Теоретическое задание №2
«Матричные вычисления»

курс «Байесовские методы в машинном обучении»
кафедра ММП ВМК МГУ

Михеев Борис, 417 группа

20 октября 2022 г.

1. Доказать тождество Вудбери:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Здесь $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times m}$, $U \in \mathbb{R}^{n\times m}$, $V \in \mathbb{R}^{m\times n}$.

► Докажем, что произведение $(A + UCV)$ и правой части тождества дает в произведении I. Достаточно будет рассмотреть умножение справа в силу единственности существования обратной матрицы.

$$\begin{aligned}&(A + UCV)(A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) = I - U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} + UCVA^{-1} -\\&- UCVA^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} = I + UCVA^{-1} - (U + UCVA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} =\\&= I + UCVA^{-1} - UC(C^{-1} + VA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1} = I + UCVA^{-1} - UCVA^{-1} = I. \quad \blacksquare\end{aligned}$$

2. Пусть $p(x) = \mathcal{N}(x|\mu, \Sigma)$, $p(y|x) = \mathcal{N}(y|Ax, \Gamma)$, $A \in \mathbb{R}^{m\times n}$. Найти распределение $p(x|y)$.

► Запишем выражение для $p(x|y)$ по теореме Байеса:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}.$$

Правдоподобие $p(y|x)$ и априорное распределение $p(x)$ являются сопряженными друг к другу, следовательно, интеграл в знаменателе аналитически берется, и апостериорное распределение $p(x|y)$ будет лежать в том же семействе, что априорное распределение $p(x)$. Т. к. $p(x) = \mathcal{N}(x|\mu, \Sigma) \Rightarrow p(x|y)$ будет также нормальным распределением: $p(x|y) = \mathcal{N}(x|\widetilde{\mu}, \widetilde{\Sigma})$. Оценим параметры этого распределения $\widetilde{\mu}$ и $\widetilde{\Sigma}$ также с помощью теоремы Байеса и выражения для $p(x|y)$. Знаменатель в нем является по сути нормировочной константой, для нахождения параметров рассматриваемого нормального распределения достаточно будет рассмотреть числитель:

$$\begin{aligned}&p(y|x)p(x) = \mathcal{N}(y|Ax, \Gamma)\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{m}{2}}\sqrt{det\Gamma}}e^{-\frac{1}{2}(y-Ax)^T\Gamma^{-1}(y-Ax)} \cdot \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{det\Sigma}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} =\\&= \frac{1}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Gamma det\Sigma}}e^{-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))}.\end{aligned}$$

Далее для определения параметров распределения достаточно будет рассмотреть лишь показатель экспоненты, в том числе и с точки зрения удобства записи и наглядности:

$$
\begin{aligned}
&(y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu)=y^T\Gamma^{-1}(y-Ax)-x^TA^T\Gamma^{-1}(y-Ax)+x^T\Sigma^{-1}(x-\mu)-\\
&-\mu^T\Sigma^{-1}(x-\mu)=x^T\Sigma^{-1}x-x^T\Sigma^{-1}\mu-x^TA^T\Gamma^{-1}y+x^TA^T\Gamma^{-1}Ax+y^T\Gamma^{-1}(y-Ax)-\mu^T\Sigma^{-1}(x-\mu)=\\
&=x^T\Sigma^{-1}x+x^TA^T\Gamma^{-1}Ax-x^T\Sigma^{-1}\mu-x^TA^T\Gamma^{-1}y+y^T\Gamma^{-1}(y-Ax)-\mu^T\Sigma^{-1}(x-\mu)=\\
&=x^T(\Sigma^{-1}+A^T\Gamma^{-1}A)x-x^T(\Sigma^{-1}\mu+A^T\Gamma^{-1}y)+y^T\Gamma^{-1}(y-Ax)-\mu^T\Sigma^{-1}(x-\mu).
\end{aligned}
$$

Матрица, стоящая в квадратичном слагаемом по x, соответствует обратной ковариационной матрице распределения $p(x|y)$. Т. о., $\widetilde{\Sigma}=(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1}$.

