Bayesian Machine Learning
Course 67564
Solution To Exercise 0: Linear Algebra and Probability

Barak Haim 021682141

10/11/2022

Contents

1 Linear Algebra

1.1 Q1. $f_1(x) = (x-\mu)^T R(x-\mu)$

Denote $g_1(x) = x - \mu$**,** $g_2(x) = Rx$ **,** $g_3(x) = x$ **and** $g_4(x) = g_3(x)^T g_2(x)$**. So** $f_1(x) = g_4(g_1(x)) = g_3^T(g_1(x))g_2(g_1(x))$**. By the chain role:**

$$\frac{\partial f_1}{\partial x}(x) = \frac{\partial (g_4 \circ g_1)}{\partial x}(x) = \frac{\partial g_1(x)}{\partial x}\frac{\partial g_4(g_1(x))}{\partial g_1(x)} = \bigstar$$

Now - $\frac{\partial g_1(x)}{\partial x} = \frac{\partial x}{\partial x} - \frac{\partial \mu}{\partial x} = 1 - 0 = 1$ **and we get** $\bigstar = 1 \cdot \frac{\partial g_4(g_1(x))}{\partial g_1(x)} = \frac{\partial g_4(y)}{\partial y}$ **for** $y = g_1(x)$**. By the product role we get:**

$$\frac{\partial g_4(y)}{\partial y} = \frac{\partial (g_3^T \cdot g_2)(y)}{\partial y} = \frac{\partial g_3(y)}{\partial y}^T g_2(y) + g_3(y)^T \frac{\partial g_2(y)}{\partial y}$$

separately we have - $\frac{\partial g_3(y)}{\partial y} = \frac{\partial y}{\partial y} = 1$ **(same as** g_1 **only without the constant) and we're left with -** $\frac{\partial g_2(y)}{\partial y}$**. Since** $Ry = \begin{bmatrix} \vdots \\ R_i^T y \\ \vdots \end{bmatrix}$ **where** R_i^T **is the i'th row vector of R (in the recitation - eq 3.3 they're marked as** $a_i's$**). For each** $1 \le i \le n$ **we get** $R_i^T y = \sum_{j=1}^{n} R_{ij} y_j$**. The partial derivative for** $1 \le k \le n$ **is:**

$$\frac{\partial}{\partial y_k} \sum_{j=1}^{n} R_{ij} y_j = R_{ik}$$

So long story short $\frac{\partial g_2(y)}{\partial y} = R$**. All Together now:**

$$\begin{aligned} &\frac{\partial g_3(y)}{\partial y}^T g_2(y) + g_3(y)^T \frac{\partial g_2(y)}{\partial y} \\ =&1Ry + (y^T R) \\ \underset{*}{=}&Ry + (y^T R)^T = y^T(R^T + R) = (R + R^T)y \end{aligned}$$

And $*$ **is just symbolic so we make sure we add to column vectors. Remember** $y = x - \mu$ **and we get:**

$$\frac{\partial f_1}{\partial x}(x) = (R + R^T)(x - \mu)$$

Now, assume $R = R^T$ **(i.e. R is symmetric), we get** $R + R^T = 2R$ **and so:**

$$\frac{\partial f_1}{\partial x}(x) = 2R(x - \mu)$$

1.2 Q2. $f_2(\theta) = \sum_{i=1}^{n} (h_i^T\theta - y_i)^2 \overset{?}{=} \|H\theta - y\|^2$

Denote

$$H = \begin{bmatrix} & \vdots & \\ [- & h_i^T & -] \\ & \vdots & \end{bmatrix}$$

so - $[H\theta]_i = h_i^T\theta$ **and** $h_i^T\theta - y_i = [H\theta]_i - y_i = [H\theta - y]_i$ **(★). Now, by definition:**

$$\|H\theta - y\|^2 = (H\theta - y)^T(H\theta - y) = \sum (H\theta - y)_i \cdot (H\theta - y) = \sum (H\theta - y)_i^2$$

Together with ★ we get $\sum_{i=1}^{n} (h_i^T\theta - y_i)^2 = \|H\theta - y\|^2$

