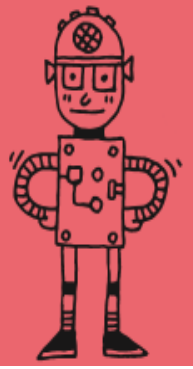
Spam Filter

데이터사이언스 학과
김 승 환
swkim4610@inha.ac.kr



1 Spam Mail 유래

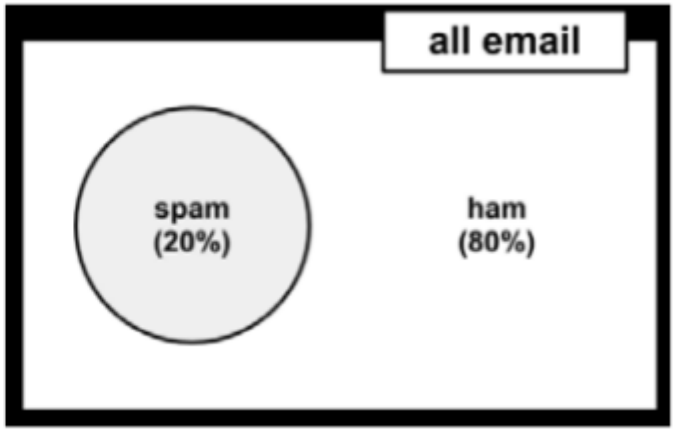
- 스팸메일의 유래: 2차 세계대전 때, 미국이 영국에 원조물자로 스팸을 공급했는데 매일 스팸만 먹어서 영국인은 스팸을 쳐다보기도 싫어하게 된것이 유래가 되어 쳐다보기도 싫은 메일이란 뜻으로 사용됨
- 스팸필터란 네이버, 구글 메일 서비스에서 스팸 메일이 도착하면 자동으로 스팸메일함으로 이동시키는 서비스임
- 가장 쉬운 스팸필터는 특정 단어 혹은 특정 발신자의 메일을 스팸 메일함으로 이동하는 방법인데 예를 들어, "대리운전", "대리", "운전" 이 포함된 메일이나 abc@email.com으로 온 메일을 스팸 메일함으로 이동하는 방법임
- 지금 여러 메일 서비스에는 이보다 지능적인 방법이 탑재되어 있음
- 우리는 나이브 베이즈라는 이론을 이용하여 스팸분류를 할 것임

2 조건부 확률

- 일반적으로 스팸메일을 받을 확률은 작지만, "비아그라" 라는 단어가 포함된 메일이라면?

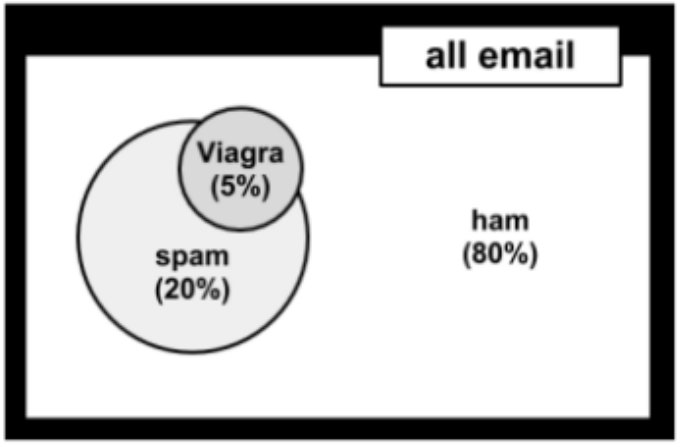
Probability

P(spam) = 0.2, P(ham) = 0.8



Conditional Probability

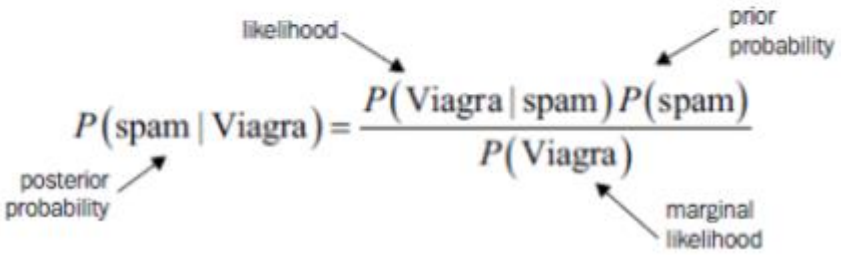
P(spam|viagra) = ?



