Machine Learning Exercise Sheet 1

Math Refresher

Group_369

Fan Xue – fan98.xue@tum.de

Xing Zhou – xing.zhou@tum.de

Jianzhe Liu – jianzhe.liu@tum.de

October 27, 2021

Problem 1

According to the matrices multiplication rule and the dimension of function f's return value, which is $\mathbb{R}$:

$$\boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{B} \in \mathbb{R}^{1\times M}, \boldsymbol{c} \in \mathbb{R}^{N\times P}, \boldsymbol{D} \in \mathbb{R}^{Q\times 1}, \boldsymbol{E} \in \mathbb{R}^{N\times N}, \boldsymbol{F} \in \mathbb{R}^{1\times 1}$$

Problem 2

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j M_{ij} = \sum_{i=1}^{N} x_i \left(\sum_{j=1}^{N} M_{ij} x_j\right) = \sum_{i=1}^{N} x_i \left(\boldsymbol{M}\boldsymbol{x}\right)_i = \boldsymbol{x}^T \boldsymbol{M}\boldsymbol{x}$$

Problem 3

a) If $M < N$ or $\text{rank}(\boldsymbol{A}) < N$, then the solution $\boldsymbol{x}$ is not always unique. If $M > N$ or $\text{rank}(\boldsymbol{A}) < M$, there would be no solution $\boldsymbol{x}$ for every $\boldsymbol{b} \in \mathbb{R}$.
If and only if $M = N$ and $\boldsymbol{A}$ is full rank, i.e. $\boldsymbol{A}$ is invertible, there is a unique solution $\boldsymbol{x}$ for any choice of $\boldsymbol{b}$.

b) No, because $\boldsymbol{A}$ has an eigenvalue equal to 0, that means $\boldsymbol{A}$ does not have full rank.

Problem 4

Since $\boldsymbol{BA} = \boldsymbol{AB} = \boldsymbol{I}$, $\boldsymbol{A}$ is invertible. The determinant of $\boldsymbol{A}$ is

$$\det\boldsymbol{A} = \prod_{i=1}^{N} \lambda_i \neq 0$$

, which indicates none of the eigenvalues of $\boldsymbol{A}$ is 0.

Problem 5

1. $\lambda_i \geqslant 0 \Rightarrow \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geqslant 0$

A is symmetric, there exists a orthonormal matrix $\boldsymbol{Q}$ such that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$. Λ is a diagonal matrix with the eigenvalues of A in its diagonal.

For any $\boldsymbol{x} \in \mathbb{R}^N$, we have

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{x} = (\boldsymbol{Q}^T\boldsymbol{x})^T\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{x}$$

Notice that $\boldsymbol{Q}^T\boldsymbol{x} \in \mathbb{R}^N$, let $\boldsymbol{y} = \boldsymbol{Q}^T\boldsymbol{x}$, we have

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} = \begin{bmatrix} y1 & \cdots & yn \end{bmatrix} \begin{bmatrix} \lambda 1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} \lambda_i y_i^2$$

Because all y_i^2 and $\lambda_i \geqslant 0$, we have $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geqslant 0$.

2. $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geqslant 0 \Rightarrow \lambda_i \geqslant 0$

Assume that $\boldsymbol{x_i}$ and λ_i are one of the eigenvectors and the corresponding eigenvalue of $\boldsymbol{A}$. We have

$$\boldsymbol{A}\boldsymbol{x_i} = \lambda_i\boldsymbol{x_i}$$

then

$$\boldsymbol{x_i}^T\boldsymbol{A}\boldsymbol{x_i} = \boldsymbol{x_i}^T\boldsymbol{\lambda_i}\boldsymbol{x_i} = \lambda_i\|\boldsymbol{x_i}\|^2 \geqslant 0$$

Because$\|\boldsymbol{x_i}\|^2 \geqslant 0$, we have $\lambda_i \geqslant 0$.

Problem 6

For any $\boldsymbol{A} \in \mathbb{R}^{M\times N}$, we have

$$\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}\|^2 \geqslant 0$$

which means $\boldsymbol{B}$ is positive semi-definite.