1.3 Q3. $f_3(\theta, \lambda) = -c \log \frac{1}{\lambda} - \frac{1}{2}\lambda \sum_{i=1}^{n} (h_i^T\theta - y_i)^2$

Denote $g_1(\theta, \lambda) = \log \frac{1}{\lambda} = -\log \lambda$ **and** $g_2(\theta, \lambda) = \frac{\lambda}{2} \sum_{i=1}^{n} (h_i^T\theta - y_i)^2$**. When deriving each function separately, we get for** g_1**:**

$$\frac{\partial g_1}{\partial \lambda}(\theta, \lambda) = \frac{\partial}{\partial \lambda} \log \frac{1}{\lambda} = -\frac{\partial}{\partial \lambda} \log \lambda = -\frac{1}{\lambda}$$

and

$$\frac{\partial g_1}{\partial \theta}(\theta, \lambda) = \frac{\partial}{\partial \theta} \log \frac{1}{\lambda} = -\frac{\partial}{\partial \theta} \log \lambda = 0$$

As for g_2**:**

$$\frac{\partial g_2}{\partial \lambda}(\theta, \lambda) = \frac{\partial}{\partial \lambda} \frac{\lambda}{2} \sum_{i=1}^{n} (h_i^T\theta - y_i)^2 = \frac{1}{2} \sum_{i=1}^{n} (h_i^T\theta - y_i)^2 \frac{\partial}{\partial \lambda}\lambda = \frac{1}{2} \sum_{i=1}^{n} (h_i^T\theta - y_i)^2$$

Lastly, note that - $\frac{\partial g_2}{\partial \theta}(\theta, \lambda) = \frac{\lambda}{2} \frac{\partial f_2}{\partial \theta}(\theta)$ **(**f_2 **from Q2). We saw in the recitation that** $\frac{\partial}{\partial y} g(y) = \frac{\partial}{\partial y} \|y\|^2 = 2y$**. Denote** $f(x) = Hx - y$ **and we get:** $f_2(x) = g(f(x))$**. By the chain role -** $\frac{\partial f_2}{\partial x}(x) = \frac{\partial f}{\partial x}(x) \frac{\partial g}{\partial f(x)}(f(x))$**. Per index:**

$$\frac{\partial (Hx - y)_j}{\partial x_i} = \frac{\partial (\sum_k H_{jk} x_k - y_j)}{\partial x_i} = \frac{\partial (\sum_k H_{jk} x_k)}{\partial x_i} - \frac{\partial y_j}{\partial x_i} = H_{ji}$$

Hence - $\frac{\partial f}{\partial x}(x) = \frac{\partial (Hx - y)}{\partial x} = H^T$ **and thus -**

$$\frac{\partial f_2}{\partial x}(x) = 2H^T(Hx - y)$$

And -

$$\frac{\partial g_2}{\partial \theta}(\theta, \lambda) = \frac{\lambda}{2} \frac{\partial f_2}{\partial \theta}(\theta) = \frac{\lambda}{2} 2H^T(H\theta - y) = \lambda H^T(H\theta - y)$$

To some it all up:

$$f_3(\theta,\lambda) = -cg_1(\theta,\lambda) - g_2(\theta,\lambda)$$

and so:

$$\frac{\partial f_3}{\partial \theta} = -c\frac{\partial g_1}{\partial \theta}(\theta,\lambda) - \frac{\partial g_2}{\partial \theta}(\theta,\lambda) = 0 - \lambda H^T(H\theta - y) = -\lambda H^T(H\theta - y)$$

and:

$$\frac{\partial f_3}{\partial \lambda} = -c\frac{\partial g_1}{\partial \lambda}(\theta,\lambda) - \frac{\partial g_2}{\partial \lambda}(\theta,\lambda) = \frac{c}{\lambda} - \frac{1}{2}\sum_{i=1}^{n}(h_i^T\theta - y_i)^2 = \frac{c}{\lambda} - \frac{1}{2}\left\|H\theta - y\right\|^2$$

A) $\hat{\theta}$ which maximizes f_3 holds $\frac{\partial f_3}{\partial \theta}(\hat{\theta},\lambda) = 0$ (as f_3 is concave):

$$\begin{aligned} &0 = -\lambda H^T(H\hat{\theta} - y) \underset{0<\lambda}{\Longrightarrow} 0 = H^T(H\hat{\theta} - y) \\ \Longleftrightarrow\ & H^T y = H^T H\hat{\theta} \\ \Longleftrightarrow\ & \hat{\theta} = \left(H^T H\right)^{-1} H^T y \end{aligned}$$

And we can see it doesn't depend on λ.

