейросети

W — матрица весов первого линейного слоя, ее размер $k \times (m+1)$

U — матрица весов второго линейного слоя, её размер $n \times (k+1)$

f — функция активации, применяемая между первым и вторым линейным слоем

Нейросеть с одним скрытым слоем и активацией работает следующим образом:

- На основе входа расширенного нейроном сдвига вычисляются состояния скрытого слоя: $\vec{h} = W^T\vec{x}'$
- К каждой компоненте вектора $\vec{h}$ применяется функция активации f — получаем вектор $\vec{a}$: $a_i = f(h_i)$
- По активированным состояниям скрытого слоя с нейроном сдвига вычисляется выход нейросети: $\vec{y} = U^T\vec{a}'$

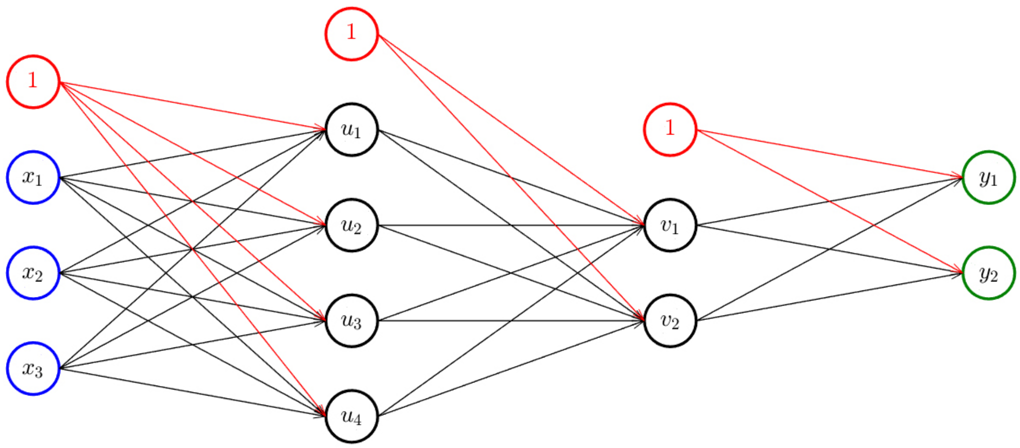
Перцептрон — это нейросеть с одним скрытым слоем, первый и второй слой линейные, между ними функция активации, выход один.

Теорема. Любая непрерывная функция многих переменных может быть приближена перцептроном с сигмоидальной функцией активации с любой наперед заданной точностью. Пример сигмоидальной функции это $f(x) = \frac{1}{1+e^{-x}}$.

Примеры функций активации

1. $\mathrm{ReLU} = \max(0, x)$
2. $\text{Leaky ReLU} = \begin{cases} x, & \text{если } x > 0 \\ 0.01x, & \text{иначе} \end{cases}$
3. $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ — гиперболический тангенс
4. $\sigma(x) = \frac{1}{1+e^{-x}}$ сигмоидальная (логистическая) функция активации
5. $\mathrm{Softmax}(z_1, z_2, \ldots, z_n) = \left(\frac{e^{z_1}}{\sum\limits_{i=1}^{n} e^{z_i}}, \frac{e^{z_2}}{\sum\limits_{i=1}^{n} e^{z_i}}, \ldots, \frac{e^{z_n}}{\sum\limits_{i=1}^{n} e^{z_i}} \right)$

Многослойные нейронные сети строятся по тем же принципам, что мы уже разобрали: последовательное умножение на матрицы и применение функций активации. Все слои, кроме нулевого и последнего называются скрытыми слоями.



Подпространства и линейные комбинации

Линейной комбинацией векторов $\vec{a}_1, \ldots, \vec{a}_k$ называется любой вектор вида $\lambda_1\vec{a}_1 + \lambda_2\vec{a}_2 + \cdots + \lambda_k\vec{a}_k$, где $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$.

Множество V всех линейных комбинаций векторов $\vec{a}_1, \ldots, \vec{a}_k \in \mathbb{R}^n$ называется линейным подпространством, *порождённым* векторами $\vec{a}_1, \ldots, \vec{a}_k$. Набор векторов $\vec{a}_1, \ldots, \vec{a}_k$ называется *порождающим*.

Комментарий. Слово "линейное" часто опускается, так что иногда мы будем писать просто "подпространство".

Нестрогое определение. *Векторное пространство* это множество V с определёнными на нём операциями сложения и умножения на число, такими что

1. Если $\vec{x} \in V$, $\vec{y} \in V$, то и $\vec{x} + \vec{y} \in V$.
2. Если $\vec{x} \in V$ и $c \in \mathbb{R}$, то $c\vec{x} \in V$.

Если V — линейное подпространство в $\mathbb{R}^n$, порождённое векторами $\vec{a}_1, \ldots, \vec{a}_k$, то оно является векторным пространством.

Линейная зависимость

Линейные комбинации, в которых все коэффициенты равны нулю, называются *тривиальными*.

Набор векторов $\vec{a}_1, \ldots, \vec{a}_k$ называется *линейно зависимым*, если $\vec{0}$ можно представить в виде нетривиальной линейной комбинации векторов из этого набора.

