Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2020

**Homework 0**

**Issued:** Saturday 12th September, 2020 **Due:** Thursday 17th September, 2020

**Tips**: It is not a formal homework and will not be graded. The primary goal is to help you remember those basic mathematics you have learnt before.

Calculus & Linear Algebra

0.1. (Inner product) If $\boldsymbol{x} \in \mathbb{R}^n$ is orthogonal to $\boldsymbol{y} \in \mathbb{R}^n$, please show that

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2$$

**Solution:**

$$\begin{aligned}
\|\boldsymbol{x}+\boldsymbol{y}\|^2 &= (\boldsymbol{x}+\boldsymbol{y})^{\mathrm{T}}(\boldsymbol{x}+\boldsymbol{y}) \\
&= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} \\
&= \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2
\end{aligned}$$

0.2. (Orthogonal) Please show that $\|\mathbf{Q}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$.

**Solution:**

$$\begin{aligned}
\|\mathbf{Q}\boldsymbol{x}\|_2 &= (\mathbf{Q}\boldsymbol{x})^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} \\
&= \boldsymbol{x}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x} \\
&= \|\boldsymbol{x}\|_2
\end{aligned}$$

0.3. (Trace) For any matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n\times n}$, please show that

(a) $\operatorname{trace}(\mathbf{AB}) = \operatorname{trace}(\mathbf{BA})$

(b) $\operatorname{trace}(\mathbf{ABC}) = \operatorname{trace}(\mathbf{CAB}) = \operatorname{trace}(\mathbf{BCA})$

(c) $\nabla_{\mathbf{A}} \operatorname{tr}(\mathbf{AB}) = \mathbf{B}^{\mathrm{T}}$.

**Solution:**

(a) From the definition of trace, we know

$$\begin{aligned}
\operatorname{trace}(\mathbf{AB}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n} b_{ji}a_{ij} \\
&= \operatorname{trace}(\mathbf{BA})
\end{aligned}$$

(b) Treat $\mathbf{AB}$ as a new matrix, then we have: $\text{trace}((\mathbf{AB})\mathbf{C}) = \text{trace}(\mathbf{C}(\mathbf{AB}))$

(c) We know $\text{trace}(\mathbf{AB}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji}$, for an arbitray element $a_{kl}$, we have

$$\frac{\partial\, \text{tr}(\mathbf{AB})}{\partial a_{kl}} = b_{lk}$$

Hence we have

$$\nabla_{\mathbf{A}} \text{tr}(\mathbf{AB}) = \begin{bmatrix} b_{11} & b_{21} & \cdots & b_{n1} \\ b_{12} & b_{22} & \cdots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{nn} \end{bmatrix} = \mathbf{B}^{\mathrm{T}}$$

0.4. (Eigenthings) Let $\boldsymbol{x}$ be an eigenvector of a matrix $\mathbf{A}$ with corresponding eigenvalue $\lambda$, then

(a) Show that for any $\gamma \in \mathbb{R}$, the $\boldsymbol{x}$ is an eigenvector of $\mathbf{A} + \gamma I$ with eigenvalue $\lambda + \gamma$.

(b) If $\mathbf{A}$ is invertible, then $\boldsymbol{x}$ is an eigenvector of $\mathbf{A}^{-1}$ with eigenvalue $\lambda^{-1}$.

(c) $\mathbf{A}^k\boldsymbol{x} = \lambda^k\boldsymbol{x}$ for any $k \in \mathbb{Z}$ ( $\mathbf{A}^0 = I$ by definition)

**Solution:**

(a) We have

$$(\mathbf{A} + \gamma I)\boldsymbol{x} = \mathbf{A}\boldsymbol{x} + \gamma\boldsymbol{x} = (\lambda + \gamma)\boldsymbol{x}$$

(b) Suppose $\mathbf{A}$ is invertible, then

$$\boldsymbol{x} = \mathbf{A}^{-1}\mathbf{A}\boldsymbol{x} = \mathbf{A}^{-1}(\lambda\boldsymbol{x}) = \lambda\mathbf{A}^{-1}\boldsymbol{x}$$

such that we have $\mathbf{A}^{-1}\boldsymbol{x} = \frac{1}{\lambda}\boldsymbol{x}$.

(c) The case $k > 0$ follows immediately by induction on $k$, as

$$\begin{aligned} \mathbf{A}\boldsymbol{x} &= \lambda\boldsymbol{x} \\ \mathbf{A}^2\boldsymbol{x} &= \mathbf{A} \cdot \mathbf{A}\boldsymbol{x} = \lambda^2\boldsymbol{x} \\ \mathbf{A}^3\boldsymbol{x} &= \mathbf{A} \cdot \mathbf{A}^2\boldsymbol{x} = \lambda^3\boldsymbol{x} \\ &\cdots \end{aligned}$$

0.5. (Chain rule) $x \in \mathbb{R}$ is a scalar, we have

$$\begin{aligned} y &= ax + b \\ z &= \frac{1}{1 + e^{-y}} \end{aligned}$$

Please give the $\frac{\partial z}{\partial x}$.

