Geophysical Data Analysis Quiz #1

05-242628 Shota Mitamura

November 2, 2025

Question #1

For questions (1) & (2), we will denote the probability of receiving a spam e-mail by $P(\text{spam})$, the probability of receiving an e-mail with the word free by $P(\text{free})$, and so on.

(1)

Here, we will try to calculate the probability that an e-mail containing the word "free" is spam, which can be written as:

$$P(\text{spam}|\text{free}) \tag{1}$$

If we apply the Bayes' theorem, (1) can be rewritten as:

$$\frac{P(\text{free}|\text{spam})P(\text{spam})}{P(\text{free})} \tag{2}$$

Then, by using some basic probability manipulations, $P(\text{free})$ can be calculated as such:

$$\begin{aligned} P(\text{free}) &= P((\text{spam} \cap \text{free}) \cup (\text{Non-spam} \cap \text{free})) \\ &= P(\text{spam} \cap \text{free}) + P(\text{Non-spam} \cap \text{free}) \\ &\qquad - P((\text{spam} \cap \text{free}) \cap (\text{Non-spam} \cap \text{free})) \\ &= P(\text{free}|\text{spam})P(\text{spam}) \\ &\qquad + P(\text{free}|\text{Non-spam})P(\text{Non-spam}) \end{aligned} \tag{3}$$

Therefore, by combining equations (2) and (3),we get:

$$P(\text{spam}|\text{free}) = \frac{P(\text{free}|\text{spam})P(\text{spam})}{P(\text{free}|\text{spam})P(\text{spam}) + P(\text{free}|\text{Non-spam})P(\text{Non-spam})} \tag{4}$$

which can then be rewritten as:

$$P(\text{spam}|\text{free}) = \left(1 + \frac{P(\text{free}|\text{Non-spam})}{P(\text{free}|\text{spam})} \frac{P(\text{Non-spam})}{P(\text{spam})} \right)^{-1} \tag{5}$$

If we look at equation (5), one shall see that $P(\text{spam}|\text{free})$ becomes higher as $\frac{P(\text{free}|\text{spam})}{P(\text{free}|\text{Non-spam})}$ and $\frac{P(\text{spam})}{P(\text{Non-spam})}$ becomes higher. This result can be interpreted as follows:

- A higher tendency of "free" appearing in spam e-mails increases the probability that an e-mail containing "free" is spam.
- A higher overall probability of receiving spam e-mails increases the probability that an e-mail containing "free" is spam.

These results show that equation (5) is consistent with reality.

Lastly, if we substitute $P(\text{spam}) = \alpha$, $P(\text{Non-spam}) = 1 - \alpha$, $P(\text{free}|\text{spam}) = 0.12$ and $P(\text{free}|\text{Non-spam}) = 0.02$ as given in the problem statement, we get the answer for question (1) as follows:

$$P(\text{spam}|\text{free}) = \frac{6\alpha}{5\alpha + 1} \tag{6}$$

(2)

Next, we will calculate the probability that an e-mail containing the word "free", "Outstanding fee", "Urgent", and "Password"is spam. This probability can be written as:

$$P\left(\text{spam}|\text{free} \cap \text{Outstanding-fee} \cap \text{Urgent} \cap \text{Password}\right) \tag{7}$$

For readability, we will define the following;

$$\text{(event)} \quad \text{all-words} := \text{free} \cap \text{Outstanding-fee} \cap \text{Urgent} \cap \text{Password} \tag{8}$$

so that (7) can be rewritten as:

$$P(\text{spam}|\text{all-words}) \tag{9}$$

Again, if we apply the Bayes' theorem, (9) can be written as:

$$\frac{P(\text{all-words}|\text{spam})P(\text{spam})}{P(\text{all-words})} \tag{10}$$

Then, for the denominator, we can use the same procedure as in question (1) to get:

$$\begin{aligned} P(\text{all-words}) &= P\left((\text{spam} \cap \text{all-words}) \cup (\text{Non-spam} \cap \text{all-words})\right) \\ &= P(\text{spam} \cap \text{all-words}) + P(\text{Non-spam} \cap \text{all-words}) \\ &\qquad - P\left((\text{spam} \cap \text{all-words}) \cap (\text{Non-spam} \cap \text{all-words})\right) \\ &= P(\text{all-words}|\text{spam})P(\text{spam}) \\ &\qquad + P(\text{all-words}|\text{Non-spam})P(\text{Non-spam}) \end{aligned} \tag{11}$$