Другими словами, набор векторов $\vec{a}_1, \ldots, \vec{a}_k$ называется линейно зависимым, если существуют такие числа $\lambda_1, \ldots, \lambda_k$, что

- $\vec{0} = \lambda_1\vec{a}_1 + \cdots + \lambda_k\vec{a}_k$
- хотя бы одно из чисел $\lambda_1, \ldots, \lambda_k$ не равно нулю

Наборы, которые не являются линейно зависимыми, называют *линейно независимыми*.

Утверждение. Если набор линейно зависим, то хотя бы один из его векторов представляется в виде линейной комбинации оставшихся векторов. Другими словами, хотя бы один из векторов набора можно выразить через другие.

Обратное утверждение тоже верно: если один из векторов набора выражается через другие, то набор линейно зависим.

Утверждение "V порождено векторами $\vec{a}_1, \ldots, \vec{a}_k$" записывают так: $V = \mathrm{Span}\{\vec{a}_1, \ldots, \vec{a}_k\}$. То же самое можно сказать так: V — *линейная оболочка* векторов $\vec{a}_1, \ldots, \vec{a}_k$.

Свойства линейной оболочки векторов

1. $\mathrm{Span}\{\vec{a}_1\} = \mathrm{Span}\{\vec{a}_1, c\vec{a}_1\}$, где c – любое число

2. $\mathrm{Span}\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_k\} = \mathrm{Span}\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_k, c_1\vec{a}_1 + c_2\vec{a}_2 + \dots + c_k\vec{a}_k\}$, где $c_1, \dots, c_k$ – любые числа

3. $\mathrm{Span}\{\vec{a}_1, \vec{a}_2, \dots, \vec{a}_k\} = \mathrm{Span}\{\vec{a}_1 + c\vec{a}_2, \vec{a}_2, \dots, \vec{a}_k\}$, где c – любое число

Базис, размерность, ранг

Базис векторного пространства это упорядоченный набор векторов, такой что любой вектор этого пространства единственным способом представляется в виде линейной комбинации векторов из этого набора.

Пусть дан вектор $\vec{x} \in V$ и базис $\vec{a}_1, \dots, \vec{a}_k$ пространства V. Тогда $\vec{x} = \lambda_1\vec{a}_1 + \dots + \lambda_k\vec{a}_k$ для некоторых $\lambda_1, \dots, \lambda_k$. Упорядоченный набор чисел $(\lambda_1, \dots, \lambda_k)$ называется *координатами* вектора $\vec{x}$ в базисе $\vec{a}_1, \dots, \vec{a}_k$.

Набор векторов $\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n$ называется *стандартным* базисом $\mathbb{R}^n$. Напомним, что $\vec{e}_i$ — это вектор, у которого стоит единица на i-ом месте, а все остальные координаты равны нулю.

Следующие три утверждения эквивалентны:

1. набор является базисом V

2. набор линейно независим и порождает V

3. набор порождает V и при удалении любого элемента из набора оставшиеся векторы не порождают V

Теорема. Все базисы векторного пространства содержат одно и то же число элементов.

Размерностью векторного пространства V называется число элементов в базисе V.

Ранг матрицы это максимальное число линейно независимых столбцов матрицы.

Выражение "ранг матрицы A равен k" записывают так: $\mathrm{rank}(A) = k$. Иногда можно встретить обозначение $\mathrm{rk}(A) = k$.

Утверждение. Ранг матрицы A равен размерности пространства, порождённого столбцами A.

Метод Гаусса

Пусть необходимо найти ранг данной матрицы. Будем построчно приводить её к виду, в котором легко найти ранг.

Введём два *элементарных преобразования* матрицы, которые не меняют ранг:

1. поменять местами два столбца матрицы

2. прибавить к одному из столбцов другой столбец, умноженный на число

Преобраование последней строки

Пусть в последней строке есть хотя бы одно ненулевое число. Применив 1-ое элементарное преобразование, переставим столбцы так, чтобы это ненулевое число стояло в последнем столбце. Тем самым, наша матрица сейчас выглядит так:

$$\begin{pmatrix} * & * & \dots & * & * \\ * & * & \dots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \dots & * & * \\ a_1 & a_2 & \dots & a_{n-1} & a_n \end{pmatrix}$$

где $a_n \neq 0$, а звёздочками обозначены числа, которые нам сейчас не интересны.

Для каждого $i < n$ вычтем из i-ого столбца последний столбец, умноженный на $\frac{a_i}{a_n}$. Получим такую матрицу:

$$\begin{pmatrix} * & * & \dots & * & * \\ * & * & \dots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \dots & * & * \\ (a_1 - \frac{a_1}{a_n}a_n) & (a_2 - \frac{a_2}{a_n}a_n) & \dots & (a_{n-1} - \frac{a_{n-1}}{a_n}a_n) & a_n \end{pmatrix} = \begin{pmatrix} * & * & \dots & * & * \\ * & * & \dots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ * & * & \dots & * & * \\ 0 & 0 & \dots & 0 & a_n \end{pmatrix}$$

Комментарий. Если у изначальной матрицы последняя строка состояла только из нулей, то найдём первую снизу строку, в которой есть хотя бы одно ненулевое число и проведём с ней описанную процедуру.

