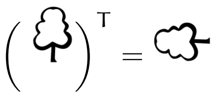
Тятя! Тятя! Нейросети заменили продавца!

Пилиф Ульянкин

https://github.com/FUlyankin/neural_nets_prob

Листочек 4: матричное дифФфФфФференцирование



«Джек и бобовый стебель» (1890)

Упражнение 1

Найдите следующие производные:

а. $f(x) = x^2$, где x скаляр

б. $f(x) = a^T x$, где a и x векторы размера $1 \times n$

в. $f(x) = x^T A x$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

г. $f(x) = \ln(x^T A x)$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

д. $f(x) = a^T X A X a$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

е. $f(x) = x x^T x$, где x вектор размера $1 \times n$

Решение:

Решение этих задач ищи в конспекте семинара: https://github.com/FUlyankin/deep_learning_tf/blob/main/week03_matrix_diff/sem03-vector-diff.pdf

Упражнение 2

Давайте пополним таблицу дифференциалов несколькими новыми функциями, специфичными для матриц. Найдём матричные дифференциалы функций:

а. $f(X) = X^{-1}$, где матрица X размера $n \times n$

б. $f(X) = \det X$, где матрица X размера $n \times n$

в. $f(X) = \operatorname{tr}(X)$, где матрица X размера $n \times n$

г. Ещё больше матричных производных можно найти в книге The Matrix Cookbook[1]

Решение:

Решение этих задач ищи в конспекте семинара: https://github.com/FUlyankin/deep_learning_tf/blob/main/week03_matrix_diff/sem03-vector-diff.pdf

Упражнение 3

Рассмотрим задачу линейной регресии

$$L(w) = (y - Xw)^T (y - Xw) \to \min_w.$$

а. Найдите $L(w)$, выведите формулу для оптимального w.

б. Как выглядит шаг градиентного спуска в матричном виде?

в. Найдите $d^2L(w)$. Убедитесь, что мы действительно в точке минимума.

Решение:

Ради интереса убедимся, что перед нами в качестве функции потерь используется именно MSE, в качестве x_i будем обозначать $i-$ую строчку матрицы X

$$(y - Xw)^T (y - Xw) = \begin{pmatrix} y_1 - x_1^T w & \dots & y_n - x_n^T w \end{pmatrix} \begin{pmatrix} y_1 - x_1^T w \\ \dots \\ y_n - x_n^T w \end{pmatrix} = \sum_{i=1}^{n} (y_i - x_i^T w)^2.$$

Найдём дифференциал для нашей функции потерь, держим в голове что производная берётся по вектору w

$$\begin{aligned} dL = d[(y - Xw)^T (y - Xw)] = d[(y - Xw)^T](y - Xw) + (y - Xw)^T d[(y - Xw)] = \\ = d[(-Xw)^T](y - Xw) - (y - Xw)^T X\, dw = \\ = -dw^T X^T (y - Xw) - (y - Xw)^T X\, dw = -2(y - Xw)^T X\, dw. \end{aligned}$$

Тут мы воспользовались тем, что $dw^T X^T (y - Xw)$ это скаляр и его можно транспонировать. Производная найдена. Шаг градиентного спуска будет выглядеть как

$$w_t = w_{t-1} + \gamma \cdot 2X^T (y - Xw).$$

[1] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Здесь γ — это скорость обучения. Приравняем производную к нулю, чтобы найти минимум для w. Получается система уравнений

$$2X^T(y - Xw) = 0 \qquad X^T y = X^T X w \qquad w = (X^T X)^{-1} X^T y.$$

При решении системы мы сделали предположение, что матрица $X^T X$ обратима. Это так, если в матрице X нет линейно зависимых столбцов, а также наблюдений больше чем переменных. Найдём вторую производную

$$d[-2X^T(y - Xw)] = 2X^T X \, dw.$$

Выходит, что $H = 2X^T X$. Так как матрица $X^T X$ положительно определена, по критерию Сильвестра, мы находимся в точке минимума.

Матрица $X^T X$ положительно определена по определению. Если для любого вектора $v \neq 0$ квадратичная форма $v^T X^T X v > 0$, матрица $X^T X$ положительно определена. При перемножении Xv у нас получается вектор. Обозначим его как z, значит $v^T X^T X v = z^T z = \sum_{i=1}^n z_i^2 > 0$.

