Теория 2. Матричные вычисления

Курс: Байесовские методы в машинном обучении
Студент: Никулин Александр

1. Доказать тождество Вудбери:

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

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

Чтобы тождество было справедливо, должна существовать невырожденная матрица $A + UCV$. По определению обратной матрицы, если домножим на правую часть, то должны получить единичную матрицу I. Проверим:

$$\begin{aligned}
&(A + UCV)(A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}) = \\
I + UCVA^{-1} - &\left[U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} + UCVA^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}\right] = \\
I + UCVA^{-1} - &\left[(U + UCVA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}VA^{-1}\right] = \\
I + UCVA^{-1} - &\left[UC\underbrace{(C^{-1} + VA^{-1}U)(C^{-1} + VA^{-1}U)^{-1}}_{=I}VA^{-1})\right] = I
\end{aligned}$$

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

По теореме Байеса:

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx} = \frac{1}{Z}p(y|x)p(x)$$

На нормировочную константу можно не смотреть т.к. она не влияет на параметры распределения (например, после нормировки не меняется дисперсия). Тогда остается разобраться с показателем экспоненты (остальные константы также уйдут в нормировку):

$$p(y|x)p(x) = \text{Const} \cdot \exp[-\frac{1}{2}((x-\mu)^T\Sigma^{-1} + (y - Ax)^T\Gamma^{-1}(y - Ax))]$$

По аналогии с лекцией, раскроем скобки:

$$\begin{aligned}
(x-\mu)^T\Sigma^{-1} + (y - Ax)^T\Gamma^{-1}(y - Ax) &= \\
x^T\Sigma^{-1}x - 2\mu^T\Sigma^{-1}x - 2y^T\Gamma^{-1}Ax + x^TA^T\Gamma^{-1}Ax + \text{Const} &= \\
x^T(\Sigma^{-1} + A^T\Gamma^{-1}A)x - 2(\mu^T\Sigma^{-1} + y^T\Gamma^{-1}A)x + \text{Const}&
\end{aligned}$$

После выделения полного квадрата получим:

$$\begin{aligned}
\Sigma_{\text{new}} &= (\Sigma^{-1} + A^T\Gamma^{-1}A)^{-1} \\
\mu_{\text{new}} &= \Sigma_{\text{new}}(\Sigma^{-1}\mu + A^T\Gamma^{-1}y)
\end{aligned}$$

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

Мы знаем, что нормальное распределение замкнуто относительно маргинализации (надеюсь это можно принять как факт). Также знаем, что нормальное полностью определяется своим средним и дисперсией. Тогда можно схитрить, представив y как $Ax + \epsilon, \epsilon \sim \mathcal{N}(0, \Gamma)$. Проверим, что при таком представлении $p(y|x)$ имеет нужный вид:

$$\begin{aligned}
\mathbb{E}[y|x = x'] &= \mathbb{E}[Ax' + \epsilon] = Ax' \\
Var[y|x = x'] &= Var[Ax' + \epsilon] = Var[\epsilon] = \Gamma \Rightarrow \\
&\qquad\qquad\qquad\quad y|x \sim \mathcal{N}(Ax, \Gamma)
\end{aligned}$$

Получили изначальное распределение. Найти мат.ожидание и дисперсию для y теперь также не составит труда:

$$\mathbb{E}[y] = \mathbb{E}[Ax + \epsilon] = A\mathbb{E}[x] + \mathbb{E}[\epsilon] = A\mu$$
$$Var[y] = Var[Ax + \epsilon] = Var[Ax] + \Gamma = AVar[x]A^T + \Gamma = A\Sigma A^T + \Gamma$$

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

   Вычислим дифференциал, после чего приведем к каноничной форме (формулки подсмотрены в материалах другого курса).

$$\begin{aligned}
d(\det(X^{-1} + A)) &= \\
\det(X^{-1} + A)\,\langle (X^{-1} + A)^{-T}, dX^{-1} + A\rangle &= \\
\det(X^{-1} + A)\,\langle (X^{-1} + A)^{-T}, -X^{-1}dXX^{-1}\rangle &= \\
\det(X^{-1} + A)\,\langle -X^{-T}(X^{-1} + A)^{-T}X^{-T}, dX\rangle &\Rightarrow \\
\frac{\partial \det(X^{-1} + A)}{\partial X} &= -\det(X^{-1} + A)(X^{-T}(X^{-1} + A)^{-T}X^{-T})
\end{aligned}$$

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

   Аналогично предыдущему. След определен только для квадратных матриц, поэтому можем циклически сдвигать и по размерностям сойдется:

$$\begin{aligned}
d(\operatorname{tr}(AX^{-T}BXC)) = \langle A^TC^T, d(X^{-T}BX\rangle &= \\
\langle A^TC^T, d(X^{-T}B)X + X^{-T}BdX\rangle &= \\
\langle A^TC^T, (-X^{-1}dXX^{-1})^TBX + X^{-T}BdX\rangle &= \\
\langle A^TC^T, -X^{-T}(X^TB^TX^{-1}dX)^T\rangle + \langle A^TC^T, X^{-T}BdX\rangle &= \\
\langle A^TC^TX^TB^TX^{-1}dX, -X^{-T}\rangle + \langle B^TX^{-1}A^TC^T, dX\rangle &= \\
\langle -X^{-T}BXCAX^{-T}, dX\rangle + \langle B^TX^{-1}A^TC^T, dX\rangle &\Rightarrow \\
\frac{\partial \operatorname{tr}(AX^{-T}BXC)}{\partial X} = B^TX^{-1}A^TC^T - X^{-T}BXCAX^{-T}&
\end{aligned}$$

Несмотря на страшный вид по размерностям все сходится, так что вероятно это правильно.