Преобраование предпоследней строки

Далее, посмотрим на часть матрицы, составленную из всех столбцов, кроме последнего. Проведём с ней такую же процедуру, как и на прошлом шаге. А именно:

- переставим столбцы так, чтобы b_{n-1} был не равен нулю: $\begin{pmatrix} * & * & \dots & * & * & * \\ * & * & \dots & * & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ * & * & \dots & * & * & * \\ b_1 & b_2 & \dots & b_{n-2} & b_{n-1} & * \\ 0 & 0 & \dots & 0 & 0 & a_n \end{pmatrix}$
- вычитая предпоследний столбец из предыдущих, занулим все числа левее b_{n-1}: $\begin{pmatrix} * & * & \dots & * & * & * \\ * & * & \dots & * & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ * & * & \dots & * & * & * \\ 0 & 0 & \dots & 0 & b_{n-1} & * \\ 0 & 0 & \dots & 0 & 0 & a_n \end{pmatrix}$

Если встречались нулевые строки, то мы их пропускаем.

Преобраование остальных строк

Зафиксируем последние два столбца, и действуем, как и на предыдущих двух шагах. Продолжаем повторять, пока не закончатся строки или столбцы.

Утверждение. Ранг матрицы, полученной методом Гаусса, равен количеству ненулевых столбцов.

Кроме того, метод Гаусса позволяет, например, исследовать линейную зависимость, находить базис и размерность пространства.

Теорема. Максимальное число линейно независимых столбцов матрицы равно максимальному числу линейно независимых строк матрицы.

Следствие из теоремы. Элементарные преобразования в методе Гаусса можно применять не только к столбцам, но и к строкам.

Третья неделя. Определитель и скалярное произведение

Определитель

Матрица 2×2

Поставим в соответствие произвольной матрице из действительных чисел $\begin{pmatrix} a & c \\ b & d \end{pmatrix}$ параллелограмм натянутый на векторы столбцы этой матрицы. Под словом "натянутый" подразумевается, что две стороны параллелограмма образуются векорами $\begin{pmatrix} a \\ b \end{pmatrix}$ и $\begin{pmatrix} c \\ d \end{pmatrix}$, а остальные получаются из них параллельным переносом.

Площадь этого параллелограмма в общем случае равна $|ad - bc|$.

Базис на плоскости $\vec{e}_1, \vec{e}_2$ называется *правым*, если кратчайший поворот от вектора $\vec{e}_1$ к вектору $\vec{e}_2$ происходит против часовой стрелки. В противном случае базис называется *левым*.

В случае, если столбцы матрицы $\begin{pmatrix} a \\ b \end{pmatrix}$ и $\begin{pmatrix} c \\ d \end{pmatrix}$ задают левую пару векторов, то *ориентированная площадь* параллелограмма будет отрицательной. Если правую пару векторов — положительной.

Ориентированная площадь параллограмма натянутого на столбцы матрицы $\begin{pmatrix} a & c \\ b & d \end{pmatrix}$ равняется $ad - bc$.

Матрица 3×3

Рассмотрим параллелепипед, натянутый на столбцы матрицы $\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}$. Его объём будет равен $|x_{11}x_{22}x_{33} + x_{12}x_{23}x_{31} + x_{13}x_{21}x_{32} - x_{13}x_{22}x_{31} - x_{23}x_{32}x_{11} - x_{33}x_{12}x_{21}|$.

Пусть $\vec{x}, \vec{y}$ и $\vec{z}$ — базис в трехмерном пространстве. Представим, что мы смотрим на плоскость, в которой лежат векторы $\vec{x}$ и $\vec{y}$, из конца вектора $\vec{z}$. Тогда ориентация трехмерного базиса совпадает с двумерной ориентацией пары векторов $\vec{x}$ и $\vec{y}$.

Ориентированный объем параллелепипеда натянутого на столбцы матрицы $\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}$ равняется $x_{11}x_{22}x_{33} + x_{12}x_{23}x_{31} + x_{13}x_{21}x_{32} - x_{13}x_{22}x_{31} - x_{23}x_{32}x_{11} - x_{33}x_{12}x_{21}$.

Матрица $\mathbf{n} \times \mathbf{n}$

Пусть M — квадратная матрица действительных чисел размера $n \times n$. Ее *определитель*, то есть n-мерный ориентированный объем параллелепипеда натянутого на столбцы этой матрицы имеет следующие свойства:

1. Ориентированный объем единичного куба равен 1, то есть определитель единичной матрицы равен 1. Или коротко $\det E = 1$.

   Единичным n-мерным кубом называют параллелепипед, натянутый на столбцы единичной матрицы $n \times n$.

2. Если один из столбцов матрицы домножить на $\lambda \in \mathbb{R}$, то n-мерный ориентированный объем (определитель) тоже домножится на λ.

3. Если поменять два столбца матрицы местами, то определитель поменяет знак. Это свойство называют *кососимметричностью* по столбцам.

4. Объем параллелепипеда натянутого на векторы $\vec{a}_1, (\vec{b} + \vec{c}), \vec{a}_3, \ldots, \vec{a}_n$ равен сумме объемов параллелограммов натянутых на векторы $\vec{a}_1, \vec{b}, \vec{a}_3, \ldots, \vec{a}_n$ и векторы $\vec{a}_1, \vec{c}, \vec{a}_3, \ldots, \vec{a}_n$.

