Понимаем математику: матрицы

Определения

Матрица — прямоугольная таблица с числами (обычно вещественными или комплексными).

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nm} \end{pmatrix}$$

Операции с матрицами:

- Сложение
- Умножение на константу
- Умножение на матрицу
- Транспонирование

Утверждение 1. *Матрицы размера $n \times n$ образуют ассоциативное кольцо с единицей относительно матричного сложения и умножения.*

Определение. Определитель матрицы A размера $n \times n$ определяется рекурсивно:

$$\det A = \sum_{k=1}^{n} (-1)^{k-1} \cdot a_{1k} \cdot \det A_{1k},$$

где A_{1k} — матрица размера $(n-1) \times (n-1)$, полученная из A удалением первой строки и k-го столбца.

Свойства определителя матрицы:

- $\det(AB) = \det A \cdot \det B$;
- $\det(A^{-1}) = \frac{1}{\det A}$;
- $\det A = \prod_{i=1}^{n} \lambda_i$ — произведение всех n собственных чисел матрицы.

Утверждение 2. *Матрица A размера $n \times n$ обратима тогда и только тогда, когда* $\det A \neq 0$.

Алгоритмы вычисления обратной матрицы за $O(n^3)$:

- Метод Гаусса.
- Разложение Холецкого (для матриц ковариации).
- SVD-декомпозиция
- ...

Лучший (в смысле асимптотики) известный алгоритм обращения матрицы работает за $O(n^{2.3728639})$.

Собственные числа и собственные векторы матрицы

Определение. Собственные числа и собственные векторы матрицы A — это такие λ_i и $\vec{v}_i$, что

$$A\vec{v}_i = \lambda_i \vec{v}_i.$$

Свойства:

- собственные числа — корни *характеристического многочлена* $\chi(\lambda) = \det(A - \lambda I)$, существует не более n комплексных λ_i;
- $\mathrm{eig}(AB) = \mathrm{eig}(BA)$ (наборы собственных чисел совпадают);
- если $\mathrm{rk}A = r$, то у матрицы A не более r ненулевых собственных чисел;
- если $A = A^T$, то матрица V, составленная из собственных векторов, ортогональна, а все λ_i вещественные;
- сумма диагональных элементов (след матрицы) равна $\mathrm{tr}A = \sum \lambda_i$.

Матричные разложения

LU-разложение (существует, когда все главные миноры матрицы A невырождены):

$$A = LU,$$

где L — нижнетреугольная матрица, U — верхнетреугольная матрица.

Разложение Холецкого (существует для симметричных положительно определённых матриц):

$$A = U^T U = LL^T,$$

где U — верхнетреугольная матрица, которая определяется единственным образом.

Спектральное разложение (существует, когда у матрицы есть n линейно независимых собственных векторов):

$$A = V \Lambda V^{-1},$$

где V — матрица, составленная из собственных векторов, $\Lambda = \mathrm{diag}(\mathrm{eig}A)$.

SVD-декомпозиция

Теорема 1. *Любая матрица A размера $n \times m$ может быть представлена в виде*

$$A = UDV^T,$$

где

$$\begin{aligned} U &= \text{ собственные векторы матрицы } AA^T && (n \times n) \\ D &= \sqrt{\operatorname{diag}\left(\operatorname{eig}\left(AA^T\right)\right)} && (n \times m) \\ V &= \text{ собственные векторы матрицы } A^T A && (m \times m) \end{aligned}$$

Матрицы U, V являются ортогональными, D — диагональная матрица.
Применения SVD:

- Наилучшее низкоранговое приближение матрицы;
- Задача уменьшения размерности (PCA);
- ...

Псевдообратная матрица

Матрица A^+ размера $m \times n$ называется псевдообратной матрицей для матрицы A размера $n \times m$, если она удовлетворяет следующим критериям:

1. $AA^+A = A$;
2. $A^+AA^+ = A^+$;
3. AA^+ — симметричная матрица;
4. A^+A — симметричная матрица.

Альтернативное определение:

$$A^+ = \lim_{\delta \to +0} \left(A^T A + \delta I\right)^{-1} A^T = \lim_{\delta \to +0} A^T \left(AA^T + \delta I\right)^{-1}.$$

Псевдообратная матрица существует и единственна. Если матрица A обратима, то $A^+ = A^{-1}$.

Псевдообратную матрицу можно вычислить через SVD-разложение $A = UDV^T$:

$$A^+ = VD^{-1}U^T.$$

Матричное дифференцирование

Рассмотрим производные (градиенты) функций вида

- $f : \mathbb{R} \to \mathbb{R}$ (скаляр);
- $f : \mathbb{R}^n \to \mathbb{R}$ (вектор);
- $f : \mathbb{R} \to \mathbb{R}^n$ (вектор);
- $f : \mathbb{R} \to \mathbb{R}^{n\times n}$ (матрица);
- $f : \mathbb{R}^{n\times n} \to \mathbb{R}$ (матрица);
- $f : \mathbb{R}^n \to \mathbb{R}^n$ (матрица);

Напоминание. Дифференциал функции $f(x)$ в точке x_0 — линейная функция, такая, что

$$d_{x_0} f(h) = f(x_0 + h) - f(x_0) + o(h).$$

Функция $f(x)$ называется дифференцируемой, если существует дифференциал $df(x)$.

