Machine Learning Applications: Learning to classify text, Spam Filter

Alessio Benavoli

CSIS
University of Limerick

Summary of probability concepts

---

joint distribution

Given two (or more) variables $p(x, y)$ is called the joint probability of x and y

---

marginal distribution

Given a joint distr. $p(x, y)$ the marginal distr. of x is defined by

$$p(x = \mathrm{x}) = \sum_{\mathrm{y} \in \mathrm{dom}(y)} p(x = \mathrm{x}, y = \mathrm{y})$$

---

conditional distribution

The probability of event $x = \mathrm{x}$ conditioned on knowing event $y = \mathrm{y}$ (or more shortly, the probability of $x = \mathrm{x}$ given $y = \mathrm{y}$) is defined as

$$p(x = \mathrm{x} | y = \mathrm{y}) \equiv \frac{p(x = \mathrm{x}, y = \mathrm{y})}{p(y = \mathrm{y})} \quad \textit{this is also called Bayes' Rule}$$

Summary

Independence

Variables x and y are independent if knowing one event gives no extra information about the other event. Mathematically, this is expressed by

$$p(x = \mathsf{x}, y = \mathsf{y}) = p(x = \mathsf{x})p(y = \mathsf{y})$$

Independence of x and y is equivalent to

$$p(x = \mathsf{x}|y = \mathsf{y}) = p(x = \mathsf{x}) \Leftrightarrow p(y = \mathsf{y}|x = \mathsf{x}) = p(y = \mathsf{y})$$

for all $\mathsf{x} \in \mathrm{dom}(x)$, $\mathsf{y} \in \mathrm{dom}(y)$, then the variables x and y are said to be independent.

Conditional independence

Two variables x and y are conditionally independent given z if

$$p(x = \mathsf{x}, y = \mathsf{y}|z = \mathsf{z}) = p(x = \mathsf{x}|z = \mathsf{z})p(y = \mathsf{y}|z = \mathsf{z})$$

Independence of x and y is equivalent to

$$p(x = \mathsf{x}|y = \mathsf{y}, z = \mathsf{z}) = p(x = \mathsf{x}|z = \mathsf{z}) \Leftrightarrow p(y = \mathsf{y}|x = \mathsf{x}, z = \mathsf{z}) = p(y = \mathsf{y}|z = \mathsf{z})$$

for all $\mathsf{x} \in \mathrm{dom}(x)$, $\mathsf{y} \in \mathrm{dom}(y)$, $\mathsf{z} \in \mathrm{dom}(z)$,

Spam

Is an email that includes the words "Bye" and "Won" Spam? We consider a simplified spam-filter. We will only use these two words, "Bye", "Won", to assess if an email is spam.

Variables:

$$S, B, W$$

Domain (possibility space):

$$S \in \{0, 1\}$$

where 0 means not-spam and 1 means spam.

$$B \in \{0, 1\}$$

where 0 means it does not include "Bye" and 1 means it includes "Bye".

$$W \in \{0, 1\}$$

where 0 means it does not include "Won" and 1 means it includes "Won".

Note that the email can also include other words, but we only focus on "Bye" and "Won".

Joint distribution

Assume we know the following joint distribution:

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

where the first row means

$$p(S = 0, B = 0, W = 0) \equiv p(S = 0 \text{ and } B = 0 \text{ and } W = 0) = 0.168$$

and similarly for the other 7 rows.

This table defines the joint probability distribution of the three variables! We will see later on how to derive this table from data (a dataset of emails), but for the moment we assume that we know it.

Why do we need probability theory?

Because we want to answer questions like:

1. What is the probability that an email is Spam? (no matter which words it includes)
2. What is the probability that an email is Spam and does not include "Won" ?
3. What is a probability that the email includes "Bye" given that the email is Spam?
4. What is the probability that an email is Spam given it includes "Won" but not "Bye"?
5. ...

These are all fundamental questions when we design a spam filter.

We will use these questions to review the probability rules learned in Week 1.