Выпишем в явном виде второй дифференциал

$$d^2 L = dw^T 2X^T X \, dw.$$

Упражнение 4

Найдите следующие производные:

а. $f(X) = \operatorname{tr}(AXB)$, где матрица A размера $p \times m$, матрица B размера $n \times p$, матрица X размера $m \times n$.

б. $f(X) = \operatorname{tr}(AX^T X)$, где матрица A размера $n \times n$, матрица X размера $m \times n$.

в. $f(X) = \ln \det X$

г. $f(X) = \operatorname{tr}(AX^T X B X^{-T})$

д. $f(X) = \det(X^T A X)$

е. $f(x) = x^T A b$, где матрица A размера $n \times n$, вектора x и b размера $n \times 1$.

ж. $f(A) = x^T A b$.

Решение:

Проверить правильность своего решения можно в матричном калькуляторе[2]. Не забывайте, что $\operatorname{tr}(A) = \operatorname{tr}(A^T)$ и что под знаком следа можно циклически переставлять матрицы, если

[2] http://www.matrixcalculus.org/

размерность не ломается.

Упражнение 5

В случае Ridge-регрессии минимизируется функция со штрафом:

$$L(w) = (y - Xw)^T (y - Xw) + \lambda w^T w,$$

где λ — положительный параметр, штрафующий функцию за слишком большие значения w.

а. Найдите $dL(w)$, выведите формулу для оптимального w.

б. Как выглядит шаг градиентного спуска в матричном виде?

в. Найдите $d^2L(w)$. Убедитесь, что мы действительно в точке минимума.

Решение:

$$\begin{aligned}
dL &= 2(y - Xw)^T X\, dw + 2\lambda w^T\, dw \\
\nabla L(w) &= 2X^T(Xw - y) + 2\lambda w \\
w &= (X^T X + \lambda I)^{-1} X^T y \\
w_t &= w_{t-1} - \gamma \cdot (2X^T(Xw_{t-1} - y) + 2\lambda w_{t-1}) \\
d^2L &= dw^T (2X^T X + 2\lambda)\, dw \\
H &= 2X^T X + 2\lambda \text{ положительно определена}
\end{aligned}$$

Упражнение 6

Пусть x_i — вектор-столбец $k \times 1$, y_i — скаляр, равный $+1$ или -1, w — вектор-столбец размера $k \times 1$. Рассмотрим логистическую функцию потерь с l_2 регуляризацией

$$L(w) = \sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^T w)) + \lambda w^T w$$

а. Найдите dL;

б. Найдите вектор-столбец ∇L.

в. Как для этой функции потерь выглядит шаг градиентного спуска в матричном виде?

Решение:

Используем весь арсенал, который обсудили выше. Начнём с одного слагаемого. Обозначим его как y. Это скаляр, значит

$$d \ln y = \frac{1}{y}\, dy = \frac{1}{\ln(1 + \exp(-y_i x_i^T w))} \cdot -y_i \exp(-y_i x_i^T w) \cdot x_i^T\, dw.$$

Выписываем дифференциал

$$dL = \left(-\sum_{i=1}^{n} \frac{y_i \exp(-y_i x_i^T w)}{1 + \exp(-y_i x_i^T w)} \cdot x_i^T + 2\lambda w^T \right) dw.$$

Можно записать градиент с помощью сигмоиды $\sigma(z) = \frac{1}{1+\exp(-z)}$. Получится, что

$$\nabla L = \sum_{i=1}^{n} -y_i \sigma(-y_i x_i^T w) x_i + 2\lambda w.$$

Выходит, что шаг градиентного спуска можно записать как

$$w_t = w_{t-1} + \gamma \cdot \nabla L.$$

Упражнение 7

Упражняемся в матричном методе максимального правдоподобия. Допустим, что выборка размера n пришла к нам из многомерного нормального распределения с неизвестными вектором средних μ и ковариационной матрицей Σ. В этом задании нужно найти оценки максимального правдоподобия для $\hat{\mu}$ и $\hat{\Sigma}$. Обратите внимание, что выборкой здесь будет не $x_1, \ldots, x_n$, а

$$\begin{pmatrix} x_{11}, \ldots, x_{n1} \\ \ldots \\ x_{n1}, \ldots, x_{nm} \end{pmatrix}$$

Решение:

Плотность распределения для m—мерного вектора y будет выглядеть как

$$f(x \mid \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^m \cdot \sqrt{\det \Sigma}} \cdot \exp\left(-\frac{1}{2} \cdot (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

