MIDDLESEX UNIVERSITY

MSc Data Science

CST4050

Submitted to: Kostanca Kovaci

Subject: Modelling, Regression and Machine Learning

Challenge Week 2

Date:08/10/2024

Department of MSc Data Science

London

Challenge 2

In this week challenge we are going to classify emails as spam or not spam based on the presence of three keywords: **WIN**, **FREE** , and **OFFER**. We have the below dataset:

Email	Contains WIN?	Contains FREE?	Contains OFFER?	Spam or Not Spam?
“WIN a FREE car”	Yes	Yes	No	Spam
“Special OFFER just for you”	No	No	Yes	Not Spam
“WIN a big FREE prize!”	Yes	Yes	No	Spam
“Limited time OFFER”	No	No	Yes	Not Spam
“Claim your FREE vacation!”	No	Yes	No	Spam
"Get an OFFER you can't refuse"	No	No	Yes	Not Spam
"WIN FREE gifts and prizes!"	Yes	Yes	No	Spam
"Exclusive OFFER ends soon"	No	No	Yes	Not Spam
"FREE upgrade on your purchase"	No	Yes	No	Not Spam
"WIN a FREE bonus now!"	Yes	Yes	No	Spam
"Special OFFER for members"	No	No	Yes	Not Spam
"WIN big with our OFFER"	Yes	No	Yes	Spam
"FREE gifts with your purchase"	No	Yes	No	Not Spam
"Huge OFFER just for you!"	No	No	Yes	Spam
"WIN a trip to Hawaii!"	Yes	No	No	Spam
"Claim your FREE OFFER now"	No	Yes	Yes	Spam
"FREE shipping on all orders"	No	Yes	No	Not Spam
"OFFER ends soon!"	No	No	Yes	Not Spam

At this point we analyze the data set and try to figure out the keywords (WIN, FREE, OFFER) and spammed or not spammed emails, for the following exercises. Consider that the incident of the three keywords (WIN, FREE, OFFER) in an email are independent events.

In math, two events A and B are independently of each other, when the probability of both occurring together is the product of their individual probabilities:

$$P(A \cap B) = P(A) \times P(B)$$

This means that the presence of one keyword does not influence the appearance of the others, as an example:

$$P(WIN \cap FREE|Spam) = P(WIN|Spam) \times P(FREE|Spam)$$

STEPS TO SOLVE

1. **Calculate Prior Probabilities**

   First, we need to define what prior probabilities is and how is cultivated. Prior probabilities usually represent the initial probability of how something has happened before we observe any data or information. It allows us to find the probability information of two independent events that work together. Prior probabilities are symbolized as P(A) and P(B), for two events A and B.

   Now, for our challenge, we have to define P(Spam) and P(NotSpam) as:

   - $P(Spam) = \frac{Only\ Spam\ Emails}{Total\ Emails} = \frac{9}{18} = \frac{1}{2} = 0.5\ or\ 50\%$
   - $P(NotSpam) = \frac{Only\ Not\ Spam\ Emails}{Total\ Emails} = \frac{9}{18} = \frac{1}{2} = 0.5\ or\ 50\%$

By counting each spam and not spam mails we find that we have **total:18 mails,** only **spam:9** and only **not spam:9.**

2. **Calculate Conditional Probabilities**

   Now we need to calculate the conditional probabilities of each keyword that we have that shows if the email is spam or it is not spam, based on the data that we have. For example: $P(WIN|Spam), P(WIN|Not\ Spam)$.

   For the keyword WIN:

   - $\boldsymbol{P(WIN|Spam) = \frac{N(WIN\ and\ Spam)}{N(Spam)} = \frac{6}{9} = \frac{2}{3} \approx 0.67}$
   - $\boldsymbol{P(WIN|NotSpam) = \frac{N(WIN\ and\ NotSpam)}{N(NotSpam)} = \frac{0}{9} = 0}$

     Total Spam: 9
     Total NotSpam: 9
     Spam(WIN): 6
     NotSpam(WIN): 0

   For the keyword FREE:

   - $\boldsymbol{P(FREE|Spam) = \frac{N(FREE\ and\ Spam)}{N(Spam)} = \frac{5}{9} \approx 0.55}$
   - $\boldsymbol{P(FREE|NotSpam) = \frac{N(FREE\ and\ NotSpam)}{N(NotSpam)} = \frac{3}{9} \approx 0.33}$

