Линейная алгебра. Векторы

1. Знакомство с линейной алгеброй

Ключевыми понятиями линейной алгебры являются «*вектор*» и «*матрица*». Для понимания того, каким образом они появляются в анализе данных, можно рассмотреть несколько примеров.

Имеется сеть магазинов, требуется предсказать прибыль для каждого из них в следующем месяце. В данной задаче магазины являются **объектами**, которые необходимо описать с помощью числовых характеристик — **признаков**.

В качестве признаков можно рассмотреть, к примеру, прибыль магазина в предыдущие четыре месяца: 50, 47, 52, 55. Будущая прибыль магазина также зависит от планируемого числа акций для трех основных категорий: кондитерские изделия, овощи и фрукты, мясо. Это следующие три признака: 5, 1, 2. Прибыль зависит и от географического положения магазина, следовательно, его координаты также являются признаками: 55,73, 37,59. Последним рассматриваемым признаком будет количество праздничных дней в следующем месяце: 1.

Все описанные числа в совокупности образуют набор, называемый *вектором*:

$$\vec{x} = (50,\ 47,\ 52,\ 55,\ 5,\ 1,\ 2,\ 55{,}73,\ 37{,}59,\ 1).$$

Другим примером является задача определения для каждой конкретной бутылки вина сорта винограда, из которого оно сделано. Это может пригодиться при определении качества вин в зависимости от дороговизны сорта винограда. Здесь признаками являются результаты химических анализов (содержание алкоголя, щелочность, насыщенность цвета). Эти признаки также образуют вектор, на основе которого можно будет делать прогнозы:

$$\vec{x} = (14{,}23,\ 1{,}17,\ 3{,}43,\ 5{,}64,\ 3{,}92).$$

Каждый человек характеризуется своим геномом. Именно геном определяет предрасположенность к некоторым наследственным заболеваниям. Можно поставить следующую задачу: предсказать по геному человека, заболеет ли он раком в течение пяти лет. Это зависит от наличия мутаций в геноме. В связи с этим можно описать каждого конкретного человека вектором, длина которого будет равна количеству возможных мутаций. Данный вектор будет состоять из нулей и единиц, причем единицы будут стоять на позициях, соответствующих произошедшим мутациям:

$$\vec{x} = (0,\ 1,\ 0,\ 0,\ 1, \dots).$$

Если рассматривается несколько магазинов или несколько бутылок вина, то данные обретают двумерную структуру. Их можно составить в таблицу, в которой строки будут соответствовать объектам, а столбцы — признакам. Такая структура называется **матрицей**.

2. Векторные пространства

Понятие «вектор» формализуется с помощью *векторных пространств* — множеств, обладающих определенными свойствами. Тогда **вектор** — это элемент векторного пространства.

Векторное пространство — это множество V, на котором заданы две операции: *сумма векторов* и *умножение вектора на число*. Эти операции *замкнуты* — и сумма любых двух векторов, и произведение любого вектора на любое число являются векторами.

Сложение и умножение на число должны удовлетворять следующим *аксиомам*:

1) $x + y = y + x \ \forall x, y \in V$ (коммутативность);
2) $x + (y + z) = (x + y) + z \ \forall x, y, z \in V$;
3) $\exists 0 \in V: \ x + 0 = x \ \forall x \in V$;
4) $\forall x \in V \ \exists -x \in V: \ x + (-x) = 0$;
5) $\alpha(\beta x) = (\alpha\beta)x$;
6) $1 \cdot x = x$;
7) $(\alpha + \beta)x = \alpha x + \beta x$;
8) $\alpha(x + y) = \alpha x + \alpha y$ (дистрибутивность).

Наиболее простые векторные пространства состоят из векторов, каждый элемент которых является вещественным числом. Такие пространства называются **евклидовы** и обозначаются $\mathbb{R}^n$, где n — это число элментов в каждом векторе данного пространства. Например, $\mathbb{R}^2$ — это все точки на плоскости, а $\mathbb{R}^3$ — все точки в пространстве.