В силу того, что все наблюдения независимы, функция правдоподобия для выборки объёма n примет вид:

$$L(x \mid \mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^{m \cdot n} \cdot \sqrt{\det \Sigma}^n} \cdot \exp\left(-\frac{1}{2} \cdot \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right).$$

Прологарифмировав правдоподобие, получим

$$\ln L(x \mid \mu, \Sigma) = -\frac{m \cdot n}{2} \ln 2\pi - \frac{n}{2} \ln \det \Sigma - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

Нам нужно найти максимум этой функции по μ и Σ. Начнём с μ. Аргумент Σ будем считать константой. Обозначим такую функцию за $f(\mu)$. Эта функция бьёт с множества векторов в множество скаляров. Значит дифференциал этой функции можно записать в виде:

$$df(\mu) = \nabla f^T d\mu.$$

Найдём этот дифференциал. Не будем забывать, что дифференциал от константы нулевой, а также что дифференциал суммы равен сумме дифференциалов

$$\begin{aligned} df(\mu) = -\frac{1}{2} \cdot d \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = -\frac{1}{2} \cdot \sum_{i=1}^{n} d[(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)] = \\ = -\frac{1}{2} \cdot \sum_{i=1}^{n} d[(x_i - \mu)^T] \Sigma^{-1} (x_i - \mu) + (x_i - \mu)^T \Sigma^{-1} d[(x_i - \mu)] = \\ = \frac{1}{2} \cdot \sum_{i=1}^{n} d\mu^T \Sigma^{-1} (x_i - \mu) + (x_i - \mu)^T \Sigma^{-1} d\mu. \end{aligned}$$

Первое слагаемое под суммой имеет размерность $1 \times m \cdot m \times m \cdot m \times 1$. Это константа. Если мы протранспонируем константу, ничего не изменится. Обратим внимание, что матрица Σ симметричная и при транспонировании не меняется. Сделаем этот трюк

$$\begin{aligned} \frac{1}{2} \cdot \sum_{i=1}^{n} d\mu^T \Sigma^{-1} (x_i - \mu) + (x_i - \mu)^T \Sigma^{-1} d\mu = \frac{1}{2} \cdot \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} d\mu + (x_i - \mu)^T \Sigma^{-1} d\mu = \\ = \frac{1}{2} \cdot \sum_{i=1}^{n} [(x_i - \mu)^T \Sigma^{-1} + (x_i - \mu)^T \Sigma^{-1}] d\mu = \left[\cdot \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} \right] d\mu \end{aligned}$$

Получается, что $f'(\mu) = \sum_{i=1}^{n} \Sigma^{-1} (x_i - \mu)$. Приравняв производную к нулю и домножив обе части уравнения слева на Σ, получим оптимальное значению μ:

$$\begin{aligned} \sum_{i=1}^{n} \Sigma^{-1} (x_i - \hat{\mu}) = 0 \\ \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \\ \sum_{i=1}^{n} x_i = n \cdot \hat{\mu} \Rightarrow \hat{\mu} = \bar{x}. \end{aligned}$$

Не будем забывать, что в записях выше x и μ были векторами-столбцами размерности $m \times 1$. В итоговом ответе они также являются векторами-столбцами такой размерности.

Займёмся оценкой для Σ. Аргумент μ будем считать константой. Обозначим такую функцию за $f(\Sigma)$

$$f(\Sigma) = -\frac{n}{2} \ln \det \Sigma - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu).$$

Эта функция бьёт с множества матриц в множество скаляров. Значит дифференциал этой функции можно записать в виде:

$$df(\Sigma) = \text{tr}(\nabla f^T dx).$$

Начнём с первого слагаемого. Для него нам понадобится вспомнить как выглядит дифференциал для определителя

$$-\frac{n}{2} \frac{1}{\det \Sigma} d[\det \Sigma] = -\frac{n}{2} \frac{1}{\det \Sigma} \text{tr}(\det \Sigma \cdot \Sigma^{-T} d\Sigma) = -\text{tr}(\frac{n}{2} \cdot \Sigma^{-1} d\Sigma).$$

Теперь поработаем со вторым слагаемым. В нём нас интересует дифференциал обратной матрицы

$$-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T d[\Sigma^{-1}] (x_i - \mu) = \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} \cdot d\Sigma \cdot \Sigma^{-1} (x_i - \mu).$$