Связь дифференциала и производной (градиента):

$$d_{x_0} f(h) = [\nabla f(x_0)]^T h \quad \text{или пишут проще:} \quad df(x) = [\nabla f(x)]^T dx.$$

Связь матричного дифференциала и градиента:

$$df(X) = \mathrm{Tr}([\nabla f(X)]^T dX).$$

Правила преобразования:

$$\begin{aligned}
dA &= 0 \\
d(\alpha X) &= \alpha(dX) \\
d(AXB) &= A(dX)B \\
d(X+Y) &= dX + dY \\
d\left(X^T\right) &= (dX)^T \\
d(XY) &= (dX)Y + X(dY) \\
d\left(\frac{X}{\phi}\right) &= \frac{\phi dX - (d\phi)X}{\phi^2}
\end{aligned}$$

Несколько стандартных дифференциалов:

$$\begin{aligned}
d\left(c^T x\right) &= c^T dx \\
d\left(x^T A x\right) &= x^T\left(A + A^T\right) dx \\
d\left(x^T A x\right) &= 2x^T A dx \quad \left(\text{если } A = A^T\right) \\
d(\mathrm{Tr}(X)) &= \mathrm{Tr}(dX) \\
d(\det(X)) &= \det(X)\,\mathrm{Tr}\left(X^{-1} dX\right) \\
d\left(X^{-1}\right) &= -X^{-1}(dX)X^{-1}
\end{aligned}$$

Одним из самых важных является правило производной композиции. Пусть $g(Y)$ и $f(X)$ — две дифференцирумые функции, и мы знаем выражения для их дифференциалов: $dg(Y)$ и $df(X)$. Чтобы посчитать производную композиции $\phi(X) := g(f(X))$, как и в скалярном случае, нужно:

- взять выражение посчитанного дифференциала $dg(Y)$;
- подставить в него вместо Y значение $f(X)$, а вместо dY значение $df(X)$.

Пример 1. Рассмотрим функцию $\phi(x) := \ln\left(x^T A x\right)$, где $A \in \mathbb{S}^n_{++}$. В данном случае

$$g(y) := \ln(y), \quad dg(y) = \frac{dy}{y}; \quad f(x) := x^T A x, \quad df(x) = 2x^T A dx.$$

Подставляем формально в $dg(y)$ вместо y выражение для $f(x) = x^T A x$, а вместо dy выражение для $df(x) = 2x^T A dx$, получаем

$$d\phi(x) = \frac{2x^T A dx}{x^T A x}.$$

Пример 2. Найти дифференциал $df(X)$, а также градиент $\nabla f(X)$ функции

$$f(X) := \ln(\det(X)),$$

заданной на множестве $\mathbb{S}^n_{++}$.

Решение. Найдём дифференциал:

$$df(X) = d(\ln\det(X)) = \left\{d(\ln(x)) = \frac{dx}{x}\right\} = \frac{d(\mathrm{Det}(X))}{\det(X)} = \frac{\det(X)\,\mathrm{Tr}\left(X^{-1} dX\right)}{\det(X)} = \mathrm{Tr}\left(X^{-1} dX\right).$$

Заметим, что $df(X)$ записан в канонической форме $df(X) = \mathrm{Tr}(\nabla f(X) dX)$, поэтому, $\nabla f(X) = X^{-1}$.

Решение задачи линейной регрессии

Дана матрица X размера $n \times d$ (n объектов, у каждого d числовых признаков) и вектор y размера n (целевые метки). Необходимо найти наилучшее линейное приближение y как функции от X (в смысле L_2-нормы):

$$\|X\theta - y\|^2 \to \min_\theta .$$

Приравняем градиент функции к нулю:

$$\begin{aligned} \frac{\partial \|X\theta - y\|^2}{\partial \theta} = \frac{\partial (X\theta - y)^T (X\theta - y)}{\partial \theta} = \\ \frac{\partial \theta^T X^T X \theta}{\partial \theta} - \frac{\partial \theta^T X^T y}{\partial \theta} - \frac{\partial y^T X \theta}{\partial \theta} + \frac{\partial y^T y}{\partial \theta} = \\ 2X^T X\theta - 2X^T y = 0. \end{aligned}$$

Находим решение $\hat{\theta} = (X^T X)^{-1} X^T y$.

Задача. Найдите решение задачи Ridge-регрессии:

$$\|X\theta - y\|^2 + \lambda \|\theta\|^2 \to \min_\theta .$$