Операции в евклидовом пространстве вводятся поэлементно. Поэлементное сложение выглядит следующим образом:

$$\vec{a} = \big(a_1, \ldots, a_n\big), \quad \vec{b} = \big(b_1, \ldots, b_n\big) \quad \Rightarrow \quad \vec{a} + \vec{b} = \big(a_1 + b_1, \ldots, a_n + b_n\big).$$

Поэлементное умножение на число задается аналогично:

$$\vec{a} = \big(a_1, \ldots, a_n\big), \quad \beta \in \mathbb{R} \quad \Rightarrow \quad \beta\vec{a} = \big(\beta a_1, \ldots, \beta a_n\big).$$

В примере с определением сорта винограда векторным пространством являются все возможные признаковые описания бутылок с вином. При этом конкретный вектор — это описание конкретной бутылки. Векторы можно усреднять, например, посчитать среднее признаковое описание для конкретного сорта:

$$\bar{\vec{x}} = \frac{1}{m}\big(x_1, \ldots, x_n\big).$$

В этой задаче можно провести некоторый анализ данных, например, искать сорт с наибольшим содержанием алкоголя.

3. Линейная независимость

Рассматриваются сто магазинов, каждый из которых описывается десятью признаками. Можно составить таблицу 100×10, в которой каждая строка будет соответствовать признаковому описанию одного магазина, а каждый столбец — набору значений одного признака для всех магазинов. В качестве вектора в данном случае будут рассматриваться *столбцы* такой таблицы.

Если, к примеру, $x_1 = 1000x_2$, то это может означать, что первый признак является весом товара в граммах, а второй — в килограммах. Это избыточная информация, так как достаточно знать вес в одной из единиц измерения. Значит, можно оставить лишь один вектор из этих двух.

Если также $x_3 = 0{,}5x_4 + 0{,}5x_5$, то это может означать, что четвертый и пятый признаки являются прибылью магазина в прошлом и позапрошлом месяцах, а третий является средней прибылью за последние два месяца. Здесь снова имеется избыточная информация.

Описанная ситуация носят название *линейная зависимость* (один вектор равен сумме других векторов с коэффициентами). Линейная зависимость ведет к лишним затратам на хранение данных и вредит некоторым методам машинного обучения.

Определение 1. Пусть дан набор векторов $\vec{x}_1, \ldots, \vec{x}_n$. Они являются **линейно зависимыми**, если существуют такие числа $\beta_1, \ldots, \beta_n$. хотя бы одно из которых не равно нулю, что сумма векторов с такими коэффициентами равна нулю:

$$\beta_1\vec{x}_1 + \cdots + \beta_n\vec{x}_n = \vec{0}.$$

Возможно и альтернативное определение линейной зависимости. Векторы **линейно зависимы**, если один из них линейно выражается через остальные:

$$\vec{x}_i = \beta_1\vec{x}_1 + \ldots \beta_{i-1}\vec{x}_{i-1} + \beta_{i+1}\vec{x}_{i+1} + \cdots + \beta_n\vec{x}_n.$$

Данное определение аналогично определению **??**.

Необходимо отметить, что если один из векторов нулевой, то все они являются линейно зависимыми: берется коэффициент при нулевом векторе равным 1, при остальных — равным 0, и определение будет выполнено. Это вполне логично, так как нулевой вектор не содержит никакой информации, а вся система становится избыточной.

Определение 2. Пусть дан набор векторов $\vec{x}_1, \ldots, \vec{x}_n$. Они являются **линейно независимыми**, если уравнение:

$$\beta_1\vec{x}_1 + \cdots + \beta_n\vec{x}_n = \vec{0}$$

выполнено только при:

$$\beta_1 = \beta_2 = \cdots = \beta_n = 0.$$

Пусть даны три вектора из евклидового пространства $\mathbb{R}^3$:

$$\vec{x}_1 = (2,\ 1,\ 2), \quad \vec{x}_2 = (4,\ 2,\ 4), \quad \vec{x}_3 = (1,\ 2,\ 3).$$

Можно заметить, что:

$$2\vec{x}_1 - \vec{x}_2 + 0\vec{x}_3 = \vec{0},$$

следовательно, система этих векторов являются линейно зависимой.

Если же заменить первую координату второго вектора:

$$\vec{x}_1 = (2,\ 1,\ 2), \quad \vec{x}_2 = (5,\ 2,\ 4), \quad \vec{x}_3 = (1,\ 2,\ 3),$$

то система становится линейно независимой. Есть два варианта это понять. Первый вариант — *метод Гаусса*, позволяющий привести систему к такому виду, что линейная независимость становится очевидной. Второй вариант является более простым и заключается в *вычислении ранга*, которое будет изучено далее.

Определение 3. Пусть дано некоторое векторное пространство V. Максимальное число линейно независимых векторов в нем называется **размерностью** и обозначается $\dim V$.