Под знаком суммы размерность каждого слагаемого $1 \times m \cdot m \times m \cdot m \times m \cdot m \times m \cdot m \times 1$. Это константа. Если мы возьмём от неё след, ничего не изменится. Взяв след, переставим внутри множители

$$\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} \cdot d\Sigma \cdot \Sigma^{-1} (x_i - \mu) = \frac{1}{2} \sum_{i=1}^{n} \text{tr}(\Sigma^{-1} (x_i - \mu) \cdot (x_i - \mu)^T \Sigma^{-1} \cdot d\Sigma).$$

Сумма следов — след суммы. Объединяем наши слагаемые в месте. В первом множитель n подменяем на сумму

$$df(\Sigma) = \text{tr} \left(\left[-\frac{1}{2} \sum_{i=1}^{n} \Sigma^{-1} + \Sigma^{-1} (x_i - \mu) \cdot (x_i - \mu)^T \Sigma^{-1} \right] d\Sigma \right)$$

Забираем себе из-под знака дифференциала производную. Под знаком суммы после транспонирования ничего не поменяется. Приравниваем производную к нулю, домножим справа каждое слагаемое на Σ. На четвёртой строчке домножим слева на Σ:

$$\frac{1}{2}\sum_{i=1}^{n} -\Sigma^{-1} + \Sigma^{-1}(x_i - \mu)\cdot(x_i-\mu)^T\Sigma^{-1} = 0$$
$$-n\cdot\Sigma^{-1} + \sum_{i=1}^{n}\Sigma^{-1}(x_i - \mu)\cdot(x_i-\mu)^T\Sigma^{-1} = 0$$
$$-n + \Sigma^{-1}\sum_{i=1}^{n}(x_i - \mu)\cdot(x_i-\mu)^T = 0$$
$$-n\Sigma + \sum_{i=1}^{n}(x_i - \mu)\cdot(x_i-\mu)^T = 0$$
$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)\cdot(x_i-\mu)^T$$

До оценок остался один шаг. Вспоминаем оценку для μ, подставляем её в уравнение и получаем, что

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\cdot(x_i-\bar{x})^T.$$

Не забываем, что x_i и $\bar{x}$ — вектора размерности $m \times 1$.

Упражнение 8

Найдите симметричную матрицу X наиболее близкую к матрице A по норме Фробениуса, $\sum_{i,j}(x_{ij} - a_{ij})^2$. Тут мы просто из каждого элемента вычитаем каждый и смотрим на сумму квадратов таких разностей. То есть решите задачку условной матричной минимизации

$$\begin{cases} ||X - A||^2 \to \min_A \\ X^T = X \end{cases}$$

Hint: Надо будет выписать Лагранджиан. А ещё пригодится тот факт, что $\sum_{i,j}(x_{ij} - a_{ij})^2 = ||X - A||^2 = \operatorname{tr}((X-A)^T(X-A))$.

Решение:

Выписываем лагранджиан

$$\mathcal{L} = \sum_{i,j}(x_{ij} - a_{ij})^2 + \sum_{ij}\lambda_{ij}(x_{ij} - x_{ji}) = \operatorname{tr}((X-A)^T(X-A)) + \operatorname{tr}(\Lambda^T(X - X^T)) =$$
$$= \operatorname{tr}(X^TX) - 2\operatorname{tr}(X^TA) + \operatorname{tr}(A^TA) + \operatorname{tr}(\Lambda^T(X-X^T))$$

Найдём все необходимые нам дифференциалы

$$\begin{aligned}
&d\,\mathrm{tr}(X^T X) = \mathrm{tr}(\,d(X^T X)) = \mathrm{tr}(X^T\,dX) + \mathrm{tr}(\,dX^T X) = \mathrm{tr}(2X^T\,dX)\\
&d\,\mathrm{tr}(X^T A) = \mathrm{tr}(A^T\,dX)\\
&d\,\mathrm{tr}(\Lambda^T X) = \mathrm{tr}(\Lambda^T\,dX)\\
&d\,\mathrm{tr}(\Lambda^T X^T) = \mathrm{tr}(\Lambda\,dX)
\end{aligned}$$

Выписываем в яном виде производную по X

$$\frac{\partial \mathcal{L}}{\partial X} = 2X^T - 2A^T + \Lambda^T - \Lambda = 0$$

Нужно избавиться от Λ, давайте транспонируем уравнение

$$\frac{\partial \mathcal{L}}{\partial X} = 2X - 2A + \Lambda - \Lambda^T = 0,$$

а после прибавим его к исходному, тогда лишние части исчезнут

$$4X - 2A^T - 2A = 0 \qquad X = \frac{1}{2}(A + A^T).$$