Утверждения

1. В результате линейного преобразования A объем любой фигуры меняется в $\det A$ раз.

2. Для квадратных матриц одного размера выполнено $\det(BA) = \det B \cdot \det A$.

Матрица, у которой все элементы ниже главной диагонали равны нулю, называется *верхнетреугольной*:
$$\begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}$$

Как считать определитель

1. Применяем к матрице метод Гаусса и получаем верхнетреугольную матрицу. Определитель этой верхнетреугольной матрицы равен или противоположен определителю изначальной матрицы (в зависимости от того, сколько раз мы использовали элементарные преобразования первого типа).
2. Если на диагонали получается 0, то определитель равен нулю. Иначе применяем преобразования второго типа, чтобы превратить верхнетреугольную матрицу в диагональную:

   Вычтем из столбцов со 2-ого по n-ый первый столбец домноженный на $\frac{1}{a_{11}}$, потом вычтем из столбцов с 3-его по n-ый второй столбец домноженный на $\frac{1}{a_{22}}$ и так далее.

   Определитель при этом не меняется.
3. Считаем определитель диагональной матрицы.

Переход в другой базис и обратная матрица

Квадратная матрица A называется *матрицей перехода* от базиса $\boldsymbol{h}$ к базису $\boldsymbol{g}$ в $\mathbb{R}^n$, если для всех векторов $\vec{v} \in \mathbb{R}^n$ выполнено следующее. Если $(v_1, v_2, \dots v_n)$ — координаты вектора в базисе h, то координаты $\vec{v}$ в базисе $\boldsymbol{g}$ — это $A \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$.

В частности, матрицей перехода от базиса $\boldsymbol{f}$ к стандартному является матрица, в которой по столбцам записаны координаты базиса $\boldsymbol{f}$.

Для квадратной матрицы A матрица B называется *обратной матрицей*, если $BA = E$. Обозначение A^{-1}.

Свойства обратной матрицы

1. Для любой матрицы полного ранга существует обратная матрица.
2. При домножении матрицы B на A^{-1} слева или справа и результат не обязательно будет одинаковый: $A^{-1}B$ не обязательно равно BA^{-1}.
3. Если $BA = E$ и $AC = E$ то $B = C$. То есть любая матрица, которая является обратной слева, будет и обратной справа. И наоборот.
4. Обратная матрица единственна. То есть, если $B_1A = E$ и $B_2A = E$, то $B_1 = B_2$.
5. Для каждой матрицы полного ранга единственным образом определена обратная матрица.

Пусть f — произвольный базис в $\mathbb{R}^n$, F — матрица перехода от f к стандартному базису.

Если $(v_1^f, v_2^f, \dots, v_n^f)$ — координаты вектора $\vec{v}$ в базисе f, то $\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = F \begin{pmatrix} v_1^f \\ v_2^f \\ \vdots \\ v_n^f \end{pmatrix}$ — координаты в стандартном базисе.

С другой стороны, F^{-1} — матрица перехода от стандартного базиса к f и при этом $\begin{pmatrix} v_1^f \\ v_2^f \\ \vdots \\ v_n^f \end{pmatrix} = F^{-1} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$

Приведение матрицы к единичной

1. Пусть дана квадратная матрица A полного ранга. С помощью первых двух элементарных преобразований приведём её к диагональному виду.

2. Применим *третье элементарное преобразование*: умножение столбца (или строки) на ненулевое число. С его помощью мы каждый элемент на диагонали матрицы сможем привести к единице, домножив i-ый столбец на $\frac{1}{a_{ii}}$.

Метод Жордана-Гаусса поиска обратной матрицы

Для всей цепочки элементарных преобразований в описанном выше алгоритме существует последовательность матриц $M_1, M_2, \dots, M_k$ такая, что M_1 соответствует первому примененному элементарному преобразованию, M_2 — второму и т.д. Так как $M_k M_{k-1} \dots M_2 M_1 A = E$. Значит, $M_k M_{k-1} \dots M_2 M_1$ — как раз и есть A^{-1}.

Матрица перехода в общем случае

Если

- T_f — матрица перехода от $\boldsymbol{f}$ к стандартному базису, то есть записанные по столбцам векторы базиса $\boldsymbol{f}$
- T_g — матрица перехода от $\boldsymbol{g}$ к стандартному базису, то есть записанные по столбцам векторы базиса $\boldsymbol{g}$

то

- $T_g^{-1} T_f$ — матрица перехода от f к g
- $T_f^{-1} T_g$ — матрица перехода от $\boldsymbol{g}$ к $\boldsymbol{f}$

Линейное преобразование записанное в другом базисе Пусть в базисе $\boldsymbol{f}$ некоторое линейное преобразование t задано при помощи матрицы A и матрица перехода от $\boldsymbol{g}$ к $\boldsymbol{f}$ это C.. Тогда матрица преобразования t в базисе $\boldsymbol{g}$ записывается как $C^{-1}AC$.

Длина, углы и скалярное произведение

Длина вектора $\vec{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ определяется так: $||\vec{x}|| = \sqrt{x_1^2 + \dots + x_n^2}$.

Свойства длины

1. Длина любого вектора неотрицательна.