В евклидовом пространстве $\mathbb{R}^n$ размерностью является n. Набор векторов, через который выражается любой другой вектор пространства, называется **базисом** векторного пространства:

$$\vec{e}_1 = (1, 0, \ldots, 0), \quad \vec{e}_2 = (0, 1, \ldots, 0), \quad \ldots, \quad \vec{e}_n = (0, 0, \ldots, 1).$$

4. Операции в векторных пространствах

Рассматривается вектор в двумерном пространстве, выходящий из нуля и идущий в точку $(1, 2)$ (см. рис. 1). *Длина* данного вектора измеряется следующим образом:

$$\sqrt{1^2 + 2^2} = \sqrt{5}.$$

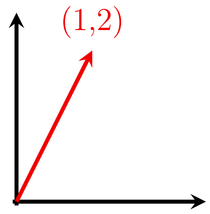

Рис. 1.

К данному вектору добавляется еще один, выходящий из нуля и идущий в точку $(-2, 1)$ (см. рис. 2). *Угол* между этими векторами равен $90°$.

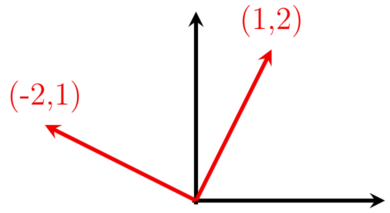

Рис. 2.

Расстояние между точками можно найти следующим образом:

$$\sqrt{\big(1 - (-2)\big)^2 + (2 - 1)^2} = \sqrt{10}.$$

Необходимо обобщить эти понятия на многомерное евклидово пространство. Длина вектора обобщается с помощью **нормы** — функции от вектора ($||\vec{x}||$). Норма должна удовлетворять трем требованиям:

1) Если $||\vec{x}|| = 0$, то $\vec{x} = 0$;
2) $||\vec{x} + \vec{y}|| \leqslant ||\vec{x}|| + ||\vec{y}||$ (неравенство треугольника);
3) $||\alpha\vec{x}|| = |\alpha| \cdot ||\vec{x}||$.

Векторное пространство, в котором введено понятие «норма», называется *нормированным.* Самым распространенным примером нормы является *евклидова норма*:

$$||\vec{x}||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Другим примером является *манхэттенская норма*:

$$||\vec{x}||_1 = \sum_{i=1}^{n} |x_i|.$$

Расстояние обобщается с помощью **метрики**:

$$\rho(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||.$$

Метрика соответствует геометрическим представлениям (см. рис. 3).

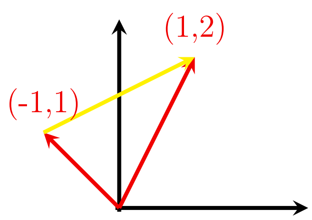

Рис. 3.

Евклидова метрика задается следующим образом:

$$\rho(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Манхэттенская метрика имеет вид:

$$\rho(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

Пусть имеются два вектора: один с координатами $(-1, 1)$ и второй с координатами $(1, 2)$. Между ними можно найти расстояние, можно найти их нормы и посчитать угол между ними. Затем оба вектора умножаются на 2. Получены векторы с координатами $(-2, 2)$ и $(2, 4)$. Изменятся их нормы и расстояние между ними, а угол останется тем же.

Таким образом, необходимо обобщить и понятие «угол», так как нормы и расстояния оказалось недостаточно.

Скалярное произведение двух векторов выглядит следующим образом:

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{n} x_i y_i.$$

Через скалярное произведение можно вычислить норму и расстояние:

$$||\vec{x}|| = \sqrt{\langle \vec{x}, \vec{x} \rangle}, \quad \rho\big(\vec{x}, \vec{y}\big) = ||\vec{x} - \vec{y}|| = \sqrt{\langle \vec{x} - \vec{y}, \vec{x} - \vec{y} \rangle}.$$

В евклидовых пространствах скалярное произведение можно вычислить с помощью косинуса угла между этими векторами:

$$\langle \vec{x}, \vec{y} \rangle = ||\vec{x}|| \cdot ||\vec{y}|| \cos\big(\vec{x}, \vec{y}\big).$$

Отсюда следует, что:

$$\cos\big(\vec{x}, \vec{y}\big) = \frac{\langle \vec{x}, \vec{y} \rangle}{||\vec{x}|| \cdot ||\vec{y}||}.$$

Косинус угла, по сути, является мерой сонаправленности векторов, так как для параллельных векторов:

$$\cos\big(\vec{x}, \vec{y}\big) = 1,$$

а для перпендикулярных векторов:

$$\cos\big(\vec{x}, \vec{y}\big) = 0.$$