Для поиска $\widetilde{\mu}$ используем факт того, что для нормального распределения (каковым, как выяснилось, является $p(x|y)$) матожидание совпадает с модой, т. е. $\mathbb{E}(x|y)=x_{MP}$. Тогда можно найти $\widetilde{\mu}=x_{MP}$ путем дифференцирования $p(x|y)$ по x и приравнивания производной к 0. Знаменатель в соответствующем выражении является константой, следовательно, достаточно продифференцировать числитель:

$$
\begin{aligned}
&\frac{\partial}{\partial x}\Big(\frac{1}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Gamma det\Sigma}}e^{-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))}\Big)=\\
&=\frac{e^{-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))}}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Gamma det\Sigma}}\cdot\frac{\partial}{\partial x}\Big(-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))\Big).
\end{aligned}
$$

Для удобства записи и наглядности достаточно рассмотреть производную показателя экспоненты:

$$
\begin{aligned}
&\frac{\partial}{\partial x}\Big(-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))\Big)=-\frac{1}{2}(-2A^T\Gamma^{-1}(y-Ax)+2\Sigma^{-1}(x-\mu))=\\
&=A^T\Gamma^{-1}(y-Ax)-\Sigma^{-1}(x-\mu)=A^T\Gamma^{-1}y-A^T\Gamma^{-1}Ax-\Sigma^{-1}x+\Sigma^{-1}\mu=A^T\Gamma^{-1}y+\Sigma^{-1}\mu-\\
&-(\Sigma^{-1}+A^T\Gamma^{-1}A)x=0\Rightarrow(\Sigma^{-1}+A^T\Gamma^{-1}A)x=\Sigma^{-1}\mu+A^T\Gamma^{-1}y\Rightarrow\\
&\Rightarrow x_{MP}=(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1}(\Sigma^{-1}\mu+A^T\Gamma^{-1}y)=\widetilde{\mu}.
\end{aligned}
$$

Т. о., мы определили вид и параметры распределения $p(x|y)$, т. е. нашли искомое распределение: $p(x|y)=\mathcal{N}(x|(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1}(\Sigma^{-1}\mu+A^T\Gamma^{-1}y),(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1})$.

3. Пусть $p(x)=\mathcal{N}(x|\mu,\Sigma)$, $p(y|x)=\mathcal{N}(y|Ax,\Gamma)$. Доказать, что $p(y)=\mathcal{N}(y|A\mu,\Gamma+A\Sigma A^T)$.

► Запишем $p(y)$ по правилу суммирования вероятностей: $p(y)=\int p(y|x)p(x)dx$. Интеграл аналитически берется, т. к. априорное распределение $p(x)$ и апостериорное $p(y|x)$ являются нормальными распределениями и сопрягаются, под интегралом будет произведение парабол под экспонентой, т. е. тоже парабола под экспонентой, и априорное распределение $p(y)$ будет нормальным распределением, как и апостериорное, т. е. будет иметь вид: $p(y)=\mathcal{N}(y|\mathbb{E}y,\mathbb{D}y)$. Можно показать это более строго, расписав выражение для $p(y)$, и использовав результаты предыдущей задачи. Используем обозначения $\widetilde{\Sigma}=(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1}$,
$\widetilde{\mu}=(\Sigma^{-1}+A^T\Gamma^{-1}A)^{-1}(\Sigma^{-1}\mu+A^T\Gamma^{-1}y)$. Тогда:

$$p(y) = \int p(y|x)p(x)dx = \frac{1}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Sigma det\Gamma}} \int e^{-\frac{1}{2}((y-Ax)^T\Gamma^{-1}(y-Ax)+(x-\mu)^T\Sigma^{-1}(x-\mu))}dx =$$
$$= \frac{1}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Sigma det\Gamma}} \int e^{-\frac{1}{2}(x^T(A^T\Gamma^{-1}A+\Sigma^{-1})x-x^T(A^T\Gamma^{-1}y+\Sigma^{-1}\mu)-(y^T\Gamma^{-1}A+\mu^T\Sigma^{-1})x+y^T\Gamma^{-1}y+\mu^T\Sigma^{-1}\mu)}dx =$$
$$= \frac{e^{-\frac{1}{2}(y^T\Gamma^{-1}y+\mu^T\Sigma^{-1}\mu)}}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Sigma det\Gamma}} \int e^{-\frac{1}{2}(x^T(A^T\Gamma^{-1}A+\Sigma^{-1})x-x^T(A^T\Gamma^{-1}y+\Sigma^{-1}\mu)-(y^T\Gamma^{-1}A+\mu^T\Sigma^{-1})x)}dx.$$