Spam(FREE): 5
NotSpam(FREE): 3

<u>For the keyword OFFER:</u>

- $\boldsymbol{P(OFFER|Spam) = \frac{N(OFFER\ and\ Spam)}{N(Spam)} = \frac{3}{9} \approx 0.33}$
- $\boldsymbol{P(OFFER|NotSpam) = \frac{N(OFFER\ and\ NotSpam)}{N(NotSpam)} = \frac{6}{9} \approx 0.67}$

Spam(OFFER): 3
NotSpam(OFFER): 6

3. Apply Bayes' Theorem:

We are going to use Bayes' Theorem to compute the probability that the email is spam, given that it contains certain keywords. For example, for an email containing WIN and FREE calculate:

$$\mathrm{P(Spam|WIN\ \cap\ FREE)} = \frac{\mathrm{P(WIN\ \cap\ FREE|Spam)} \times\ \mathrm{P(Spam)}}{\mathrm{P(WIN \cap\ FREE)}}$$

- **<u>WIN and Free:</u>**
  Total containing WIN and FREE: 4

1. Calculate total WIN and FREE for Spam email:

$$\boldsymbol{P(WIN \cap FREE|Spam) = P(WIN|Spam) \times P(FREE|Spam)}$$
$$\boldsymbol{= \frac{6}{9} \times \frac{5}{9} = \frac{30}{81} = \frac{10}{27} \approx 0.37}$$

2. Calculate total WIN and FREE:

$$\boldsymbol{P(WIN \cap FREE)}$$
$$\boldsymbol{= P(WIN \cap FREE|Spam) \times P(Spam)}$$
$$\boldsymbol{+ P(WIN \cap FREE|NotSpam) \times P(NotSpam)}$$
$$\boldsymbol{= \frac{10}{27} \times \frac{1}{2} + \frac{3}{9} = \frac{10}{54} + \frac{3}{9} = \frac{10}{54} + \frac{18}{54} = \frac{28}{54} \approx 0.52}$$

We know:

$$\boldsymbol{P(Spam) = \frac{1}{2} = 0.5}$$

We find:

$$P(WIN \cap FREE|NotSpam) = P(WIN|NotSpam) \times P(FREE|NotSpam) = 0 + \frac{3}{9} = \frac{3}{9}$$

3. The total function that we need:

$$P(Spam|WIN \cap FREE) = \frac{P(WIN \cap FREE|Spam) \times P(Spam)}{P(WIN \cap FREE)} = \frac{\left(\frac{10}{27}\right) \times \left(\frac{1}{2}\right)}{\frac{28}{54}} = \frac{\frac{10}{54}}{\frac{28}{54}} = \frac{10}{28} = 0.55$$

After the results of our function, we have to:

1. Classify the Email: if our probability of being spam is greater than 50%,we are going to classify the email as spam, otherwise, if it is lower than 50% we are going to classify it as not spam.

2. Apply Your Classifier: use your classifier to infer which of the following emails is spam or not spam based on the keywords they contain:

Email	Contains WIN?	Contains FREE?	Contains OFFER?	Spam or Not Spam?
"WIN a FREE trip!"	Yes	Yes	No	?
"Exclusive OFFER for you!"	No	No	Yes	?
"WIN big with this special OFFER!"	Yes	No	Yes	?
"Get your FREE OFFER now!"	No	Yes	Yes	?
"WIN a FREE prize with our OFFER!"	Yes	Yes	Yes	?

SOLUTION

Solving the email dataset that is given to find if it the email is spam or not spam:

- **For the 1st email we have WIN and OFFER:**
  From our previous solution when we have WIN and FREE, but no OFFER, we found the probability greater than 0.5 or 50%. So, for the **1st email** we have a probability 0.55, which means that we have a **spam** email.

- **For the 2nd email, we have an OFFER keyword only so:**

$$P(OFFER|Spam) = \frac{3}{9} \approx 0.33$$

According to our probability result the **2nd email** is **not spam**, because **0.33** < 0.5.