2. Длина вектора равна нулю тогда и только тогда, когда вектор равен $\vec{0}$.

3. Длина вектора $\alpha\vec{x}$ равна $|\alpha| \cdot ||\vec{x}||$.

4. Неравенство треугольника: $||\vec{x}|| + ||\vec{y}|| \geq ||\vec{x} + \vec{y}||$ для любых $\vec{x}, \vec{y} \in \mathbb{R}^n$.

Теорема Пифагора. Дан треугольник со сторонами a, b, c. Угол между сторонами с длинами a и b прямой, если и только если $a^2 + b^2 = c^2$.

Векторы, угол между которыми прямой, называются *ортогональными* или *перпендикулярными.*
Нулевой вектор считается одновременно ортогональным и параллельным любому вектору.

Теорема косинусов. Дан треугольник со сторонами a, b, c. Угол между сторонами с длинами a и b равен α. Тогда выполнено соотношение $a^2 + b^2 - 2ab\cos\alpha = c^2$.
В частности, $\cos\alpha = \frac{||\vec{x}||^2 + ||\vec{y}||^2 - ||\vec{x} - \vec{y}||^2}{2||\vec{x}|| \cdot ||\vec{y}||}$ при $||\vec{x}|| > 0$ и $||\vec{y}|| > 0$.

Скалярным произведением векторов $\vec{x} \in \mathbb{R}^n$ и $\vec{y} \in \mathbb{R}^n$ называется число
$\langle \vec{x}, \vec{y} \rangle := x_1 y_1 + x_2 y_2 + \dots + x_n y_n = \sum\limits_{i=1}^{n} x_i y_i$.

Свойства скалярного произведения

1. $\langle \vec{x}, \vec{x} \rangle = ||\vec{x}||^2$
2. $\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle$, то есть скалярное произведение коммутативно.
3. $\langle c\vec{x}, \vec{y} \rangle = c\langle \vec{x}, \vec{y} \rangle$ для любого $c \in \mathbb{R}$
4. $\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle$
5. $||\vec{x} + \vec{y}||^2 = \langle \vec{x}, \vec{x} \rangle + 2\langle \vec{x}, \vec{y} \rangle + \langle \vec{y}, \vec{y} \rangle = ||x||^2 + 2\langle \vec{x}, \vec{y} \rangle + ||y||^2$.
6. $\langle \vec{x}, \vec{y} \rangle = \vec{x}^T \vec{y}$.
7. Для любого линейного отображения $f : \mathbb{R}^n \to \mathbb{R}$ можно найти вектор $\vec{a}$, такой что $f(\vec{x}) = \langle \vec{a}, \vec{x} \rangle$
8. $\langle \vec{x}, \vec{y} \rangle = 0$ если и только если вектора $\vec{x}$ и $\vec{y}$ перпендикулярны.
9. Все $\vec{e}_i$ ортогональны друг другу.
10. $\langle \vec{x}, \vec{y} \rangle = \cos\alpha \cdot ||\vec{x}|| \cdot ||\vec{y}||$, где α это угол между векторами $\vec{x}$ и $\vec{y}$

Следующие определения скалярного произведения эквивалентны:

1. $\langle \vec{x}, \vec{y} \rangle = \sum\limits_{i=1}^{n} x_i y_i$ — алгебраическое определение
2. $\langle \vec{x}, \vec{y} \rangle = \cos\alpha \cdot ||\vec{x}|| \cdot ||\vec{y}||$ — геометрическое определение
3. $\langle \vec{x}, \vec{y} \rangle = \vec{x}^T \vec{y}$ — определение через "строка на столбец"
4. $\langle \vec{x}, \vec{y} \rangle = \frac{1}{2}(||\vec{x} + \vec{y}||^2 - ||\vec{x}||^2 - ||\vec{y}||^2)$ — определение через длины

Ортогональные матрицы

Линейное преобразование $Q : \mathbb{R}^n \to \mathbb{R}^n$, сохраняющее длины всех векторов, называется *ортогональным.* Другими словами, Q ортогонально, если $||Q\vec{x}|| = ||\vec{x}||$ для всех $\vec{x} \in \mathbb{R}^n$.

Теорема. Линейное преобразование Q сохраняет длины векторов тогда и только тогда, когда оно сохраняет скалярные произведения. То есть когда $\langle Q\vec{x}, Q\vec{y} \rangle = \langle \vec{x}, \vec{y} \rangle$ для любых $\vec{x}, \vec{y}$.

Любой набор векторов в $\mathbb{R}^n$, удовлетворяющий свойствам

1. в наборе n элементов
2. длина каждого вектора набора равна 1
3. эти векторы попарно ортогональны (то есть любые два различных вектора из набора ортогональны друг другу)

называется *ортонормированным базисом.*

Следующие утверждения равносильны:

1. Q сохраняет длины векторов, то есть $||Q\vec{x}|| = ||\vec{x}||$
2. Q сохраняет скалярные произведения, то есть $\langle Q\vec{x}, Q\vec{y} \rangle = \langle \vec{x}, \vec{y} \rangle$
3. $Q^T Q = E$
4. $QQ^T = E$
5. $Q^T = Q^{-1}$
6. Q переводит базис $\vec{e}_1, \ldots, \vec{e}_n$ в ортонормированный базис

