Знакомство с линейной алгеброй

1. Признаковое описание

В машинном обучении широко используется понятие признака. Признаком называется отображение из множества объектов в множество допустимых значений этого признака. Если задано множество объектов и некоторый набор признаков, для каждого объекта можно построить его признаковое описание — вектор, составленный из значений этого набора признаков на данном объекте.

Пусть, например, для каждого магазина торговой сети требуется предсказать прибыль в следующем месяце. Эта задача анализа данных имеет огромную практическую ценность. Действительно, если будет выяснено, что прибыль некоторого магазина упадет, можно будет заблаговременно принять меры по предотвращению этого. В этой задаче множеством объектов является множество магазинов. В качестве признаков разумно выбрать следующие:

- Прибыль магазина в каждом из 4-ех последних месяцев (4 признака)
- Планируемое число акций для каждой из трех основных категорий (3 признака)
- Географические координаты магазина: широта и долгота (2 признака)
- Число дней, когда магазин будет открыт в ближайшем месяце. (1 признак)

Признаковым описанием магазина будет вектор, в котором находятся значения данных 10 признаков для данного конкретного магазина.

Если необходимо работать с признаковыми описаниями сразу нескольких объектов, удобно ввести двумерную структуру данных — матрицу, в которой каждая строка соответствует одному объекту, а каждый столбец — признаку. Работать с векторами и матрицами позволяет аппарат линейной алгебры.

2. Векторное пространство

Векторное пространство V представляет собой набор элементов, называемых векторами, для которых определены операции сложения друг с другом и умножения на скаляр, причем эти операции замкнуты и подчинены восьми аксиомам:

1. Коммутативность сложения
2. Ассоциативность сложения
3. Существование нейтрального элемента относительно сложения
4. Существование для каждого вектора x противоположенного вектора $-x$
5. Ассоциативность умножения на скаляр
6. Унитарность: умножение на единичный скаляр не меняет вектор
7. Дистрибутивность умножения на вектор относительно сложения скаляров
8. Дистрибутивность умножения на скаляр относительно сложения векторов

3. Линейная независимость

Линейная зависимость является одним из основополагающих понятий линейной алгебры. Конечный набор элементов векторного пространства называется линейно зависимым, если существует нетривиальная линейная комбинация элементов из этого набора, равная нулевому элементу. Линейная комбинация называется тривиальной, если все коэффициенты в ней равны нулю.

Оказывается, что конечный набор элементов векторного пространства линейно зависим тогда и только тогда, когда один из элементов этого набора может быть выражен через оставшиеся.

С помощью понятия линейной независимости вводится понятие размерности векторного пространства. А именно: размерностью $dimV$ векторного пространства V называется максимальное число линейно независимых векторов в нем.

Пусть дан некоторый набор объектов X и набор $f_j : X \to \mathbb{R}$ вещественнозначных признаков. В этом случае набор векторов признаков может быть линейно зависим. Например, признаки вес товара на первом складе, вес товара на втором складе и вес товара на обоих складах линейно зависимы. Другой случай — два вещественнозначных признака отличаются множителем, например являются одной и той же величиной в разных единицах измерения. В обоих случаях наблюдается избыточность информации.

Такая избыточность приводит к дополнительным затратам ресурсов. Более того, линейная зависимость векторов признаков приводит к возникновению проблем при обучении линейной регрессионной модели — об этом пойдет речь в следующем курсе.

Чтобы проверить, являются ли система векторов линейно зависимой, можно составить из них матрицу и вычислить ее ранг. Об этом будет рассказано далее.

4. Норма и скалярное произведение векторов

Нормированные пространства

Для обобщения понятия длины вектора используется понятие нормы. Функция $\|\cdot\| : V \to \mathbb{R}$ называется нормой в векторном пространстве V, если для нее выполняются аксиомы нормы:

1. $\|x\| = 0 \iff x = 0$ (Нулевую норму имеет только нулевой вектор)
2. $\forall x, y \in L : \|x + y\| \leq \|x\| + \|y\|$ (Неравенство треугольника)
3. $\forall \alpha \in \mathbb{R}\ \forall x \in V : \|\alpha x\| = |\alpha|\|x\|$ (Условие однородности)

Пространство с введенной на нем нормой называют нормированным пространством. Обычно используется Евклидова норма $\|x\|_2$, другой пример нормы — Манхэттенская норма $\|x\|_1$:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}, \qquad \|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

Метрические пространства

Понятие расстояние обобщается с помощью понятия метрики. Пусть X — некоторое множество, а числовая функция $d : X \times X \to \mathbb{R}$, которая называется метрикой, удовлетворяет следующим условиям:

1. $d(x,\ y) = 0 \Leftrightarrow x = y$ (аксиома тождества).
2. $d(x,\ y) = d(y,\ x)$ (аксиома симметрии).
3. $d(x,\ z) \leqslant d(x,\ y) + d(y,\ z)$ (неравенство треугольника).

Любое нормированное пространство можно превратить в метрическое, определив функцию расстояния $d(x,\ y) = \|y - x\|$. Например, Евклидова и Манхэттенская метрики имеют вид:

$$\rho_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \qquad \rho_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Название «манхэттенская метрика» связано с уличной планировкой Манхэттена.

Скалярное произведение

Скалярным произведением называется функция $\langle x, y\rangle : V \times V \to \mathbb{R}$, удовлетворяющая следующим условиям:

1. $\langle \alpha x_1 + \beta x_2, y\rangle = \alpha\langle x_1, y\rangle + \beta\langle x_2, y\rangle; \quad \forall \alpha, \beta \in \mathbb{R}\ \forall x_1, x_2, y \in V$
2. $\langle y, x\rangle = \langle x, y\rangle; \quad \forall x, y \in V$
3. $\langle x, x\rangle > 0; \quad \forall x \in V \setminus \{0\}; \quad \langle 0, 0\rangle = 0;$

В современном аксиоматическом подходе уже на основе понятия скалярного произведения векторов вводятся следующие производные понятия:

1. Длина вектора $\|x\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$
2. Угол между векторами $\alpha = \arccos \frac{\langle \mathbf{a},\mathbf{b}\rangle}{\sqrt{\langle \mathbf{a},\mathbf{a}\rangle \langle \mathbf{b},\mathbf{b}\rangle}}$.

В случае Евклидового пространства скалярное произведение задается формулой $\langle \mathbf{a}, \mathbf{b}\rangle = \sum_{i=1}^{n} a_i b_i$ и можно убедиться, что такое определение согласуется с введенной ранее Евклидовой нормой:

$$\|x\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2} = \|x\|_2$$