2 조건부 확률

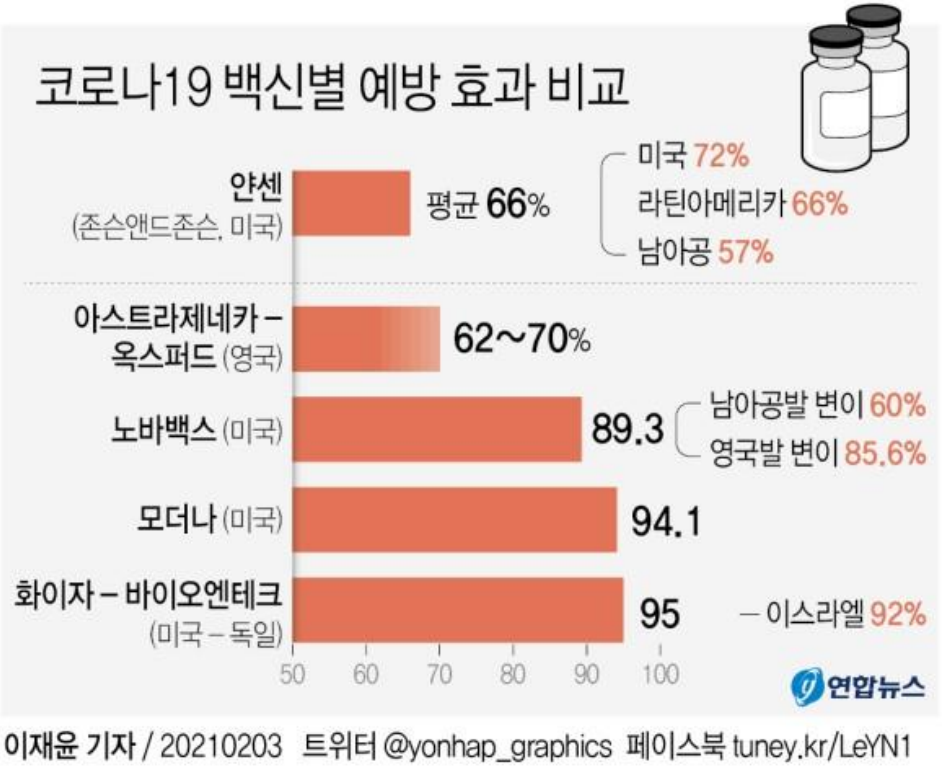
- Tomas Bayes는 1701~1761의 영국 목사로 1763년 “An Essay towards solving a Problem in the Doctrine of Chances”라는 책에 아래와 같은 수식을 소개하였다.
- 이 수식은 B라는 정보가 사실이 되었을 때, A의 확률에 어떤 영향을 미치는가에 대한 정리로 현대 정보과학이론에 큰 영향을 준 이론이다.
- 우리의 뇌 신경계(Neuron)는 불확실성의 세계에서 주어진 정보를 토대로 최적에 가까운 의사결정을 내리게 진화해 왔다. 즉, 우리가 이전에 알고 있던 정보(prior)에 새로 습득한 정보(likelihood)를 조합해 이를 바탕으로 사후 확률(posterior)을 예측해 결정을 내린다.

Bayes Law $$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$



2 조건부 확률

- 화이자 백신의 예방효과 95%는?
- 이 수치는 코로나에 감염이 되었다는 가정하에 이 백신을 통해 감염된 바이러스가 발현되지 않을 확률을 계산한 것이다.
- 계산 방법을 알아보자.
- 백신을 테스트를 할 사람을 모집하고 사람들을 두 그룹으로 랜덤하게 나눈다.
- 첫번째 그룹 백신 투여군은 백신을 투여하고, 두번째 그룹 대조군은 위약(placebo)을 투여한다. 계산 편의를 위해 각각 1000명이라고 하자.
- 충분한 시간 관찰 후에 백신 투여군에서 5명이 걸리고, 대조군에서는 100명이 걸렸다면 대조군처럼 100명이 걸려야할 상황에 5명만 걸렸으니 95명이 예방된 것이다.



2 조건부 확률

- '비아그라' 라는 단어가 포함된 메일이 수신되면 스팸이 확률이 80%이다.

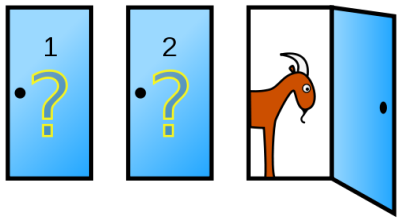
	Viagra		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

	Viagra		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

$$P(Spam|Viagra) = \frac{P(Spam \cap Viagra)}{P(Viagra)} = \frac{4/100}{5/100} = 0.8$$

- Monty Hall Problem(스포츠카는 어디에?)

  방 3개 중 하나에 스포츠카, 두 곳에는 염소가 있다. 퀴즈 참가자가 1번에 스포츠카가 있다고 응답했다. 이 때, 사회자는 3번 방을 열어보이면서 선택을 바꾸겠냐고 질문했다. 숙제: 여러분의 선택은?



3 Naïve Bayes Theorem

- Naïve: 소박하고 천진하다~ (부정적 의미)

- 나이브 베이즈는 조건부 확률을 이용하여 미리 특정 단어들이 관측되었을 때, Spam일 확률을 계산해 놓고 메일이 도착했을 때,특정 단어의 포함여부를 계산하여 스팸 확률을 추정하는 알고리즘이다.
- 문서에서 단어가 나타날 확률이 서로 독립이 아니다. 즉, "사랑" 과 "행복"은 동시에 나올 수 있는 가능성이 높은 단어이기 때문에 $P("love" \cap "happy") = P("love") \cdot P("happy")$ 가 아니다. 즉, 서로 독립이 아니라는 뜻이다.
- 하지만, 수많은 단어의 조합에 대해 확률을 추정하는 것은 어려운 일이다.
- 어렵기 때문에 독립이 아닌 것을 독립이라고 가정하고 계산하자라는 아이디어가 나이브 베이즈 알고리즘이다.