Marginalisation

What is the probability that an email is spam?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

We want to know $p(S = 1)$, we do not have it but we can compute it by marginalising out the other variables:

$$
\begin{aligned}
p(S = 1) &\equiv \sum_{b,w \in \{0,1\}} p(S = 1, B = b, W = w) \\
&= p(S = 1, B = 0, W = 0) + p(S = 1, B = 0, W = 1) \\
&+ p(S = 1, B = 1, W = 0) + p(S = 1, B = 1, W = 1) \\
&= 0.032 + 0.128 + 0.048 + 0.192 = 0.4
\end{aligned}
$$

Marginalisation

What is the probability that an email is spam and does not include "Won" ?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

We want to know $p(S = 1, W = 0)$, we do not have it but we can compute it by marginalising out the other variable:

$$\begin{aligned} p(S = 1, W = 0) &\equiv \sum_{b \in \{0,1\}} p(S = 1, B = b, W = 0) \\ &= p(S = 1, B = 0, W = 0) + p(S = 1, B = 1, W = 0) \\ &= 0.032 + 0.048 = 0.08 \end{aligned}$$

Conditional probability

What is the probability that an email includes the word Bye and Won given that is spam?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

We want to know $p(B = 1, W = 1|S = 1)$, we do not have it but we can compute it by applying Bayes' rule (conditional probability):

$$p(B = 1, W = 1|S = 1) = \frac{p(B = 1, W = 1, S = 1)}{p(S = 1)} = \frac{0.192}{0.4} = 0.48$$

Alternative equivalent formulas

There are alternative formulas we could use, they are equivalent. We can choose the one that is more convenient to use.

$$
\begin{aligned}
p(B=1,W=1|S=1) &= \frac{p(B=1,W=1,S=1)}{p(S=1)} \\
&= \frac{p(B=1|W=1,S=1)p(W=1,S=1)}{p(S=1)} \\
&= \frac{p(B=1|W=1,S=1)p(W=1|S=1)p(S=1)}{p(S=1)} \\
&= p(B=1|W=1,S=1)p(W=1|S=1)
\end{aligned}
$$

or

$$
\begin{aligned}
p(B=1,W=1|S=1) &= \frac{p(B=1,W=1,S=1)}{p(S=1)} \\
&= \frac{p(W=1|B=1,S=1)p(B=1,S=1)}{p(S=1)} \\
&= \frac{p(W=1|B=1,S=1)p(B=1|S=1)p(S=1)}{p(S=1)} \\
&= p(W=1|B=1,S=1)p(B=1|S=1)
\end{aligned}
$$

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

What is $p(B = 1|S = 1)$: probability that there is "Bye" given the email is spam?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(B = 1|S = 1) = \sum_{w\in\{0,1\}} p(B = 1, W = w|S = 1) = \sum_{w\in\{0,1\}} \frac{p(B{=}1,W{=}w,S{=}1)}{p(S{=}1)}$$

$$= \frac{p(S{=}1,B{=}1,W{=}0)}{p(S{=}1)} + \frac{p(S{=}1,B{=}1,W{=}1)}{p(S{=}1)}$$

$$= \frac{0.048}{0.4} + \frac{0.192}{0.4} = \frac{0.24}{0.4} = 0.6$$

and so $p(B = 0|S = 1) = 1 - p(B = 1|S = 1) = 0.4$.

$p(W = 1|S = 1)$

What is $p(W = 1|S = 1)$: probability that there is the word "Won" given the email is spam?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(W = 1|S = 1) = \sum_{b\in\{0,1\}} p(B = b, W = 1|S = 1) = \sum_{b\in\{0,1\}} \frac{p(B{=}b, W{=}w, S{=}1)}{p(S{=}1)}$$

$$= \frac{p(S{=}1, B{=}0, W{=}1)}{p(S{=}1)} + \frac{p(S{=}1, B{=}1, W{=}1)}{p(S{=}1)}$$

$$= \frac{0.128}{0.4} + \frac{0.192}{0.4} = \frac{0.32}{0.4} = 0.8$$

and so $p(W = 0|S = 1) = 1 - p(W = 1|S = 1) = 0.2$.

$p(W = 1|S = 0)$

What is $p(W = 1|S = 0)$: probability that there is the word "Won" given the email is not spam?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(W = 1|S = 0) = \sum_{b\in\{0,1\}} p(B = b, W = 1|S = 0) = \sum_{b\in\{0,1\}} \frac{p(B{=}b,W{=}1,S{=}0)}{p(S{=}0)}$$

$$= \frac{p(S{=}0,B{=}0,W{=}1)}{p(S{=}0)} + \frac{p(S{=}0,B{=}1,W{=}1)}{p(S{=}0)}$$

$$= \frac{0.072}{0.6} + \frac{0.108}{0.6} = \frac{0.18}{0.6} = 0.3$$

and so $p(W = 0|S = 0) = 1 - p(W = 1|S = 0) = 0.7$.

