Школа анализа данных

Семинар 4. Матричные вычисления

Курс: Байесовские методы в машинном обучении, осень 2015

Дифференцирование по вектору и матрице

Задача 1. $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^T\boldsymbol{a} = \boldsymbol{a}$.

$$\frac{\partial}{\partial x_p}\boldsymbol{x}^T\boldsymbol{a} = \frac{\partial}{\partial x_p}\sum_{i=1}^n x_i a_i = a_p.$$

Задача 2. $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^T A\boldsymbol{x} = (A + A^T)\boldsymbol{x}$.

$$\frac{\partial}{\partial x_p}\boldsymbol{x}^T A\boldsymbol{x} = \frac{\partial}{\partial x_p}\sum_{i,j=1}^n a_{ij}x_i x_j = 2a_{pp}x_p + \sum_{j\neq p} a_{pj}x_j + \sum_{i\neq p} a_{ip}x_i = \sum_{j=1}^n a_{pj}x_j + \sum_{i=1}^n a_{ip}x_i = (A\boldsymbol{x})_p + (A^T\boldsymbol{x})_p.$$

Задача 3. $\frac{\partial}{\partial \boldsymbol{x}}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = 2A^T(A\boldsymbol{x} - \boldsymbol{b})$.

$$\frac{\partial}{\partial \boldsymbol{x}}\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \frac{\partial}{\partial \boldsymbol{x}}(A\boldsymbol{x} - \boldsymbol{b})^T(A\boldsymbol{x} - \boldsymbol{b}) = \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^T A^T A\boldsymbol{x} - \boldsymbol{x}^T A^T\boldsymbol{b} - \underbrace{\boldsymbol{b}^T A\boldsymbol{x}}_{=\boldsymbol{x}^T A^T\boldsymbol{b}} + \boldsymbol{b}^T\boldsymbol{b}\right) =$$

$$= \{\text{используем результаты задач 1 и 2}\} = 2A^T A\boldsymbol{x} - 2A^T\boldsymbol{b} = 2A^T(A\boldsymbol{x} - \boldsymbol{b}).$$

Задача 4. $\frac{\partial}{\partial A}\det A = (\det A)A^T$.

$$\frac{\partial}{\partial a_{ij}}\det A = \{\text{разложим определитель по } i\text{-ой строке}\} = \frac{\partial}{\partial a_{ij}}\sum_{k=1}^n a_{ik}M_{ik} =$$

$$= \{M_{ik} - \text{алгебраическое дополнение к элементу } a_{ik}, \text{ оно не зависит от } a_{ij}\ \forall k\} = M_{ij} = (\det A)(A^{-1})_{ji}.$$

Задача 5. $\frac{\partial}{\partial A}\operatorname{tr}(AB) = B^T$.

$$\frac{\partial}{\partial a_{pq}}\operatorname{tr}(AB) = \frac{\partial}{\partial a_{pq}}\sum_{i,j=1}^n a_{ij}b_{ji} = b_{qp}.$$

Задача 6. $\frac{\partial}{\partial A}\boldsymbol{x}^T A\boldsymbol{y} = \boldsymbol{x}\boldsymbol{y}^T$.

$$\frac{\partial}{\partial A}\boldsymbol{x}^T A\boldsymbol{y} = \frac{\partial}{\partial A}\operatorname{tr}(\boldsymbol{x}^T A\boldsymbol{y}) = \{\text{используем свойство } \operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)\} =$$

$$= \frac{\partial}{\partial A}\operatorname{tr}(A\boldsymbol{y}\boldsymbol{x}^T) = \{\text{используем результат задачи 5}\} = (\boldsymbol{y}\boldsymbol{x}^T)^T = \boldsymbol{x}\boldsymbol{y}^T.$$

Задача 7. $\frac{\partial}{\partial x}\log\det A = \operatorname{tr}\left(\frac{\partial A}{\partial x}A^{-1}\right)$. Здесь $x \in \mathbb{R}$, каждый элемент A зависит от x.

$$\frac{\partial}{\partial x}\log\det A = \frac{1}{\det A}\frac{\partial \det A}{\partial x} = \left\{\text{производная сложной функции, зависящей от матрицы}\right.$$

$$\left.\frac{\partial}{\partial x}f(a_{11}(x), a_{12}(x), \ldots, a_{nm}(x)) = \sum_{i,j=1}^{n,m}\frac{\partial f}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial x}\right\} = \frac{1}{\det A}\sum_{i,j=1}^{n,m}\frac{\partial \det A}{\partial a_{ij}}\frac{\partial a_{ij}}{\partial x} =$$

$$= \{\text{используем результат задачи 4}\} = \frac{1}{\det A}\sum_{i,j=1}^{n,m}\det A(A^{-T})_{ij}\left(\frac{\partial A}{\partial x}\right)_{ij} = \sum_{i,j=1}^{n,m}\left(\frac{\partial A}{\partial x}\right)_{ij}(A^{-1})_{ji} = \operatorname{tr}\left(\frac{\partial A}{\partial x}A^{-1}\right).$$

Задача 8. $\frac{\partial}{\partial x}A^{-1} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$. Здесь $x \in \mathbb{R}$, каждый элемент A зависит от x.