Теорема. Для любого ортогонального преобразования существует ортонормированный базис, в котором матрица преобразования выглядит так:

$$\begin{pmatrix}
1 & & & & & & & & & & & \\
& \ddots & & & & & & & & & & \\
& & 1 & & & & & & & & & \\
& & & -1 & & & & & & & & \\
& & & & \ddots & & & & & & & \\
& & & & & -1 & & & & & & \\
& & & & & & \cos\alpha_1 & -\sin\alpha_1 & & & & \\
& & & & & & \sin\alpha_1 & -\cos\alpha_1 & & & & \\
& & & & & & & & \ddots & & & \\
& & & & & & & & & & \cos\alpha_k & -\sin\alpha_k \\
& & & & & & & & & & \sin\alpha_k & -\cos\alpha_k
\end{pmatrix}$$

где все не написанные коэффициенты матрицы равны 0.

То есть есть в этом базисе ортогональное преобразование

- оставляет неподвижными первые несколько элементов базиса
- следующие несколько элементов базиса заменяет на противоположные
- оставшиеся элементы базиса разбивает на пары и делает поворот в каждой из плоскостей, заданной парой элементов базиса

То есть любое ортогональное преобразование состоит из простых отображений: отражений и поворотов.

Четвёртая неделя. Матричные разложения

Собственные векторы

Собственным вектором преобразования $A : \mathbb{R}^n \to \mathbb{R}^n$ называется ненулевой вектор $\vec{v} \in \mathbb{R}^n$, такой что $A\vec{v} = \lambda\vec{v}$ для какого-нибудь числа $\lambda \in \mathbb{R}$.

Если $\det(A - \lambda E) = 0$, то число λ называется *собственным числом* преобразования A. Собственный вектор затем ищется из уравнения $A\vec{v} = \lambda\vec{v}$.

Свойства собственных векторов.

1. В базисе из собственных векторов матрица имеет диагональный вид
2. Если $\vec{v}$ это собственный вектор A, то и $c\vec{v}$ это собственный вектор A, где c – любое ненулевое число
3. Каждому собственному числу соответствует хотя бы один собственный вектор
4. Пусть $\vec{v}_1, \ldots, \vec{v}_k$ это собственные векторы с различными собственными числами $\lambda_1, \ldots, \lambda_k$. Тогда $\vec{v}_1, \ldots, \vec{v}_k$ линейно независимы.

Спектральное разложение

Утверждение. Выражение $\det(A - \lambda E)$ это многочлен степени n от λ. То есть

$$\det(A - \lambda E) = b_n\lambda^n + b_{n-1}\lambda^{n-1} + \cdots + b_1\lambda + b_0,$$

где $b_n, b_{n-1}, \ldots, b_0$ – какие-то действительные числа, определяемые по коэффициентам матрицы A, $b_n = (-1)^n$.

Пусть у многочлена $\det(A - \lambda E)$ существует n различных корней $\lambda_1, \ldots, \lambda_n$. Найдём соответствующие числам $\lambda_1, \ldots, \lambda_n$ собственные векторы $\vec{v}_1, \ldots, \vec{v}_n$.

В базисе $\vec{v}_1, \ldots, \vec{v}_n$ преобразование A имеет диагональную матрицу:

$$\Lambda := \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}.$$

Пусть V — матрица перехода из базиса $\vec{v}_1, \ldots, \vec{v}_n$ в стандартный базис, то есть её столбцы равны $\vec{v}_1, \ldots, \vec{v}_n$. Значит, выполнено: $A = V \cdot \Lambda \cdot V^{-1}$.

Множество собственных чисел матрицы называется *спектром* матрицы. Построенное выше разложение $A = V \cdot \Lambda \cdot V^{-1}$ называется *спектральным разложением* матрицы A.

Когда действительных корней меньше, чем n, то надо переходить к комплексным числам. Почти все конструкции из курса переносятся с $\mathbb{R}$ на $\mathbb{C}$, кроме:

- Для скалярного произведения, углов и ортогональных преобразований требуется операция *сопряжения*, которая меняет знак у мнимой части: если $z = a + bi$, то сопряжённое число это $\bar{z} := a - bi$.
- Определителем диагональной комплексной матрицы называют произведение элементов на диагонали. А определитель любой недиагональной матрицы вычисляют через метод Гаусса.

Если некоторые корни $\det(A - \lambda E)$ всё же совпадают, то спектрального разложения может не существовать. Обобщение спектрального разложения матрицы это *Жорданово разложение* матрицы, которое существует всегда.

Низкоранговое приближение матрицы

Рекомендательная система

Строки матрицы A размера $n \times m$ отвечают за пользователей, а столбцы — за фильмы. В каждой клетке матрицы A стоит оценка от -2 до 2, либо знак звёздочка $*$, если оценки нет.

Цель — уметь предсказывать, какую оценку пользователь даст фильму, который он ещё не смотрел, чтобы затем решать, какие фильмы ему рекомендовать.

Есть два подхода к решению задачи рекомендаций: Content based filtering (основан на знании контента) и Collaborative filtering (основан только на оценках).