- **For the 3rd email we have a WIN and an OFFER, so:**

$$\mathbf{P(Spam|WIN \cap OFFER)} = \frac{\mathbf{P(WIN \cap OFFER|Spam) \times P(Spam)}}{\mathbf{P(WIN \cap OFFER)}}$$

a. $P(WIN \cap OFFER|Spam) = P(WIN|Spam) \times P(OFFER|Spam) = \frac{6}{9} \times \frac{3}{9} = \frac{18}{81} \approx 0.22$

b. $P(WIN \cap FREE) = P(WIN \cap FREE|Spam) \times P(Spam) + P(WIN \cap FREE|NotSpam) \times P(NotSpam) = \frac{18}{81} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2} = \frac{18}{162} + \frac{2}{6} = \frac{18}{162} + \frac{54}{162} = \frac{72}{162} = 0.44$

c. $P(WIN \cap FREE|NotSpam) = P(WIN|NotSpam) \times P(FREE|NotSpam) = 0 + \frac{6}{9} = \frac{2}{3} = 0.67$

From all the above we have:

$$\mathbf{P(Spam|WIN \cap OFFER)} = \frac{\mathbf{P(WIN \cap OFFER|Spam) \times P(Spam)}}{\mathbf{P(WIN \cap OFFER)}}$$
$$= \frac{\mathbf{0.22 \times 0.5}}{\mathbf{0.44}} = \frac{\mathbf{0.11}}{\mathbf{0.44}} = \mathbf{0.25}$$

According to the result the **3rd email** is **not a spam**, because **0.25 < 0.50**.

- **For the 4th email we have a FREE and a OFFER:**

  a. $P(FREE \cap OFFER|Spam) = P(FREE|Spam) \times P(OFFER|Spam) = \frac{5}{9} \times \frac{3}{9} = \frac{15}{81} \approx 0.18$

  b. $P(FREE \cap OFFER|NotSpam) = P(FREE|NotSpam) \times P(OFFER|NotSpam) = \frac{3}{9} \times \frac{6}{9} = \frac{18}{81} \approx 0.22$

  c. $P(FREE \cap OFFER) = P(FREE \cap OFFER|Spam) \times P(Spam) + P(FREE \cap OFFER|NotSpam) \times P(NotSpam) = 0.18 \times 0.5 + 0.22 \times 0.5 = 0.09 + 0.11 = 0.2$

$$\mathbf{P(Spam|FREE \cap OFFER)} = \frac{\mathbf{P(FREE \cap OFFER|Spam) \times P(Spam)}}{\mathbf{P(FREE \cap OFFER)}} = \frac{\mathbf{0.18 \times 0.5}}{\mathbf{0.2}} = \frac{\mathbf{0.09}}{\mathbf{0.2}} = \mathbf{0.45}$$

According to the result **4th email** is **not a spam**, since **0.45 < 0.50**.

- **For the 5th email, we have WIN, FREE, OFFER, so:**

  a. $P(WIN \cap FREE \cap OFFER|Spam) = P(WIN|Spam) \times P(FREE|Spam) \times P(OFFER|Spam) = \frac{6}{9} \times \frac{5}{9} \times \frac{3}{9} = \frac{90}{729} = 0.12$

  b. $P(WIN \cap FREE \cap OFFER) = P(WIN \cap FRE \cap OFFER|Spam) \times P(Spam) + P(WIN \cap FREE \cap OFFER|NotSpam) \times P(NotSpam) = 0.12 \times 0.5 + 0 = 0.06$

  c. $P(WIN \cap FREE \cap OFFER|NotSpam) = 0 \times \frac{3}{9} \times \frac{6}{9} = 0$

  d. $P(\,Spam \mid WIN \cap FREE \cap OFFER\,) = \frac{P(WIN \cap FREE \cap OFFER|Spam) \times P(Spam)}{P(WIN \cap FREE \cap OFFER)} = \frac{0.12 \times 0.5}{0.06} = \frac{0.06}{0.06} = 1$

According to the results the **5th email** is considered as **spam**, because **1 > 0.5**.

Results:

Email	Contains WIN?	Contains FREE?	Contains OFFER?	Spam or Not Spam?
"WIN a FREE trip!"	Yes	Yes	No	Spam
"Exclusive OFFER for you!"	No	No	Yes	Not Spam
"WIN big with this special OFFER!"	Yes	No	Yes	Not Spam
"Get your FREE OFFER now!"	No	Yes	Yes	Not Spam
"WIN a FREE prize with our OFFER!"	Yes	Yes	Yes	Spam

CONCLUSION

1. After all the results that we have we understand that even with one parameter of WIN, FREE or OFFER we can find and recognize, with the help of probability, if an email is spam or not spam.

2. However, that doesn't mean that it is always correct. We also need to know the other keywords so we can define and calculate again if an email is a spam or not spam.

3. Moreover, we now know that when we have the three keywords WIN, FREE and OFFER as our parameters, our probability resolution will be a spam.

4. In conclusion, even with the dataset that we have we are never 100% sure when our probability is in the meddle, such as 0.5, if it totally a spam or if it is not spam.