По определению обратной матрицы
$$A^{-1}A = I.$$
Дифференцируя обе части равенства по x, получаем:
$$\frac{\partial I}{\partial x} = 0 = \frac{\partial}{\partial x}(A^{-1}A) = \frac{\partial A^{-1}}{\partial x}A + A^{-1}\frac{\partial A}{\partial x}.$$
Выражая из последнего равенства $\frac{\partial A^{-1}}{\partial x}$, получаем требуемый результат.

Задача 9. $\frac{\partial}{\partial A}\operatorname{tr}(A^{-1}B) = -(A^{-1}BA^{-1})^T$.

$$\frac{\partial}{\partial a_{pq}}\operatorname{tr}(A^{-1}B) = \operatorname{tr}\left(\frac{\partial A^{-1}B}{\partial a_{pq}}\right) = \operatorname{tr}\left(\frac{\partial A^{-1}}{\partial a_{pq}}B\right) = \{\text{используем результат задачи 8}\} = \operatorname{tr}\left(-A^{-1}\frac{\partial A}{\partial a_{pq}}A^{-1}B\right) =$$
$$= \{\text{обозначим через } I_{pq} \text{ матрицу, в которой в позиции } pq \text{ стоит 1, а в остальных нули}\} =$$
$$= -\operatorname{tr}(A^{-1}BA^{-1}I_{pq}) = -\sum_{i,j=1}^{n}(A^{-1}BA^{-1})_{ij}(I_{pq})_{ji} = -(A^{-1}BA^{-1})_{qp}.$$

Нормальное распределение

По определению плотность нормального распределения выглядит как
$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{2\pi}^D\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$
Здесь
$$\mathbb{E}\boldsymbol{x} = \boldsymbol{\mu},$$
$$\mathbb{E}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T = \Sigma.$$

Задача 10. $\mathbb{E}\boldsymbol{x}^T\boldsymbol{x} = \operatorname{tr}\Sigma + \boldsymbol{\mu}^T\boldsymbol{\mu}$.

$$\Sigma = \mathbb{E}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T) =$$
$$= \mathbb{E}\boldsymbol{x}\boldsymbol{x}^T - \underbrace{\mathbb{E}\boldsymbol{x}}_{=\boldsymbol{\mu}}\boldsymbol{\mu}^T - \mathbb{E}\boldsymbol{\mu}\boldsymbol{x}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{\mu}\boldsymbol{\mu}^T.$$
$$\mathbb{E}\boldsymbol{x}^T\boldsymbol{x} = \mathbb{E}\operatorname{tr}(\boldsymbol{x}^T\boldsymbol{x}) = \mathbb{E}\operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^T) = \operatorname{tr}\mathbb{E}\boldsymbol{x}\boldsymbol{x}^T = \operatorname{tr}(\Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T) = \operatorname{tr}\Sigma + \operatorname{tr}(\boldsymbol{\mu}^T\boldsymbol{\mu}) = \operatorname{tr}\Sigma + \boldsymbol{\mu}^T\boldsymbol{\mu}.$$

Задача 11. Пусть имеется независимая выборка $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ из $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$. Требуется найти оценки максимального правдоподобия для параметров $\boldsymbol{\mu},\Sigma$.

Функция правдоподобия:
$$p(X|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N}p(\boldsymbol{x}_n|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu},\Sigma) = \prod_{n=1}^{N}\frac{1}{\sqrt{2\pi}^D\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right) =$$
$$= \left(\frac{1}{\sqrt{2\pi}^D\sqrt{\det\Sigma}}\right)^N\exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right).$$
$$\log p(X|\boldsymbol{\mu},\Sigma) = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log\det\Sigma - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}).$$
Дифференцируя $\log p(X|\boldsymbol{\mu},\Sigma)$ по $\boldsymbol{\mu}$, получаем:
$$\frac{\partial}{\partial\boldsymbol{\mu}}\log p(X|\boldsymbol{\mu},\Sigma) = -\frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial\boldsymbol{\mu}}\left(\boldsymbol{x}_n^T\Sigma^{-1}\boldsymbol{x}_n - 2\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x}_n + \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right) =$$
$$= -\frac{1}{2}\sum_{n=1}^{N}(-2\Sigma^{-1}\boldsymbol{x}_n + 2\Sigma^{-1}\boldsymbol{\mu}) = \Sigma^{-1}\left(\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})\right) = 0.$$

