Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2019

Problem Set 0

Issued: Thursday 12th September, 2019 **Due:** Monday 16th September, 2019

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.

Probability Theory Part

0.1. (Conditional Probability) For discrete random variables, the conditional probability can be derived by Product Rule.

$$p(X,Y) = p(Y|X)\,p(X)$$

We can define the conditional expectation as

$$\mathbb{E}[Y|X=x] \triangleq \sum_{y\in\mathcal{Y}} y\cdot p(Y=y|X=x)$$

Explain that

(a) $\mathbb{E}[X|X] = X$

(b) $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}(X)$

(c) $\mathbb{E}[g(X)h(Y)|Y] = h(Y)\mathbb{E}[g(X)|Y]$ $\quad g(x)$ and $h(y)$ are bounded functions

Solution:

(a)

$$p(X=x|X=x) = 1$$
$$\mathbb{E}[X|X] = X\cdot p(X|X) = X$$

(b)

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

(c)

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

0.2. (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**.

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

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

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

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

$$p(C|B) = 1$$

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

OK, now let's come to the questions.

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

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

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

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

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.

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

0.3. (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:

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

$$C = 2$$

0.4. (Covariance) For two random variables X and Y, the covariance is defined by

$$\mathrm{Cov}\left[X,Y\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

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:

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

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

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

Thus,

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

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

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

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

When the covariance of X and Y is 0, we call them uncorrelated variables.
For continuous random variables, when the joint pdf can be written as the product of two RVs' pdf

$$f\left(x,y\right) = f_X\left(x\right)f_Y\left(y\right),$$

we call them independent.
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,

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

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

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

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

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

0.6. (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)

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

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

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

Therefore,

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

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

Calculus & Linear Algebra

0.7. (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

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

0.8. (Orthogonal) The $Q \in \mathbb{R}^{n \times n}$ is said to be **orthogonal** if its columns are pairwise orthogonal, which implies that

$$
QQ^\top = Q^\top Q = I
$$

Please show that $\|Q\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$.

Solution:

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

0.9. (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:

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

0.10. (Determinant) If a matrix $A \in \mathbb{R}^{n\times n}$ is invertible, the A^* is said to be the **adjoint matrix** of A where $A^{-1} = \frac{A^*}{\det A}$. Please prove that if $\det A = 0$, then we have $\det A^* = 0$.

Solution: Here we use proof by contradiction, that is, assume that when $\det A$ $= 0$ and the $\det A^* \neq 0$, such that the A^* is invertible. We know

$$A^*A = \det A \cdot I$$

hence $A^* \cdot \frac{A}{\det A} = I$, which means $\frac{A}{\det A}$ is the invertible matrix of A^*, and further we get that A is invertible. It contradicts to the assumption that $\det A = 0$, so the $\det A^* = 0$ must hold.

0.11. (Invertibility) Given a matrix $A \in \mathbb{R}^{n\times n}$ and $A^3 = 4I$, please give the invertible matrix of $A - I$.

Solution:

$$\begin{aligned}A^3 - 4I &= 0 \\ A^3 - I &= 3I \\ A^3 - I^3 &= 3I \\ (A-I)(A^2+A+I) &= 3I \\ (A-I)\cdot\frac{1}{3}(A^2+A+I) &= I\end{aligned}$$

Therefore we know the inversion of $A - I$ is $\frac{1}{3}(A^2+A+I)$.

0.12. (Trace) The **trace** of a matrix $A \in \mathbb{R}^{n \times n}$ is defined as sum of diagonal elements of A:

$$\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

(a) Show that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

(b) Show that $\nabla_A \operatorname{tr}(AB) = B^\top$.

Solution:

(a) From the definition of trace, we know

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

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

$$\frac{\partial \operatorname{tr}(AB)}{\partial a_{kl}} = b_{lk}$$

Hence we have

$$\nabla_A \operatorname{tr}(AB) = \begin{bmatrix} b_{11} & b_{21} & \cdots & b_{n1} \\ b_{12} & b_{22} & \cdots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{nn} \end{bmatrix} = B^\top$$

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

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

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

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

Solution:

(a) We have

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

(b) Suppose A is invertible, then

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

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

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

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

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

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{w}^\top\boldsymbol{x}$$

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

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

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

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

$$
\begin{aligned}
df(\boldsymbol{x}) &= d\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top A d\boldsymbol{x} + \boldsymbol{w}^\top d\boldsymbol{x} \\
&= \operatorname{tr}(d\boldsymbol{x}^\top A\boldsymbol{x}) + \operatorname{tr}(\boldsymbol{x}^\top A d\boldsymbol{x} + \boldsymbol{w}^\top d\boldsymbol{x}) \\
&= \operatorname{tr}(\boldsymbol{x}^\top A^\top d\boldsymbol{x}) + \operatorname{tr}(\boldsymbol{x}^\top A d\boldsymbol{x} + \boldsymbol{w}^\top d\boldsymbol{x}) \\
&= \operatorname{tr}\left((\boldsymbol{x}^\top A^\top + \boldsymbol{x}^\top A + \boldsymbol{w}^\top) d\boldsymbol{x}\right)
\end{aligned}
$$

Refer to the above two equations, we have

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

which means

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

Or simply, you can remember the result for convenience

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