$p(W = 1|B = 1, S = 1)$

What is the probability that there is "Won" given the email is spam and includes "Bye"?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(W = 1|B = 1, S = 1) = \frac{p(W=1,B=1,S=1)}{p(B=1,S=1)}$$
$$= \frac{p(W=1,B=1,S=1)}{p(B=1|S=1)p(S=1)} = \frac{0.192}{0.6 \cdot 0.4} = 0.8$$

and so $p(W = 0|B = 1, S = 1) = 1 - p(W = 1|B = 1, S = 1) = 0.2$.

$p(W = 1|B = 0, S = 1)$

What is the probability that there is "Won" given the email is spam and does not include "Bye"?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(W = 1|B = 0, S = 1) = \frac{p(W=1,B=0,S=1)}{p(B=0,S=1)}$$
$$= \frac{p(W=1,B=0,S=1)}{p(B=0|S=1)p(S=1)} = \frac{0.128}{0.4 \cdot 0.4} = 0.8$$

and so $p(W = 0|B = 0, S = 1) = 1 - p(W = 1|B = 0, S = 1) = 0.2$.

Conditional independence

Are the variables B and W conditional independent given S? This is true if

$$p(W = w|B = b, S = s) = p(W = w|S = s) \text{ or } p(B = b|W = w, S = s) = p(B = b|S = s)$$

for all $w, b, s \in \{0, 1\}$. If we choose the definition at the left, we must verify that the probability in the following two columns are equal:

$$\begin{aligned}
p(W = 1|B = 1, S = 1) &= 0.8, \quad p(W = 1|S = 1) = 0.8, \\
p(W = 0|B = 1, S = 1) &= 0.2, \quad p(W = 0|S = 1) = 0.2, \\
p(W = 1|B = 0, S = 1) &= 0.8, \quad p(W = 1|S = 1) = 0.8, \\
p(W = 0|B = 0, S = 1) &= 0.2, \quad p(W = 0|S = 1) = 0.2, \\
p(W = 1|B = 1, S = 0) &= 0.3, \quad p(W = 1|S = 0) = 0.3, \\
p(W = 0|B = 1, S = 0) &= 0.7, \quad p(W = 0|S = 0) = 0.7, \\
p(W = 1|B = 0, S = 0) &= 0.3, \quad p(W = 1|S = 0) = 0.3, \\
p(W = 0|B = 0, S = 0) &= 0.7, \quad p(W = 0|S = 0) = 0.7
\end{aligned}$$

So B, W are conditional independent given S.

Independence

Are the variables B and S (in)dependent? That is, does the knowledge that an email is spam changes our belief that the email includes or doesn't include the word Bye?

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

$$p(B=1|S=1)=0.6 \quad p(B=1)=0.6$$
$$p(B=0|S=1)=0.4 \quad p(B=0)=0.4$$
$$p(B=1|S=0)=0.6 \quad p(B=1)=0.6$$
$$p(B=0|S=0)=0.4 \quad p(B=0)=0.4$$

So B, S are independent

Machine Learning

Decision: What is the probability that an email that includes the word "Bye" and "Won" is spam?

We have seen that if we know this joint probability

S	B	W	$Prob$
0	0	0	0.168
1	0	0	0.032
0	1	0	0.252
1	1	0	0.048
0	0	1	0.072
1	0	1	0.128
0	1	1	0.108
1	1	1	0.192

we can compute

$$p(S = s|B = b, W = w)$$

that is what a spam filter computes to make decisions. This probability helps us to filter all the mails which are marked as Spam and then store them in a Spam folder.

Machine Learning

The problem is that we do not know the joint distribution:

S	B	W	$Prob$
0	0	0	?
1	0	0	?
0	1	0	?
1	1	0	?
0	0	1	?
1	0	1	?
0	1	1	?
1	1	1	?

but we have past data, that is user-annotated emails as spam or not-spam (the class column in the table).

	text	class
0	Did you hear about the new "Divorce Barbie"? I...	1
1	Will u meet ur dream partner soon? Is ur caree...	1
2	Maybe i could get book out tomo then return it...	0
3	Boltblue tones for 150p Reply POLY# or MONO# e...	1
4	Your credits have been topped up for http://ww...	1
5	22 days to kick off! For Euro2004 U will be ke...	1
6	Hi I'm sue. I am 20 years old and work as a la...	1
7	08714712388 between 10am-7pm Cost 10p	1

A ML algorithm can utilize such data to learn the "?" probabilities and use them to predict the correct label (spam or not-spam) of unseen new emails.