Домножая в последнем равенстве на невырожденную матрицу Σ, получаем

$$\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})=0 \ \Rightarrow\ \boldsymbol{\mu}_{ML}=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n.$$

Теперь дифференцируем $\log p(X|\boldsymbol{\mu},\Sigma)$ по Σ:

$$\frac{\partial}{\partial\Sigma}\log p(X|\boldsymbol{\mu},\Sigma)=-\frac{N}{2}\frac{\partial}{\partial\Sigma}\log\det\Sigma-\frac{1}{2}\frac{\partial}{\partial\Sigma}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})=$$

$$=-\frac{N}{2}\frac{1}{\det\Sigma}\frac{\partial\det\Sigma}{\partial\Sigma}-\frac{1}{2}\frac{\partial}{\partial\Sigma}\operatorname{tr}\left(\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})\right)=\{\text{используем результат задачи 4}\}=$$

$$=-\frac{N}{2}\Sigma^{-T}-\frac{1}{2}\frac{\partial}{\partial\Sigma}\operatorname{tr}\left(\Sigma^{-1}\underbrace{\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T}_{\text{обозначим через } S}\right)=\{\text{используем результат задачи 9}\}=-\frac{N}{2}\Sigma^{-T}+\frac{1}{2}\Sigma^{-T}S^T\Sigma^{-T}=0.$$

Домножая в последнем равенстве слева и справа на Σ^T, получаем:

$$-\frac{N}{2}\Sigma^T+\frac{1}{2}S^T=0 \ \Rightarrow\ \Sigma_{ML}=\frac{1}{N}S.$$

Заметим, что при взятии производной $\log p(X|\boldsymbol{\mu},\Sigma)$ по Σ нигде не использовалось свойство симметричности Σ. Этого оказалось достаточно, т.к. итоговый ответ $\frac{1}{N}S$ получился симметричной матрицей. Если бы это оказалось не так, то тогда пришлось бы вычислять производную с дополнительным учётом свойства симметричности. Тогда, в частности, при решении задачи 9 при вычислении $\frac{\partial A}{\partial a_{pq}}$ возникла бы матрица с единицей в позициях pq и qp.

Задача 12. Рассмотрим многомерное нормальное распределение $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$ и разобьём вектор $\boldsymbol{x}$ на два подвектора $\boldsymbol{x}_a$ и $\boldsymbol{x}_b$. Тогда

$$\boldsymbol{x}=\begin{bmatrix}\boldsymbol{x}_a\\ \boldsymbol{x}_b\end{bmatrix},\quad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix},\quad \Sigma=\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ab}^T & \Sigma_{bb}\end{bmatrix},\quad \Lambda:=\Sigma^{-1}=\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ab}^T & \Lambda_{bb}\end{bmatrix}.$$

Требуется найти распределение $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$.

$$p(\boldsymbol{x}_a|\boldsymbol{x}_b)\propto p(\boldsymbol{x}_a,\boldsymbol{x}_b)=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)\propto$$

$$\propto\exp\left[-\frac{1}{2}\left((\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{aa}(\boldsymbol{x}_a-\boldsymbol{\mu}_a)+(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)+(\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\Lambda_{ab}^T(\boldsymbol{x}_a-\boldsymbol{\mu}_a)\right)\right]\propto$$

$$\propto\exp\left[-\frac{1}{2}\left(\boldsymbol{x}_a^T\Lambda_{aa}\boldsymbol{x}_a-2\boldsymbol{x}_a^T(\Lambda_{aa}\boldsymbol{\mu}_a-\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b))\right)\right].$$

Видно, что искомое распределение с точностью до нормировочной константы имеет вид «экспонента от квадратичной функции». Отсюда заключаем, что $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ является многомерным нормальным распределением. Осталось в последнем выражении выделить полный квадрат и получить параметры распределения $\hat{\boldsymbol{\mu}},\hat{\Sigma}$:

$$\hat{\Sigma}^{-1}=\Lambda_{aa} \ \Rightarrow\ \hat{\Sigma}=\Lambda_{aa}^{-1},$$
$$\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}=\Lambda_{aa}\boldsymbol{\mu}_a-\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) \ \Rightarrow\ \hat{\boldsymbol{\mu}}=\Lambda_{aa}^{-1}(\Lambda_{aa}\boldsymbol{\mu}_a-\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b))=\boldsymbol{\mu}_a-\Lambda_{aa}^{-1}\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b).$$

В итоге получаем $p(\boldsymbol{x}_a|\boldsymbol{x}_b)=\mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a-\Lambda_{aa}^{-1}\Lambda_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b),\Lambda_{aa}^{-1})$.