Therefore, by combining equations (10) and (11), we get

$$\begin{aligned} &P(\text{spam}|\text{all-words}) \\ &= \frac{P(\text{all-words}|\text{spam})P(\text{spam})}{P(\text{all-words}|\text{spam})P(\text{spam}) + P(\text{all-words}|\text{Non-spam})P(\text{Non-spam})} \end{aligned} \tag{12}$$

which then can be rewritten as:

$$P(\text{spam}|\text{all-words}) = \left(1 + \frac{P(\text{all-words}|\text{Non-spam})}{P(\text{all-words}|\text{spam})} \frac{P(\text{Non-spam})}{P(\text{spam})} \right)^{-1} \tag{13}$$

Since we know that the probability of receiving an e-mail with word "X" is independent from that with word "Y", $P(\text{all-words}|\text{spam})$ can be written as such:

$$\begin{aligned} &P(\text{all-words}|\text{spam}) \\ &\quad = P(\text{free}|\text{spam})P(\text{Outstanding-fee}|\text{spam})P(\text{Urgent}|\text{spam})P(\text{Password}|\text{spam}) \end{aligned} \tag{14}$$

(Note that "spam" can be replaced for "Non-spam" to get the same result.)

Therefore, if we substitute (14) for (13), we get:

$$P(\text{spam}|\text{all-words}) = \left\{ 1 + \left(\prod_{X=\text{spam-words}} \frac{P(\text{X}|\text{Non-spam})}{P(\text{X}|\text{spam})} \right) \frac{P(\text{Non-spam})}{P(\text{spam})} \right\}^{-1} \tag{15}$$

where X substitutes for spam words "free", "Outstanding-fee" and etc. [*1]

[*1] By expanding the product, we will get

$$P(\text{spam}|\text{all-words}) = \left(1 + \frac{P(\text{free}|\text{Non-spam})}{P(\text{free}|\text{spam})} \frac{P(\text{Outstanding-fee}|\text{Non-spam})}{P(\text{Outstanding-fee}|\text{spam})} \frac{P(\text{Urgent}|\text{Non-spam})}{P(\text{Urgent}|\text{spam})} \frac{P(\text{Password}|\text{Non-spam})}{P(\text{Password}|\text{spam})} \frac{P(\text{Non-spam})}{P(\text{spam})} \right)^{-1}$$

If we look at equation (15), we can see that $P(\text{spam}|\text{all-words})$ becomes higher as $\dfrac{P(\text{X}|\text{spam})}{P(\text{X}|\text{Non-spam})}$ becomes higher for any word X. Also, $P(\text{spam}|\text{all-words})$ becomes higher when $\dfrac{P(\text{spam})}{P(\text{Non-spam})}$ becomes higher. These results can be interpreted as follows:

- For any spam word, a higher tendency of such spam words appearing in spam e-mails increases the probability of an e-mail containing spam words actually being a spam.
- A higher overall probability of receiving spam e-mails increases the probability of an e-mail containing spam words actually being a spam.

These results show that equation (15) is consistent with reality.

Lastly, if we substitute various values for equation (15) as given in the problem statement, we get the answer for question (2) as follows:

$$P(\text{spam}|\text{all-words}) = \frac{9 \times 10^3 \alpha}{1 + 9 \times 10^3 \alpha} \tag{16}$$

remark:

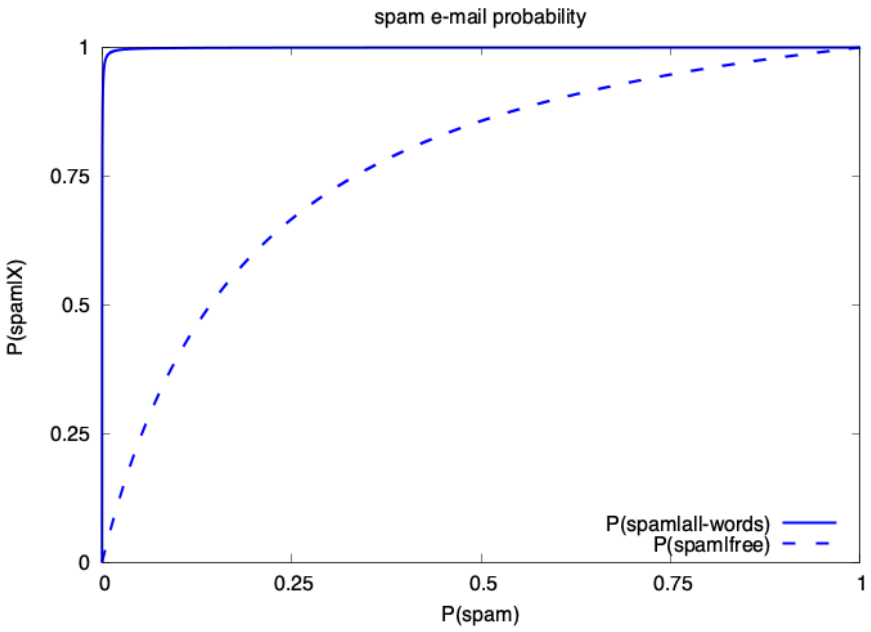

Figure1: **Probability of an e-mail being spam**

Figure 1 shows $P(\text{spam}|\text{free})$ and $P(\text{spam}|\text{all-words})$ as $P(\text{spam}) = \alpha$ varies from 0 to 1. We can see that both probability reaches 1 as $P(\text{spam})$ reaches 1, meaning when there is a 100% chace of receiving a spam e-mail, all e-mails are spam regardless of the words included. Also, the graph

shows that $P(\text{spam}|\text{all-words})$ is higher than $P(\text{spam}|\text{free})$, which indicates that e-mails with multiple spam-words have a higher chance of being a spam than an e-mail with just the word "free".

Question #2

By looking at equation (3.36) from the lecture note, we can compute Σ_m as follows:

$$\begin{aligned} \Sigma_m &= E\left(\mathbf{M}\left(\mathbf{d}-\mu\right)\left(\mathbf{M}\left(\mathbf{d}-\mu\right)\right)^T\right) \\ &= E\left(\mathbf{M}\left(\mathbf{d}-\mu\right)\left(\mathbf{d}-\mu\right)^T\mathbf{M}^T\right) \\ &= \mathbf{M}\,E\left(\mathbf{d}-\mu\left(\mathbf{d}-\mu\right)^T\right)\mathbf{M}^T \\ &= \mathbf{M}\Sigma_d\mathbf{M}^T \end{aligned} \tag{17}$$

Then, by substituting $\Sigma_d = \sigma_d^2\mathbf{I}$, Σ_m can be expressed as:

$$\Sigma_m = \sigma_d^2\mathbf{M}\mathbf{M}^T \tag{18}$$

Therefore, for the model parameters to be uncorrelated with uniform variance σ_m^2, Σ_m must satisfy the following:

$$\Sigma_m = \sigma_m^2\mathbf{I} \tag{19}$$

Thus, if we substitute (19) for (18), we get:

$$\begin{aligned} \sigma_m^2\mathbf{I} &= \sigma_d^2\mathbf{M}\mathbf{M}^T \\ \Leftrightarrow\quad \mathbf{M}\mathbf{M}^T &= \left(\frac{\sigma_m}{\sigma_d}\right)^2\mathbf{I} \end{aligned} \tag{20}$$

which is the condition we were looking for.

Question #3

In this question, we are asked to find the probability distribution of x^2 when x follows the normal distribution $N(0,1)$. If we denote x^2 by y (≥ 0), the question is equivalent to finding the probability distribution $f(y)$ when $\sqrt{y}$ follows the normal distribution $N(0,1)$. By the definition of a continuous probability distribution, $f(y)$ satisfies:

$$\text{Prob}\left(y \leq y' \leq y+dy\right) = f(y)dy \tag{21}$$

[*2]We know that:

$$y \leq y' \leq y + dy \;\Leftrightarrow\; \begin{cases} \sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy} & \text{(22a)} \\ -\sqrt{y + dy} \leq \sqrt{y'} \leq -\sqrt{y} & \text{(22b)} \end{cases}$$

Therefore, from (21), we get:

$$f(y)dy = \text{Prob}\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy}\right) + \text{Prob}\left(-\sqrt{y + dy} \leq \sqrt{y'} \leq -\sqrt{y}\right) \tag{23}$$

Since $\sqrt{y'}$ follows the normal distribution which is symmetric around its mean (in this case 0), we have:

$$\text{Prob}\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy}\right) = \text{Prob}\left(-\sqrt{y + dy} \leq \sqrt{y'} \leq -\sqrt{y}\right) \tag{24}$$

(Figure 2 below shows the symmetry of normal distribution $N(0,1)$. Prob $\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy}\right)$ and Prob $\left(-\sqrt{y + dy} \leq \sqrt{y'} \leq -\sqrt{y}\right)$ are represented by the narrow area in between two vertical lines.)