- 질문: 요즘같이 컴퓨팅 성능이 좋은 시대에 왜 독립을 가정하고 정확하지 않게 계산할까?

3 Naïve Bayes Theorem

- 아래는 비아그라, 구독철회 단어가 있고, 머니, 채소는 없는 메일이 스팸일 확률을 계산하는 나이브 베이즈 식이다.

Using Bayes' theorem, we can define the problem as shown in the following formula, which captures the probability that a message is spam, given that Viagra = Yes, Money = No, Groceries = No, and Unsubscribe = Yes:

$$P(\text{Spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) = \frac{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 \mid \text{spam})\,P(\text{spam})}{P(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4)}$$

- 나이브(순진)하게 단어들끼리 서로 독립이라고 가정하면 ...

$$= \frac{P(W_1|spam)P(!W_2|spam)P(!W_3|spam)P(W_4|spam)P(spam)}{P(W_1 \cap !W_2 \cap !W_3 \cap W_4)}$$

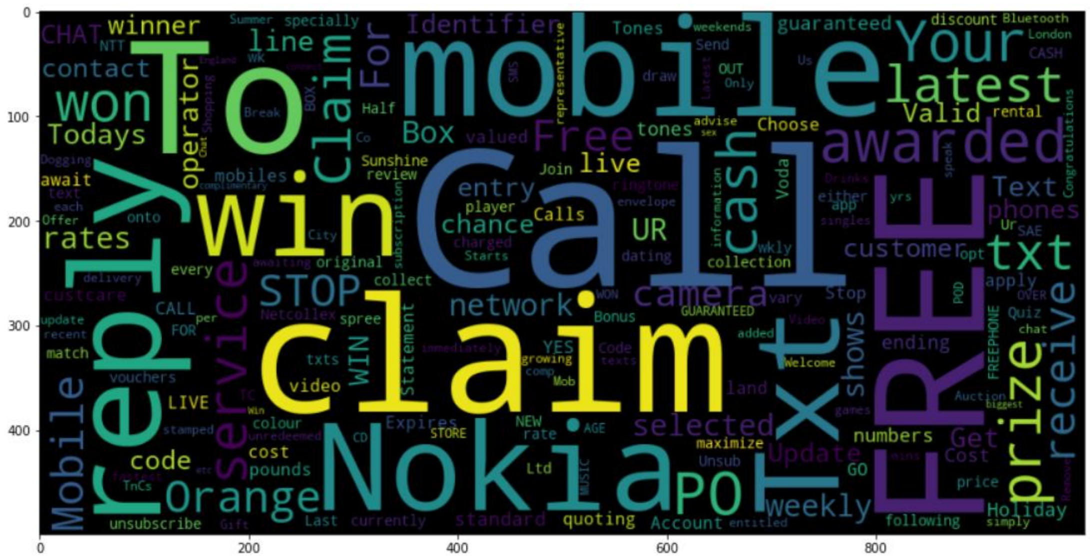
$$= \frac{P(W_1|spam)P(!W_2|spam)P(!W_3|spam)P(W_4|spam)P(spam)}{P(W_1 \cap !W_2 \cap !W_3 \cap W_4|spam)P(spam) + P(W_1 \cap !W_2 \cap !W_3 \cap W_4|ham)P(ham)}$$

4 TDM(Term Document Matrix)

- 자연어 메일을 학습 가능한 형태로 만들기 위해 메일의 특징을 독립변수로 만들고 종속변수인 타겟 변수를 만들어야 한다.
- 자연어 처리에서는 아래와 같은 TDM을 주로 사용한다.
- 아래는 각 메일에 대해 단어가 몇 번 나왔는지를 카운트하여 특징을 만드는 행렬이다.
- 이는 대표적인 Sparse Matrix(희소행렬)로 CS 분야 자료구조에서 중요한 내용이다.

Email#	"online"	"Viagra"	"Order now!!!"	"offer"	"win"	SPAM?
1	1	0	1	0	1	YES
2	1	1	1	1	0	YES
3	1	0	0	1	0	NO
4	0	1	1	1	1	YES
5	0	0	1	1	0	NO

5 실습



Call
claim
FREE
mobile
Nokia
reply
To
win
Txt
awarded
latest
Your
prize
receive
service
txt
STOP
network
Mobile
Orange
PO
weekly
won
Todays
rates
code
cash
chance
entry
camera
customer
Text
phones
selected
Update
0
100
200
300
400
0
200
400
600
800

you
the
to
in
that
but
can
like
me
my
and
are
it
get
we
how
if
so
not
for
is
at
up
your
have
be
its
go
just
with
was
all
want
of
about
need
know
do
on
back
will
come
still
no
0
100
200
300
400
0
200
400
600
800

감사합니다