Problem 7

a) We observe the first and second derivative of $f(x)$, we have

$$f'(x) = ax + b$$
$$f''(x) = a$$

(i) a unique solution means there exists a $x \in \mathbb{R}$, such that $f'(x) = 0$ and $f''(x) > 0$, which is

$$a > 0$$

(ii) infinitely many solutions means there exist infinitely many $x \in \mathbb{R}$, such that $f'(x) = 0$, which is

$$a = b = 0$$

(iii) no solution means $f'(x) = 0$ have no solution or $f'(x) = 0$ have solution but at that point $f''(x) < 0$, which is

$$a = 0, b \neq 0 \quad \text{or} \quad a < 0$$

b) let $f'(x) = ax + b = 0$, we have

$$x^* = -\frac{b}{a}$$

Problem 8

a) Consider the term including x_k and x_l factors to take the partial derivative, under the constrain of $\boldsymbol{A} \in \mathbb{S}^N$, i.e. $A_{ij} = A_{ji}$:

$$\begin{aligned}
\frac{\partial g(\boldsymbol{x})}{\partial x_k \partial x_l} &= \frac{\partial}{\partial x_k \partial x_l}\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}x_ix_j + \sum_{i=1}^{N} b_ix_i + c\right) \\
&= \frac{\partial}{\partial x_k}\left(\frac{1}{2}\sum_{i=1}^{N} A_{il}x_i + \frac{1}{2}\sum_{j=1}^{N} A_{lj}x_j + b_l\right) \\
&= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{N} A_{il}x_i + b_l\right) = A_{kl}
\end{aligned}$$

This result indicates the Hessian of $g(\boldsymbol{x})$ is:

$$\nabla^2 g(\boldsymbol{x}) = \boldsymbol{A}$$

To have a unique solution of this optimization problem, $\boldsymbol{A}$ should be positive definite.

b) Since $\boldsymbol{A} \in \mathbb{S}^N$, $\boldsymbol{A}$ can be represented an $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, in which $\boldsymbol{U}$ is the matrix of orthonormal eigenvectors of $\boldsymbol{A}$, $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of $\boldsymbol{A}$.

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{x} = \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y} = \sum_{i=1}^{N} \lambda_i y_i^2$$

where $\boldsymbol{y} = \boldsymbol{U}^T\boldsymbol{x}$ (and since $\boldsymbol{U}$ is full rank, any vector $\boldsymbol{y} \in \mathbb{R}^N$ can be represented in this form).
If there exists a negative eigenvalue of $\boldsymbol{A}$, which we assume $\lambda_k < 0$, then as we take $y_k \to \infty$, $g(\boldsymbol{x})$ would be infinite negative, which means the function has no minimum. Therefore the matrix $\boldsymbol{A}$ is PSD should be well-defined.

c) Since the matrix $\boldsymbol{A}$ is PD, we could minimize the objective function by setting the gradient to zero. To get the gradient of $g(\boldsymbol{x})$ we first consider the term of the partial derivative of x_k:

$$\begin{aligned}\frac{\partial g(\boldsymbol{x})}{\partial x_k} &= \frac{\partial}{\partial x_k}\left(\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}x_i x_j + \sum_{i=1}^{N} b_i x_i + c\right)\\ &= \frac{1}{2}\sum_{i=1}^{N} A_{ik}x_i + \frac{1}{2}\sum_{j=1}^{N} A_{kj}x_j + b_k\\ &= \sum_{i=1}^{N} A_{ik}x_i + b_k\end{aligned}$$

therefore the gradiant of $g(\boldsymbol{x})$ is:

$$\nabla g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$$

By setting $\nabla g(\boldsymbol{x}) = 0$, we can get

$$\boldsymbol{x}^\star = -\boldsymbol{A}^{-1}\boldsymbol{b}.$$

Problem 9

The equation can be **disproved** by following example: let's say we have a coin, and after one toss we define 3 events:

Event A: the coin is back-side. $p(a) = 0.5$

Event B: the coin is front-side. $p(b) = 0.5$

Event C: the coin is front-side. $p(c) = 0.5$

(Here we use an extreme example, where event B and event C are equivalent)

In this case, Event A and Event B (or Event C) are mutually exclusive, which means:

$$p(a|b, c) = p(a|c) = 0$$

But obviously,

$$p(a|b) = 0 \neq p(a) = 0.5$$

Problem 10

The equation can be **disproved** by following example: let's say we have 2 dices, and after rolling we define 3 events:

Event A: the first dice shows 1 or 2. $p(a) = \frac{1}{3}$

Event B: the second dice shows 1 or 3. $p(b) = \frac{1}{3}$

Event C: the result of both dices are 2 and 3. $p(c) = \frac{1}{18}$

In this case, Event A and Event B are independent, which means:

$$p(a|b) = p(a) = \frac{1}{3}$$

But when we see the other side of the equation,
when Event B and Event C are determined, $p(a) = 1$.
when only Event C is determined, $p(a) = \frac{1}{2}$.
That is to say:

$$p(a|b, c) = 1 \neq p(a|c) = \frac{1}{2}$$

Problem 11

When the joint PDF p(a, b, c) of three continuous random variables are given, we can solve the following problems according to the definition of the probability density function and conditional Probability:

1. p(a) can be obtained by integrating b and c

$$p(a) = \iint p(a, b, c)\,\mathrm{d}b\,\mathrm{d}c$$

2. Since a, b and c are independent to each other, so:

$$p(c|a, b) = \frac{p(a, b, c)}{p(a, b)} = \frac{p(a, b, c)}{\int p(a, b, c)\,\mathrm{d}c}$$

3. According to the question 1 and question 2:

$$p(b|c) = \frac{p(b, c)}{p(c)} = \frac{\int p(a, b, c)\,\mathrm{d}a}{\iint p(a, b, c)\,\mathrm{d}a\,\mathrm{d}b}$$

Problem 12

The probability that the person has the disease can be calculated by following steps:
If he has the disease while being tested positive, then:

$$p(\text{disease}) * p(\text{positive with disease}) = 0.1\% * 95\%$$

If he has no disease while being tested positive, then:

$$p(\text{healthy}) * p(\text{positive when healthy}) = 99.9\% * 5\%$$

According to these 2 equations, the probability that he really has the disease is:

$$\frac{0.1\% * 95\%}{0.1\% * 95\% + 99.9\% * 5\%} = 1.87\%$$

(This probabilistic model is also known as a typical example in Bayes' theorem)

Problem 13

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then we have following equations:

$$\mathbb{E}(x) = \mu$$

$$var(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 = \sigma^2$$

Then for $f(x) = ax + bx^2 + c$, we have:

$$\mathbb{E}(f(x)) = \mathbb{E}(ax) + \mathbb{E}(bx^2) + \mathbb{E}(c) = a\mu + b(\mu^2 + \sigma^2) + c$$

Problem 14

a)

$$\mathbb{E}[g(\boldsymbol{x})] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{\mu}$$

b)

$$\mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x})^T] = \mathbb{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T$$

we have $(\boldsymbol{x}\boldsymbol{x}^T)_{ij} = x_i x_j$, so $(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T])_{ij} = \mathbb{E}[x_i x_j]$.
if $i = j$, we have

$$\mathbb{E}[x_i^2] = Var[x_i] + (\mathbb{E}[x_i])^2 = \Sigma_{ii} + \mu_i^2$$

if $i \neq j$, x_i and x_j are independent, we have

$$\mathbb{E}[x_i x_j] = \mathbb{E}[x_i]\mathbb{E}[x_j] = \mu_i \mu_j$$

Combining the two cases, we have

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$$

So

$$\mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{A}^T = \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T$$

c)

$$\mathbb{E}[g(\boldsymbol{x})^T g(\boldsymbol{x})] = \mathbb{E}[(\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\boldsymbol{x}] = \mathbb{E}[\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}]$$

Because $\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}] &= \mathbb{E}[Tr(\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x})] = \mathbb{E}[Tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T)] \\ &= Tr(\mathbb{E}[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{A}^T]) \\ &= Tr(\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T)\end{aligned}$$

d)

$$\begin{aligned}\operatorname{Cov}[g(\boldsymbol{x})] &= \mathbb{E}[g(\boldsymbol{x})g(\boldsymbol{x})^T] - \mathbb{E}[g(\boldsymbol{x})](\mathbb{E}[g(\boldsymbol{x})])^T \\ &= \boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{A}^T - \boldsymbol{A}\boldsymbol{\mu}(\boldsymbol{A}\boldsymbol{\mu})^T \\ &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\end{aligned}$$