Из векторов-строк $\vec{u}_i$, отвечающих за предпочтения i-ого пользователя, составим матрицу U размера $n \times k$. А из векторов-столбцов $\vec{m}_j$, описывающих j-ый фильм, составим матрицу M размера $k \times m$. Мы пытаемся приблизить A произведением UM, а это то же самое, что приблизить матрицу A матрицей ранга не больше k.

Функция потерь в данном случае равна $L(U, M) = \sum\limits_{a_{ij} \neq *} (a_{ij} - \langle \vec{u}_i, \vec{m}_j \rangle)^2$.

Минимум функции потерь ищется либо градиентным спуском, либо с помощью EM-алгоритма (находим минимум $L(U, M)$ как многочлена второй степени от M при фиксированной U, а затем наоборот фиксируем M и т.д.)

Сингулярное разложение — SVD

У любой матрицы A размера m на n (у которой известны все элементы) существует такое разложение в произведение трёх матриц $A = U\Sigma V^T$, где

- U — ортогональная матрица размера m на m
- V^T — ортогональная матрица размера n на n
- Σ — диагональная матрица размера m на n, при этом диагональные элементы $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq \dots \geq 0$ неотрицательны и упорядочены по убыванию.

Числа σ_i называются *сингулярными числами* матрицы A. Поэтому такое разложение называется *сингулярным разложением*. На английский это переводится как *singular value decomposition*, или коротко *SVD*.

Пусть дана диагональная матрица Σ с диагональными элементами $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$. Сохраним первые k диагональных элементов, а остальные заменим на нули. Полученную матрицу обозначим Σ_k.

$$\Sigma = \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_k & & & \\ & & & \sigma_{k+1} & & \\ & & & & \sigma_{k+2} & \\ & & & & & \ddots \end{pmatrix}, \Sigma_k := \begin{pmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_k & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & \ddots \end{pmatrix}$$

Обозначим $A_k := U\Sigma_k V^T$.

Для двух матриц X и Y одинакового размера *норма Фробениуса* равна $||X - Y||_F := \sqrt{\sum\limits_{i,j} (x_{ij} - y_{ij})^2}$.

Теорема. Дана матрица A. Минимум $||A - B||_F^2$ при ограничении $\operatorname{rank}(B) \leq k$ достигается при $B = A_k$.

Доля *объяснённой дисперсии* при этом равна $\frac{||A_k||_F^2}{||A||_F^2} = \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{l} \sigma_i^2}$.

Другие матричные разложения

Главная диагональ матрицы A размера m на n — это все элементы a_{ii}, где $i \leq \min(n, m)$.

Матрица, у которой все элементы ниже главной диагонали равны нулю, называется *верхнетреугольной*:

$$\begin{pmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{pmatrix}$$

Аналогично, матрица, у которой все элементы выше главной диагонали равны нулю, называется *нижнетреугольной*.

LU-разложение

Если A это квадратная обратимая матрица (с некоторым незначительным условием), то существует разложение $A = LU$, где

- L это нижнетреугольная матрица (L от "Lower triangular")
- U это верхнетреугольная матрица (U от "Upper triangular")

- размеры L и U совпадают с размером A

Иногда используют модификации LU, называемые LDP и LUP.

QR-разложение

Если A имеет размер m на n, где $m \geq n$, то существует разложение $A = QR$, где

- Q это ортогональная матрица размера m на m
- R это верхнетреугольная матрица размера m на n

Заметьте, что из-за верхнетреугольности последние $(m-n)$ строк матрицы R всегда оказываются заполнены нулями:

$$\begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Пятая неделя. Backpropagation

Одномерный backpropagation

Backpropagation — это метод вычисления градиентов функции потерь для нейросети путем обратного распространения ошибки.

Рассмотрим одномерную нейросеть в общем случае. Пусть она состоит из k линейных слоев с одним нейроном (без нейрона сдвига), веса этих слоев — $w_1, w_2, \dots, w_k$, функции активации $f_1, f_2, \dots, f_k$.

$$x \xrightarrow{w_1, f_1} h_1 \xrightarrow{w_2, f_2} h_2 \xrightarrow{w_3, f_3} \dots \xrightarrow{w_{k-1}, f_{k-1}} h_{k-1} \xrightarrow{w_k, f_k} \hat{y}$$

Вход обозначим за x, правильный ответ — за y, функцию потерь — за L. Обозначим также выходы линейных слоев с активацией:

$$h_1 = f_1(w_1 x), h_2 = f_2(w_2 h_1), \dots, h_{k-1} = f_{k-1}(w_{k-1} h_{k-2}), \hat{y} = f_k(w_k h_{k-1})$$

Частная производная по слою номер i равна

$$L(y, \hat{y})'_{w_i} = L(y, \hat{y})'_{\hat{y}} \cdot f'_k(w_k h_{k-1}) \cdot w_k \cdot f'_{k-1}(w_{k-1} h_{k-2}) \cdot w_{k-1} \cdot \dots \cdot f'_i(w_i h_{i-1}) h_{i-1}$$

Верны более компактные формулы:

$$L(y, \hat{y})'_{w_i} = L(y, \hat{y})'_{h_i} h_{i-1}$$

Чтобы эта формула работала и для $i = 1$, будем считать, что $h_0 = x$.