B) $\hat{\lambda}$ which maximizes f_3 holds $\frac{\partial f_3}{\partial \theta}(\theta,\hat{\lambda}) = 0$ (as f_3 is concave):

$$\begin{aligned} 0 = \frac{\partial f_3}{\partial \lambda}(\theta,\hat{\lambda}) = \frac{c}{\hat{\lambda}} - \frac{1}{2}\left\|H\theta - y\right\|^2 \Longrightarrow \frac{1}{2}\left\|H\theta - y\right\|^2 = \frac{c}{\hat{\lambda}} \\ \Longrightarrow \hat{\lambda} = \frac{2c}{\left\|H\theta - y\right\|^2} \end{aligned}$$

And here $\hat{\lambda}$ depends on θ.

C) Maximize f_3:

We choose $\hat{\theta} = \left(H^T H\right)^{-1} H^T y$. Plug it in to λ and we get:

$$\hat{\lambda} = \frac{2c}{\left\|H\theta - y\right\|^2} = \frac{2c}{\left\|H\left(H^T H\right)^{-1} H^T y - y\right\|^2}$$

2 Probability

2.1 Q4. Will it rain?

We want P("it will rain" | "machine said it won't rain"). Denote A={"it will rain"}, B={"machine said it won't rain"}. What is P(A|B). According to Bayes' law: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$**. P(B|A)=P("machine said it won't rain" | "it will rain")** $= p_{FN}$**. Furthermore - P(A)=P("it will rain")**$=p_r$**. All we need now is P(B)=P("machine said it won't rain"). By the law of total probability we get** $P(B) = P(B|A)p_r + P(B|!A)(1-p_r) = p_{FN} \cdot p_r + P(B|!A)(1-p_r)$**. So we want - P(B|**$!A$**) = P("machine said it won't rain"| NOT "it will rain") = P("machine said it won't rain"| "it will not rain") i.e. the question is what's the probability to get a True Positive which is** $1 - p_{FP}$**. Thus:**

$$\begin{aligned}
P(B) = p_{FN} \cdot p_r + (1-p_{FP})(1-p_r) = \\
=& p_{FN} \cdot p_r + (1-p_r) - p_{FP}(1-p_r) \\
=& p_{FN}p_r + 1 - p_r - p_{FP} + p_r p_{FP} \\
=& 1 - p_r(1-p_{FN}) + p_{FP}(1-p_r)
\end{aligned}$$

So:

$$P(A|B) = \frac{p_r \cdot p_{FN}}{1 - p_r(1-p_{FN}) + p_{FP}(1-p_r)}$$

2.2 Q5. Uniform random variable on a segment in $\mathbb{R}$.

A) We know $1 = \int\limits_{-\infty}^{\infty} PDF(x)dx$**.**

So, because the p(x) is non zero only in the interval $[m - \frac{d}{2}, m + \frac{d}{2}]$**:**

$$1 = \int\limits_{-\infty}^{m-\frac{d}{2}} 0dx + \int\limits_{m-\frac{d}{2}}^{m+\frac{d}{2}} \frac{1}{c}dx + \int\limits_{m+\frac{d}{2}}^{\infty} 0dx = \int\limits_{m-\frac{d}{2}}^{m+\frac{d}{2}} \frac{1}{c}dx = \frac{1}{c}\int\limits_{m-\frac{d}{2}}^{m+\frac{d}{2}} dx = \frac{1}{c} \cdot x|_{m-\frac{d}{2}}^{m+\frac{d}{2}} = \frac{1}{c}(m + \frac{d}{2} - (m - \frac{d}{2})) = \frac{d}{c}$$

Multiply by c and we get c=d.