Classifier

A ML algorithm that uses *annotated (labeled)* past-data to predict the label (also called class) of unseen data is called **classifier**.

This task is called **classification**.

In the spam filter:

- Class: spam or not-spam
- Features/attributes/inputs: words in the email.

Note that a label is a categorical variable. Therefore, we can say that a classifier is a ML algorithm that uses the inputs (features) to predict a categorical variable.

Machine Learning

Here we need to learn

S	B	W	$Prob$
0	0	0	?
1	0	0	?
0	1	0	?
1	1	0	?
0	0	1	?
1	0	1	?
0	1	1	?
1	1	1	?

We have $2^3 - 1$ unknown quantities (because they sum up to one, we have only 7 unknowns and not 8). In a real spam filter, we must consider hundreds of words. For instance, if we consider the 300 most common words in English, then

$2^{300} - 1$ is bigger than the number of atoms in the Universe!

We cannot even store that table in the Universe. What do we do?

Conditional Independence

$$p(S=1|B=1,W=1) = \frac{p(B=1,W=1|S=1)p(S=1)}{p(B=1,W=1|S=1)p(S=1)+p(B=1,W=1|S=0)p(S=0)}$$

Given that B, W are independent given S, we can simplify it as

$$= \frac{p(W=1|S=1)p(B=1|S=1)p(S=1)}{p(W=1|S=1)p(B=1|S=1)p(S=1)+p(W=1|S=0)p(B=1|S=0)p(S=0)}$$

Conditional independence of the features given the class

In ML, we often assume that the features are conditional independent given the class (this is called "Naive" hypothesis because it is in general not true but it allows us to reduce the number of unknowns"?"). In this case we can only consider:

S	B	$Prob$
0	0	$?$
1	0	$?$
0	1	$?$
1	1	$?$

S	W	$Prob$
0	0	$?$
1	0	$?$
0	1	$?$
1	1	$?$

so we have a probability table for each feature, that is $(4-1)\cdot 300 = 900$ probabilities "?" that we need to estimate from data.

This is more than feasible.

Conditional independence

That is why we care about

Conditional Independence

because it allows us to reduce the size of the unknowns.

Independence of the features and the class

If B and S are independent, B does not give us any information about S so we can discard it:

S	W	$Prob$
0	0	?
1	0	?
0	1	?
1	1	?

$(4-1) \cdot 299 = 897$ probabilities.

This is what we usually call "feature selection": we can get rid of all features that are independent of the class variable S.

Unknowns

The unknowns

S	W	$Prob$
0	0	?
1	0	?
0	1	?
1	1	?

are probabilities that are continuous variables.

We will model them as the bias of the coin and so we employ the same model we used for the Coin to design a spam filter.

Unknowns

We first apply Bayes's rule and write the joint distribution in the previous table as

$$p(S, W) = p(W|S)p(S)$$

Note that S is a binary variable ($S = 1$ means spam and 0 means not-spam). We introduce two variables $0 \leq \theta_{S=1}, \theta_{S=0} \leq 1$ to model these two probabilities

$$p(S = 1) = \theta_{S=1}, \quad p(S = 0) = \theta_{S=0}$$

with $\theta_{S=0} = 1 - \theta_{S=1}$. Similarly for all the four values of $p(W|S)$ we introduce a variable

$$p(W = 1|S = 0) = \theta_{W=1|S=0}, \quad p(W = 0|S = 0) = \theta_{W=0|S=0} = 1 - \theta_{W=1|S=0}$$

and

$$p(W = 1|S = 1) = \theta_{W=1|S=1}, \quad p(W = 0|S = 1) = \theta_{W=0|S=1} = 1 - \theta_{W=1|S=1}$$

are probabilities that are continuous variables. They are like the θ in the coin example in Week 2.

Other quantities

We can then derive all the other conditional and marginal probabilities, for instance the probability that the email is spam given it includes "Won":

$$p(S = 1|W = 1) = \frac{\theta_{W=1|S=1}\theta_{S=1}}{\theta_{W=1|S=1}\theta_{S=1} + \theta_{W=1|S=0}\theta_{S=0}}$$

Unknown

Therefore, we can write the unknowns "?" probabilities as

S	W	$Prob$
0	0	$\theta_{S=0}\theta_{W=0\|S=0}$
1	0	$\theta_{S=1}\theta_{W=0\|S=1}$
0	1	$\theta_{S=0}\theta_{W=1\|S=0}$
1	1	$\theta_{S=1}\theta_{W=1\|S=0}$

note that $\theta_{S=0}\theta_{W=0|S=0}$ is the product of $\theta_{S=0}$ and $\theta_{W=0|S=0}$.