Матричное дифференцирование

Если функция отображает матрицу в число $f : \mathbb{R}^{n \times m} \to \mathbb{R}$, то будем записывать её градиент в матрицу:

$$\nabla_A f(A) = \begin{pmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1m}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \cdots & \frac{\partial f}{\partial a_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \frac{\partial f}{\partial a_{n2}} & \cdots & \frac{\partial f}{\partial a_{nm}} \end{pmatrix}$$

Верны следующие формулы:

- $\nabla_x x^T a = a$, или, что то же самое, $\nabla_x a^T x = a$
- $\nabla_x x^T A x = (A + A^T) x$
- $\nabla_A x^T A y = x y^T$

Точное решение для линейной регрессии с MSE

Пусть X — матрица, в которой по строкам записаны признаки объектов, y — вектор ответов для этих объектов, $w = (w_1, w_2, \dots, w_n)$ — веса линейной регрессии, $\hat{y} = Xw$ — вектор ответов модели для обучающей выборки. Тогда минимум среднеквадратичной функции потерь $L(y, \hat{y}) = \frac{1}{m}(y - \hat{y})^T (y - \hat{y})$ достигается при $w = (X^T X)^{-1} X^T y$ (если матрица $X^T X$ имеет полный ранг, в частности, если матрица X имеет ранг n).

Непрерывная всюду определенная функция $f : \mathbb{R}^n \to \mathbb{R}$ является *выпуклой*, если для любой пары точек $x, y \in \mathbb{R}^n$ выполнено $f(\frac{x}{2} + \frac{y}{2}) \leq f(\frac{x}{2}) + f(\frac{y}{2})$.

Среднеквадратичная функция потерь выпукла, поэтому у неё существует глобальный минимум.

Backpropagation для многомерного случая

Рассмотрим нейронную сеть с одним линейным слоем и функцией активации $f : \mathbb{R} \to \mathbb{R}$. Веса линейного слоя задаются матрицей $W \in \mathbb{R}^{n \times m}$, m — размер входа с учетом нейрона сдвига, n — размер выхода.

По функции f построим функцию $F : \mathbb{R}^n \to \mathbb{R}^n$, которая будет отвечать покоординатному применению функции f к вектору. То есть для $v \in \mathbb{R}^n$ выполнено $F(v) = (f(v_1), f(v_2), \ldots, f(v_n))$.

Для входного вектора $x \in \mathbb{R}^m$ и весов W значение функции потерь вычисляется как $L(y, \hat{y}) = L(y, F(Wx))$.

Утверждение. Пусть даны функции $f : \mathbb{R}^n \to \mathbb{R}^m$ и $g : \mathbb{R}^m \to \mathbb{R}$. Для $x \in \mathbb{R}^n$ пусть $y = f(x)$ и $z = g(y)$. Функцию f можно представлять себе, как набор функций $f_1, f_2, \ldots, f_m$ таких, что $f(x) = (y_1, y_2, \ldots, y_m) = (f_1(x), f_2(x), \ldots, f_m(x))$.

Градиент $\nabla_x z$ равен $J(y, x)^T \nabla_y z$, где $J_x(y)$ — матрица из частных производных (её ещё называют *матрицей Якоби*):

$$J_x(y) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}.$$

Правила дифференцирования многомерной сложной функции позволяет записать соотношения:

$$\frac{\partial L}{\partial w_i} = J_{w_i}(h_i)^T \frac{\partial L}{\partial h_i}$$

$$\frac{\partial L}{\partial x} = J_x(h_i)^T \frac{\partial L}{\partial h_i}$$

Комплексные числа

Комплексное число z — это упорядоченная пара действительных чисел (a, b). Запись: $z = a + bi$.

Число a называют *действительной частью* числа $a + bi$ и обозначают $\mathrm{Re}(z)$. Число b называют *мнимой частью* числа $a + bi$ и обозначают $\mathrm{Im}(z)$.

Множество всех комплексных чисел обозначают буквой $\mathbb{C}$.

Складываются и вычитаются комплексные числа покоординатно: $(a + bi) + (c + di) := (a + c) + (b + d)i$

Умножение комплексных чисел определяется так: $(a + bi)(c + di) := (ac - bd) + (ad + bc)i$

Комплексное число $z = a + bi$ можно представить как вектор из действительных чисел $(a, b) \in \mathbb{R}^2$. Длина вектора z называется *модулем* числа z и обозначается $|z|$. Угол между вектором z и осью OX называется *аргументом* числа z и обозначается $\arg(z)$. У числа 0 длина определена и равна 0, но аргумент у него не определён.

Теорема. При умножении комплексных чисел модули перемножаются, а аргументы складываются. То есть для любых не равных нулю $z_1, z_2 \in \mathbb{C}$ выполнено:

1. $|z_1 z_2| = |z_1| \cdot |z_2|$
2. $\arg(z_1 z_2) = \arg(z_1) + \arg(z_2)$

Многочлен — это функция вида: $P(x) := a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$.

Корнем многочлена P называется число, при подстановке которого вместо x в выражение $P(x)$ получается 0. Другими словами, число c называется коренем многочлена P, если и только если $P(c) = 0$.

Утверждение. $\mathbb{C}$ алгебраически замкнуто, то есть любой многочлен с комплексными коэффициентами имеет корень в $\mathbb{C}$.