B) Mean and Variance

For the mean we use $E[x] = \int\limits_{-\infty}^{\infty} xp(x)dx$**. As above:**

$$\begin{aligned}
E[x] = \int\limits_{m-\frac{d}{2}}^{m+\frac{d}{2}} \frac{x}{c}dx = \frac{1}{c}\int\limits_{m-\frac{d}{2}}^{m+\frac{d}{2}} xdx = \frac{1}{2c} \cdot x^2|_{m-\frac{d}{2}}^{m+\frac{d}{2}} = \frac{1}{2c}((m + \frac{d}{2})^2 - (m - \frac{d}{2})^2) = \\
= \frac{1}{2c}(m^2 + md + \frac{d^2}{4} - (m^2 - md + \frac{d^2}{4})) = \frac{1}{2c}(md + md) = \frac{md}{c} \underset{c=d}{=} m = E[x]
\end{aligned}$$

As for the variance - $var(x) = E[x^2] - E[x]^2$**. By the computation above -** $E[x]^2 = m^2$**.**

According to 4.10 in the recitation $E[x^2] = E[f(x)] = \int p(x)f(x)dx$ for $f(x) = x^2$. We compute:

$$E[f(x)] = \int_{-\infty}^{m-\frac{d}{2}} 0 \cdot x^2 dx + \int_{m-\frac{d}{2}}^{m+\frac{d}{2}} \frac{1}{c} x^2 dx + \int_{m+\frac{d}{2}}^{\infty} 0 \cdot x^2 dx = \frac{1}{c} \int_{m-\frac{d}{2}}^{m+\frac{d}{2}} x^2 dx =$$

$$\underset{c=d}{=} \frac{1}{3d}((m+\frac{d}{2})^3 - (m-\frac{d}{2})^3) = \frac{d^3}{12d} + \frac{dm^2}{d} = \frac{d^2}{12} + m^2 = E[x^2]$$

Finally - $var(x) = E[x^2] - E[x]^2 = \frac{d^3}{12d} + m^2 - m^2 = \frac{d^3}{12d} = \frac{d^2}{12}$**. To sum up:**

$$E[x] = m$$

$$var[x] = \frac{d^2}{12}$$

C) $y = x + \delta$

According to the "Change of variable" rule from the recitation (4.24) we know $p_y(y) = p_x(f^{-1}(y)) \cdot det(J_y(f^{-1}(y)))$**. As** $f^{-1}(y) = y - \delta$ **and we're talking about 1D real functions -** $det(J_y(f^{-1}(y))) = \frac{\partial f^{-1}}{\partial y}(y) = 1$ **and so**

$$p_y(y) = p_x(f^{-1}(y)) = \begin{cases} \frac{1}{d} & m - \frac{d}{2} \le f^{-1}(y) \le m + \frac{d}{2} \\ 0 & else \end{cases} = \begin{cases} \frac{1}{d} & m - \frac{d}{2} + \delta \le y \le m + \frac{d}{2} + \delta \\ 0 & else \end{cases}$$

Now note that if we take $\hat{m} = m + \delta$ **and**

$$\hat{p}_z(z) = \begin{cases} \frac{1}{d} & m - \frac{d}{2} \le z \le m + \frac{d}{2} \\ 0 & else \end{cases}$$

We're back to the conditions of our original question, meaning the mean and variance of a uniform random variable in a real segment are invariant to shifts of the segment. I.e. $E(y) = m^2$ **and** $var(y) = \frac{d^2}{12}$**.**

2.3 Q6. $cov(x + y) = cov[x] + cov[y]$

Let x, y **be independent continuous random vectors. For each indexes** i, j **we get:**

$$E[xy^T]_{ij} = E[x_i y_j] = \int \int x_i y_j p(x_i, y_j) dx_i dy_j =$$

$$= \int \int x y_j p(x_i) p(y_j) dx_i dy_j = \int x_i p(x_i) \left[\int y_j p(y_j) dy_j \right] dx_i =$$

$$= \left[\int y_j p(y_j) dy_j \right] \int x_i p(x_i) dx_i = \left[\int y_j p(y_j) dy_j \right] \left[\int x_i p(x_i) dx_i \right] =$$

$$= E[x_i] E[y_j]$$

Hence - $E[xy^T] = E[x]E[y^T]$**. Now, denote** $z = x + y$ **(note that for the summation defined we must have** $dim(x) = dim(y)$ **), So**