Заметим, что $(A^T\Gamma^{-1}y + \Sigma^{-1}\mu) = \widetilde{\Sigma}^{-1}\widetilde{\mu}$. Прибавим и вычтем к показателю экспоненты слагаемое $\widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu}$, и воспользуемся фактом симметричности ковариационной матрицы:

$$p(y) = \frac{e^{-\frac{1}{2}(y^T\Gamma^{-1}y+\mu^T\Sigma^{-1}\mu)}}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Sigma det\Gamma}} \int e^{-\frac{1}{2}(x^T\widetilde{\Sigma}^{-1}x-x^T\widetilde{\Sigma}^{-1}\widetilde{\mu}-\widetilde{\mu}^T\widetilde{\Sigma}^{-1}x+\widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu}-\widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu})}dx =$$
$$= \frac{e^{-\frac{1}{2}(y^T\Gamma^{-1}y+\mu^T\Sigma^{-1}\mu-\widetilde{\mu}^T\widetilde{\Sigma}^{-1}\widetilde{\mu})}}{(2\pi)^{\frac{m+n}{2}}\sqrt{det\Sigma det\Gamma}} \int e^{-\frac{1}{2}(x-\widetilde{\mu})^T\widetilde{\Sigma}^{-1}(x-\widetilde{\mu})}dx.$$

В полученном выражении под интегралом стоит ненормированная плотность многомерного нормального распределения $\mathcal{N}(x|\widetilde{\mu}, \widetilde{\Sigma})$, дробь перед ним не зависит от x, зависит от y, содержит экспоненту с квадратичным слагаемым по y в показателе. Таким образом, $p(y)$ также будет многомерным нормальным распределением вида $\mathcal{N}(y|\mathbb{E}y, \mathbb{D}y)$.

Найдем теперь параметры данного распределения. Т. к. $p(y|x) = \mathcal{N}(y|Ax, \Gamma)$, то справедливо следующее представление:

$$y = Ax + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Gamma) \Rightarrow \mathbb{E}y = \mathbb{E}(Ax) + \mathbb{E}\varepsilon = A\mathbb{E}x = A\mu.$$

$$\mathbb{D}y = \mathbb{D}(Ax + \varepsilon) = \mathbb{D}(Ax) + \mathbb{D}\varepsilon = A\mathbb{D}xA^T + \Gamma = A\Sigma A^T + \Gamma.$$

Т. о., нашли распределение $p(y)$: $p(y) = \mathcal{N}(y|A\mu, A\Sigma A^T + \Gamma)$. ■

4. Вычислить $\frac{\partial}{\partial x}det(X^{-1} + A)$ (все матрицы не являются симметричными).

► Дифференцируемое выражение можно представить в виде $det(X^{-1} + A) = f_2(f_1(X))$, где $f_1(Y) = (Y^{-1} + A)$, $f_2(Y) = detY$. Тогда для вычисления производной применим правило дифференцирования сложной функции: $D(f_2 \circ f_1)(x)[h] = D(f_2(f_1(x)))[D(f_1(x))[h]]$.

Рассмотрим $f_2(Y) = detY$, найдем $\frac{\partial}{\partial Y}detY$. Используем разложение определителя матрицы по i-ой строке: $detY = \sum\limits_{k=1}^{n} y_{ik}Y_{ik}$, где n – размерность квадратной матрицы $Y \in \mathbb{R}^{n\times n}$, Y_{ik} – алгебраическое дополнение к элементу y_{ik}. Продифференцируем поэлементно:

$$\frac{\partial}{\partial y_{ij}}detY = \frac{\partial}{\partial y_{ij}}(\sum_{k=1}^{n} y_{ik}Y_{ik}) = Y_{ij}.$$

Вспомним также, что по определению обратной матрицы: $Y^{-1} = \frac{1}{detY}(Y_{ij})_{i,j=1}^n$, где $(Y_{ij})_{i,j=1}^n$ – матрица, составленная из алгебраических дополнений к соответствующим элементам y_{ij} матрицы Y. Тогда:

$$\frac{\partial}{\partial y_{ij}}detY = Y_{ij} = detY \cdot (Y^{-1})_{ij} \Rightarrow \frac{\partial}{\partial Y}detY = detY \cdot Y^{-1}.$$

Далее можем записать дифференциал в каноническом виде:

$$D(f_2(Y))[h] = detY \cdot Y^{-1}h = \left\langle detY \cdot Y^{-T}, h \right\rangle = tr(detY \cdot Y^{-1}h).$$