**Solution:** According to the chain rule, we have

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \times \frac{\partial y}{\partial x} = z(1-z) \times a$$

0.6. (Matrix derivative) $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{n\times n}$. We have $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$$

Please give the $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$.

**Solution:** The standard solution is, first, we give the differential of $f(\boldsymbol{x})$:

$$\begin{aligned} df(\boldsymbol{x}) &= \sum_{i=1}^{n} \frac{\partial f(\boldsymbol{x})}{\partial x_i} dx_i \\ &= \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}^{\mathrm{T}} d\boldsymbol{x} \\ &= \mathrm{tr}\left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}^{\mathrm{T}} d\boldsymbol{x}\right) \end{aligned}$$

Here we use the trace trick, that is, for a scalar $a$ we have $\mathrm{tr}(a) = a$. Then, for the function above we derive its differential

$$\begin{aligned} df(\boldsymbol{x}) &= d\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}}\mathbf{A}d\boldsymbol{x} + \boldsymbol{w}^{\mathrm{T}}d\boldsymbol{x} \\ &= \mathrm{tr}(d\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}) + \mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}\mathbf{A}d\boldsymbol{x} + \boldsymbol{w}^{\mathrm{T}}d\boldsymbol{x}) \\ &= \mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}d\boldsymbol{x}) + \mathrm{tr}(\boldsymbol{x}^{\mathrm{T}}\mathbf{A}d\boldsymbol{x} + \boldsymbol{w}^{\mathrm{T}}d\boldsymbol{x}) \\ &= \mathrm{tr}\big((\boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{x}^{\mathrm{T}}\mathbf{A} + \boldsymbol{w}^{\mathrm{T}})d\boldsymbol{x}\big) \end{aligned}$$

Refer to the above two equations, we have

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}^{\mathrm{T}} d\boldsymbol{x} = (\boldsymbol{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} + \boldsymbol{x}^{\mathrm{T}}\mathbf{A} + \boldsymbol{w}^{\mathrm{T}})d\boldsymbol{x}$$

which means

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{A}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{w}$$

Or simply, you can remember the result for convenience

$$\begin{aligned} \frac{\partial \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}} &= (\mathbf{A}^{\mathrm{T}} + \mathbf{A})\boldsymbol{x} \\ \frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} &= \boldsymbol{w} \end{aligned}$$

Probability Theory Part

0.7. (Conditional Probability) Explain that $\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] = \mathbb{E}[X]$

**Solution:**

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[X|Y\right]\right] &= \mathbb{E}\left[g(Y)\right] \\ &= \sum_{y\in\mathcal{Y}} p\left(Y=y\right)\cdot\left[\sum_{x\in\mathcal{X}} x\cdot p\left(X=x|Y=y\right)\right] \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} x\cdot p\left(X=x, Y=y\right) \\ &= \mathbb{E}\left[X\right]\end{aligned}$$

0.8. (Bayes) A city has a 50% chance to rain everyday and the weather report has a 90% chance to correctly forecast.
You will take an umbrella when the report says it will rain and you have a 50% chance to take an umbrella when the report says it will not rain.
Compute

(a) the probability of raining when you don't take an umbralla;

(b) the probability of not rainning when you take an umbrella.

**Solution:** Let's evaluate the question. $\overline{A}$ denotes the opposite events of $A$.
Let A be the event **Rain**.

$$p\left(A\right) = p\left(\overline{A}\right) = 0.5$$

Let B be the event **Forecasting Rain**.

$$p\left(B|A\right) = p\left(\overline{B}|\overline{A}\right) = 0.9$$

Let C be the event **Taking Umbrella**.

$$p\left(C|B\right) = 1$$

$$p\left(C|\overline{B}\right) = 0.5$$

OK, now let's come to the questions.

(a) the probability of raining when you don't take an umbralla $= p\left(A|\overline{C}\right)$

$$p\left(A|\overline{C}\right) = \frac{p\left(A\right)p\left(\overline{C}|A\right)}{p\left(A\right)p\left(\overline{C}|A\right) + p\left(\overline{A}\right)p\left(\overline{C}|\overline{A}\right)}$$

$$p\left(\overline{C}|A\right) = p\left(\overline{C}|AB\right)p\left(B|A\right) + p\left(\overline{C}|A\overline{B}\right)p\left(\overline{B}|A\right) = 0*0.9 + 0.5*0.1 = 0.05$$

$$p\left(\overline{C}|\overline{A}\right) = p\left(\overline{C}|\overline{A}B\right)p\left(B|\overline{A}\right) + p\left(\overline{C}|\overline{AB}\right)p\left(\overline{B}|\overline{A}\right) = 0*0.1 + 0.5*0.9 = 0.45$$

Here, we use that

$$p\left(\overline{C}|AB\right) = p\left(\overline{C}|B\right)$$

$$p\left(A|\overline{C}\right) = \frac{0.5 * 0.05}{0.5 * 0.05 + 0.5 * 0.45} = 0.1$$