$$(x+y)(x+y)^T = \begin{bmatrix} [x+y]_1[x+y]_1 & ... & [x+y]_1[x+y]_n \\ & \vdots & \\ [x+y]_n[x+y]_1 & ... & [x+y]_n[x+y]_n \end{bmatrix}$$
$$= \begin{bmatrix} (x_1+y_1)(x_1+y_1) & ... & (x_1+y_1)(x_m+y_m) \\ & \vdots & \\ (x_n+y_n)(x_1+y_1) & ... & (x_n+y_n)(x_n+y_n) \end{bmatrix}$$
$$= \begin{bmatrix} (x_1x_1+y_1x_1+y_1x_1+y_1y_1) & ... & (x_1x_n+y_1x_n+y_1x_n+y_1y_n) \\ & \vdots & \\ (x_nx_1+y_nx_1+x_ny_1+y_1y_n) & ... & (x_nx_n+y_nx_n+y_nx_n+y_ny_n) \end{bmatrix}$$
$$= \begin{bmatrix} x_1x_1 & ... & x_1x_n \\ & \vdots & \\ x_nx_n & ... & x_nx_n \end{bmatrix} + \begin{bmatrix} y_1y_1 & ... & y_1y_n \\ & \vdots & \\ y_ny_n & ... & y_ny_n \end{bmatrix}$$
$$+ \begin{bmatrix} x_1y_1 & ... & x_1y_n \\ & \vdots & \\ x_ny_n & ... & x_ny_n \end{bmatrix} + \begin{bmatrix} y_1x_1 & ... & x_1y_n \\ & \vdots & \\ x_ny_n & ... & x_ny_n \end{bmatrix}$$
$$= xx^T + yy^T + 2xy^T$$

Now:

$$cov(z) = cov(z,z) = E[zz^T] - E[z]E[z]^T =_1$$
$$E[xx^T + 2xy^T + yy^T] - E[x+y]^2 =_2 E[xx^T] + 2E[xy^T] + E[yy^T] - (E[x]+E[y])^2 = \blacklozenge$$

Where 1 is due to the fact computed above that $(x+y)(x+y)^T = xx^T + yy^T + 2xy^T$ **and 2 is due to E's linearity. Remember that** $E[xy^T] = E[x]E[y^T]$ **and** $(E[x]+E[y])^2 = E[x]^2 + 2E[x]E[y] + E[y]^2$ **we get:**

$$\blacklozenge = E[xx^T] + 2E[xy^T] + E[yy^T] - E[x]^2 - 2E[x]E[y] - E[y]^2$$
$$= E[xx^T] - E[x]^2 + E[yy^T] - E[y]^2$$
$$= \left[E[xx^T] - E[x]E[x]^T\right] + \left[E[yy^T] - E[y]E[y^T]\right]$$
$$= cov[x] + cov[y]$$

And thus $cov(x+y) = cov[x] + cov[y]\blacksquare$

2.4 Q7. $cov\left[Hx + \eta\right] = Hcov[x]H^T + cov(\eta)$

First let's look at $E[Hx]$**. As**

$$Hx = \begin{bmatrix} ... & H_i^T x & ... \end{bmatrix}^T = \begin{bmatrix} ... & \sum H_{ik}x_k & ... \end{bmatrix}^T = \sum_{k=1}^{n} \begin{bmatrix} H_{1k}x_k & ... & H_{qk}x_k \end{bmatrix}^T$$

and by the definition $E[Hx]_i = E[[Hx]_i]$ **we get:**

$$E[[Hx]_i] = E\left[\sum H_{ik}x_k\right] = \sum H_{ik}E\left[x_k\right]$$

where the last transition is due to E's linearity. So we can write:

$$E[Hx] = \begin{bmatrix} E\left[\sum H_{1k}x_k\right] \\ \vdots \\ E\left[\sum H_{qk}x_k\right] \end{bmatrix} = \begin{bmatrix} \sum H_{1k}E\left[x_k\right] \\ \vdots \\ \sum H_{qk}E\left[x_k\right] \end{bmatrix} = H \begin{bmatrix} E\left[x_1\right] \\ \vdots \\ E\left[x_n\right] \end{bmatrix} = H \cdot E[x]$$

Now, using Q6 we now $cov(Hx + \eta) = cov(Hx) + cov(\eta)$**. From definition:**

$$cov(Hx) = E\left[(Hx - E[Hx])(Hx - E[Hx])^T\right] =$$
$$E\left[(Hx - HE[x])(Hx - HE[x])^T\right] = E\left[(H(x - E[x]))\left((x - E[x])^T\right)H^T\right] =$$
$$HE\left[((x - E[x]))\left((x - E[x])^T\right)\right]H^T = Hcov[x]H^T$$

Finally:

$$cov(Hx + \eta) = Hcov[x]H^T + cov(\eta)$$