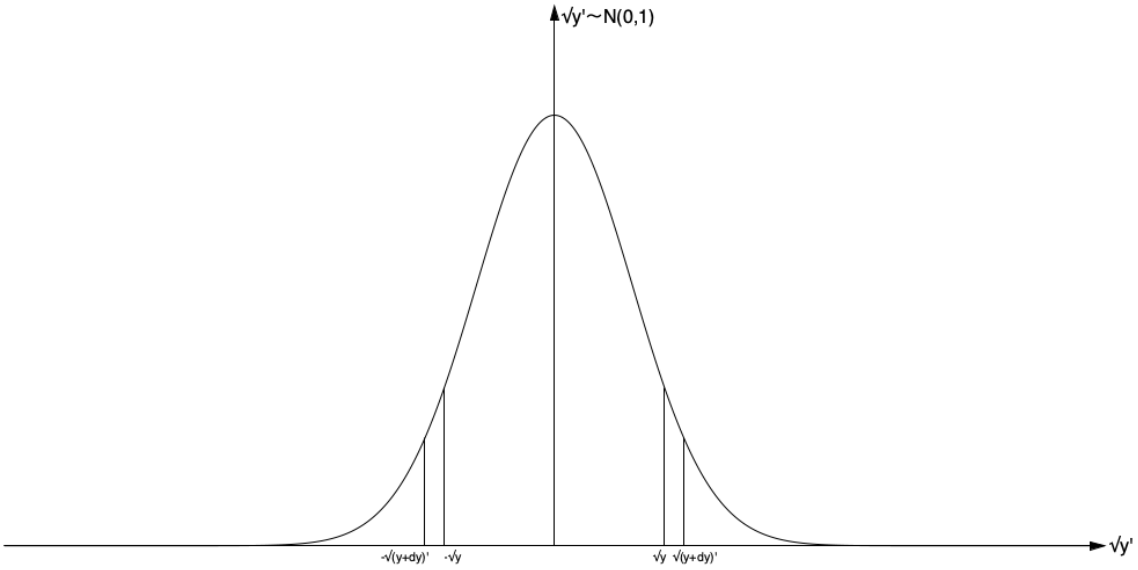

Figure2: **Symmetry of normal distribution N(0,1)**

Thus, by using equation (23) and (24), we get:

$$f(y)dy = 2 \cdot \text{Prob}\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy}\right) \tag{25}$$

Since $\sqrt{y'}$ follows the normal distribution, Prob $\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y + dy}\right)$ can be calculated as:

[*2]In this case, y' is the random variable.

$$\begin{aligned}\text{Prob}\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y+dy}\right) &= \int_{\sqrt{y}}^{\sqrt{y+dy}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\sqrt{y'}^2}{2}\right) d\left(\sqrt{y'}\right) \\ &= \int_{y}^{y+dy} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y'}{2}\right) \frac{dy'}{2\sqrt{y'}}\end{aligned} \tag{26}$$

and since dy is an infinitesimal, (26) can be further calculated as:

$$\text{Prob}\left(\sqrt{y} \leq \sqrt{y'} \leq \sqrt{y+dy}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y}{2}\right) \frac{dy}{2\sqrt{y}} \tag{27}$$

Therefore, by substituting (27) into (25),we obtain:

$$\begin{aligned}f(y) &= 2 \times \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y}{2}\right) \frac{1}{2\sqrt{y}} \\ &= \frac{1}{2^{1/2}\Gamma(1/2)}\, y^{\frac{1}{2}-1} e^{-y/2}\end{aligned} \tag{28}$$

[*3] If we consider the χ^2-distribution with k degrees of freedom which is given by:

$$f_\chi(t;k) = \frac{1}{2^{k/2}\Gamma(k/2)}\, t^{\frac{k}{2}-1} e^{-t/2} \tag{29}$$

one shall see that (28) is equivalent to χ^2-distribution with 1 degree of freedom, and thus:

$$y \sim \chi_1^2 \tag{30}$$

Therefore, if we recall that $y = x^2$, the probability distribution we are looking for is:

$$x^2 \sim \chi_1^2 \tag{31}$$

[*3] We used $\Gamma(1/2) = \sqrt{\pi}$