(b) the probability of not rainning when you take an umbrella $= p\left(\overline{A}|C\right)$
The deduction is the same, so let me omit some steps.

$$p\left(\overline{A}|C\right) = \frac{0.5 * 0.55}{0.5 * 0.55 + 0.5 * 0.95} = \frac{11}{30}$$

0.9. (Joint Distribution) Random Variables $X$ and $Y$ have a joint distribution with joint probability density function

$$f(x, y) = \begin{cases} Ce^{-(2x+y)} & x > 0, y > 0 \\ 0 & ow. \end{cases}$$

Please find $C$ by

$$\int_0^\infty \int_0^\infty f(x, y)\mathrm{d}x\mathrm{d}y = 1$$

**Solution:**

$$C \int_0^\infty \int_0^\infty e^{-(2x+y)}\mathrm{d}x\mathrm{d}y = C \cdot \frac{1}{2} \cdot 1 = 1$$

$$C = 2$$

0.10. (Covariance) Now we have a joint pdf

$$f(x, y) = \begin{cases} 4xy & 0 < x < 1, 0 < y < 1 \\ 0 & ow. \end{cases}$$

Please show that the covariance of $X$ and $Y$ is 0.

**Solution:**

$$\mathbb{E}[X] = \int_0^1 \int_0^1 x \cdot 4xy\mathrm{d}x\mathrm{d}y = \frac{2}{3}$$

$$\mathbb{E}[Y] = \int_0^1 \int_0^1 y \cdot 4xy\mathrm{d}x\mathrm{d}y = \frac{2}{3}$$

$$\mathbb{E}[XY] = \int_0^1 \int_0^1 xy \cdot 4xy\mathrm{d}x\mathrm{d}y = \frac{4}{9}$$

Thus,

$$\mathrm{Cov}[X, Y] = 0$$

Of course, if you are clever enough, you will see that they are independent.

0.11. (Uncorrelated and independent RVs) We have a uniform distribution of $X$ and $Y$ on a disk. The pdf is

$$f(x,y) = \frac{1}{\pi} \qquad x^2 + y^2 \leq 1$$

Please show that $X$ and $Y$ are uncorrelated but not independent.

**Solution:**

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\mathrm{d}y = \frac{2\sqrt{1-x^2}}{\pi} \qquad -1 < x < 1$$

Similarly,

$$f_Y(y) = \frac{2\sqrt{1-y^2}}{\pi} \qquad -1 < y < 1$$

Obviusly,

$$f(x,y) \neq f_X(x)f_Y(y) \Rightarrow \text{Not Independent}$$

$$\mathbb{E}[X] = \int_{-1}^{1} x\frac{2\sqrt{1-x^2}}{\pi}\mathrm{d}x = 0$$

$$\mathbb{E}[Y] = \int_{-1}^{1} y\frac{2\sqrt{1-y^2}}{\pi}\mathrm{d}y = 0$$

$$\mathbb{E}[XY] = \int_{x^2+y^2\leq 1} \frac{xy}{\pi}\mathrm{d}x\mathrm{d}y = 0$$

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \Rightarrow \text{Uncorrelated}$$

0.12. (Guassian Distribution) There is a famous integral here

$$\int_{-\infty}^{\infty} e^{-x^2}\mathrm{d}x = \sqrt{\pi}$$

It is called Guassian Integral. Based on it, please find some results of the Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad -\infty < x < \infty$$

(a) Prove it is a pdf ($\sigma > 0$)

(b) Compute the expectation and variance

**Solution:**

(a)

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}}\exp\left(-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right)\mathrm{d}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) = 1$$

(b)

$$\begin{aligned}
\int_{-\infty}^{\infty} xf(x)\mathrm{d}x &= \int_{-\infty}^{\infty} (x-\mu+\mu)\,\frac{1}{\sqrt{\pi}}\exp\left(-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right)\mathrm{d}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\\
&= \mu + \int_{-\infty}^{\infty} \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right)\mathrm{d}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\\
&= \mu
\end{aligned}$$

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2 e^{-x^2}\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{1}{2}xe^{-x^2}\mathrm{d}x^2\\
&= -\int_{-\infty}^{\infty} \frac{1}{2}x\mathrm{d}e^{-x^2}\\
&= \int_{-\infty}^{\infty} \frac{1}{2}e^{-x^2}\mathrm{d}x - \frac{1}{2}xe^{-x^2}\Big|_{-\infty}^{+\infty}\\
&= \frac{\sqrt{\pi}}{2}
\end{aligned}$$

$$\begin{aligned}
\int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x &= \int_{-\infty}^{\infty} \frac{(x-\mu)^2+2\mu x-\mu^2}{\sqrt{\pi}}\exp\left(-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right)\mathrm{d}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\\
&= \frac{\sqrt{\pi}}{2}\cdot\frac{2\sigma^2}{\sqrt{\pi}} + 2\mu\cdot\mu - \mu\cdot\mu\\
&= \sigma^2+\mu^2
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[X\right] = \mu$$

$$\mathrm{Var}\left[X\right] = \sigma^2+\mu^2-\mu^2 = \sigma^2$$