Рассмотрим $f_1(Y) = (Y^{-1} + A)$. По правилу дифференцирования суммы получим:

$$\frac{\partial}{\partial Y}(Y^{-1} + A) = \frac{\partial}{\partial Y}Y^{-1} + \frac{\partial}{\partial Y}A = \frac{\partial}{\partial Y}Y^{-1}.$$

По определению, $YY^{-1} = I$. Возьмем дифференциал от обеих частей равенства и применим правило дифференцирования произведения:

$$D(YY^{-1})[h] = D(Y)[h]Y^{-1} + YD(Y^{-1})[h] = hY^{-1} + YD(Y^{-1})[h] = 0 \Rightarrow$$
$$\Rightarrow YD(Y^{-1})[h] = -hY^{-1} \Rightarrow D(Y^{-1})[h] = -Y^{-1}hY^{-1}.$$

Теперь можем подставить найденные величины в выражение производной сложной функции, применив в дальнейшем циклическое свойство следа:

$$D(f_2 \circ f_1)(x)[h] = tr(det(X^{-1} + A) \cdot (X^{-1} + A)^{-1}(-X^{-1}hX^{-1})) =$$
$$= tr(-det(X^{-1} + A) \cdot (X^{-1} + A)^{-1}X^{-1}hX^{-1}) = tr(-det(X^{-1} + A) \cdot X^{-1}(X^{-1} + A)^{-1}X^{-1}h) \Rightarrow$$
$$\Rightarrow \frac{\partial}{\partial X}det(X^{-1} + A) = (-det(X^{-1} + A) \cdot X^{-1}(X^{-1} + A)^{-1}X^{-1})^T = -det(X^{-1} + A) \cdot X^{-T}(X^{-1} + A)^{-T}X^{-T} \Rightarrow$$
$$\Rightarrow \frac{\partial}{\partial X}det(X^{-1} + A) = -det(X^{-1} + A) \cdot X^{-T}(X^{-1} + A)^{-T}X^{-T}.$$

5. Вычислить $\frac{\partial}{\partial X}tr(AX^{-T}BXC)$ (все матрицы не являются симметричными, матрицы A, C не являются квадратными).

► Запишем выражение для следа матрицы в виде скалярного произведения:

$$tr(AX^{-T}BXC) = \left\langle I_{\widetilde{n}}, AX^{-T}BXC \right\rangle,$$

где $I_{\widetilde{n}}$ – единичная матрица необходимого и соответствующего размера (т. к. размеры матриц A, B, C явно не заданы). Запишем выражения для дифференциала, применим правила дифференцирования произведения и другие правила дифференцирования, а также результаты предыдущей задачи (найденное выражения для дифференциала обратной матрицы), и выразим значение искомого градиента через каноническую форму записи дифференциала:

$$dtr(AX^{-T}BXC) = d(\left\langle I_{\widetilde{n}}, AX^{-T}BXC \right\rangle) = \left\langle I_{\widetilde{n}}, d(AX^{-T}BXC) \right\rangle =$$
$$= \left\langle I_{\widetilde{n}}, A(d(X^{-T}))BXC + AX^{-T}B(dX)C \right\rangle = \left\langle I_{\widetilde{n}}, A(dX^{-1})^T BXC + AX^{-T}B(dX)C \right\rangle =$$
$$= \left\langle I_{\widetilde{n}}, A(-X^{-1}dXX^{-1})^T BXC + AX^{-T}B(dX)C \right\rangle = \left\langle I_{\widetilde{n}}, -AX^{-T}(dX)^T X^{-T}BXC + AX^{-T}B(dX)C \right\rangle =$$
$$= -\left\langle I_{\widetilde{n}}, AX^{-T}(dX)^T X^{-T}BXC \right\rangle + \left\langle I_{\widetilde{n}}, AX^{-T}B(dX)C \right\rangle = -\left\langle X^{-1}A^TC^TX^TB^TX^{-1}, (dX)^T \right\rangle +$$
$$+ \left\langle B^TX^{-1}A^TC^T, dX \right\rangle = -\left\langle X^{-T}BXCAX^{-T}, dX \right\rangle + \left\langle B^TX^{-1}A^TC^T, dX \right\rangle \Rightarrow$$
$$\Rightarrow \frac{\partial}{\partial X}tr(AX^{-T}BXC) = B^TX^{-1}A^TC^T - X^{-T}BXCAX^{-T}.$$