Observe that, when we defined

$$p(W = 1|S = 1) = \theta_{W=1|S=1}$$

it means that the value of the conditional probability mass function $p(W = 1|S = 1)$ is equal to $\theta_{W=1|S=1}$. We are now going to infer this unknown value $\theta_{W=1|S=1}$ from Data.

Dataset

	text	class
0	Did you hear about the new "Divorce Barbie"? I...	1
1	Will u meet ur dream partner soon? Is ur caree...	1
2	Maybe i could get book out tomo then return it...	0
3	Boltblue tones for 150p Reply POLY# or MONO# e...	1
4	Your credits have been topped up for http://ww...	1
5	22 days to kick off! For Euro2004 U will be ke...	1
6	Hi I'm sue. I am 20 years old and work as a la...	1
7	08714712388 between 10am-7pm Cost 10p	1

We assume a *bag-of-words* model (a representation used in natural language processing). In this model, a text (such as an email) is represented as the bag of its words. In other words, we only look at the words that are included in the email and we disregard the grammar, the word order, but we keep their multiplicity. Multiplicity means the number of times a word appears in an email.

Dataset

	text	class
0	Did you hear about the new "Divorce Barbie"? I...	1
1	Will u meet ur dream partner soon? Is ur caree...	1
2	Maybe i could get book out tomo then return it...	0
3	Boltblue tones for 150p Reply POLY# or MONO# e...	1
4	Your credits have been topped up for http://ww...	1
5	22 days to kick off! For Euro2004 U will be ke...	1
6	Hi I'm sue. I am 20 years old and work as a la...	1
7	08714712388 between 10am-7pm Cost 10p	1

Imagine the email #0,1,2,5 includes "Won" only once; the the email #3,4,6,7 does not include "Won"; the email #0,1,3,4,5,6 is spam and the email #2,7 is not. Then, under the *bag-of-words* model, the above set of emails reduce to the following dataset denoted as D:

Won	$Spam$
1	1
1	1
1	0
0	1
0	1
1	1
0	1
0	0

The first row means that the first email includes the word "Won" and it is "Spam" and so on for the other rows (remember that we only care about two words "Bye", "Won" and we discarded "Bye"). We assume that we have only 8 emails for simplicity.

Summing up

If we denote our dataset with D and the unknown probabilities as the vector of parameters

$$\theta = [\theta_{W=1|S=1}, \theta_{W=0|S=1}, \theta_{W=1|S=0}, \theta_{W=0|S=0}, \theta_{S=1}, \theta_{S=0}]$$

Our goal is to apply Bayes' rule:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}$$

to obtain $p(\theta|D)$, that means learning θ from data.

Now θ is a vector of continuous variables and, therefore, $p(\theta)$ cannot be a probability mass function, it must be a probability density function. To make this clear, we will use $f(\theta|D), f(D|\theta), f(\theta)$ instead of $p(\theta|D), p(D|\theta), p(\theta)$

Likelihood $f(D|\theta)$

Let us consider just the first row in D then

$$\begin{aligned} f((1,1)|\theta) &= \theta^1_{W=1|S=1}\theta^0_{W=0|S=1} \\ &\cdot \theta^0_{W=1|S=0}\theta^0_{W=0|S=0} \\ &\cdot \theta^0_{S=0}\theta^1_{S=1} \\ &= \theta^1_{W=1|S=1}\theta^1_{S=1} \end{aligned}$$

this is true by definition of the parameters $\theta_{W|S}$ and θ_S.
Now the first two rows

$$\begin{aligned} f((1,1),(1,1)|\theta) &= \theta^2_{W=1|S=1}\theta^0_{W=0|S=1} \\ &\cdot \theta^0_{W=1|S=0}\theta^0_{W=0|S=0} \\ &\cdot \theta^0_{S=0}\theta^2_{S=1} \\ &= \theta^2_{W=1|S=1}\theta^2_{S=1} \end{aligned}$$

Likelihood $p(D|\theta)$

Now the first three rows

$$\begin{aligned} f((1,1),(1,1),(1,0)|\theta) = & \theta^2_{W=1|S=1}\theta^0_{W=0|S=1} \\ & \cdot \theta^1_{W=1|S=0}\theta^0_{W=0|S=0} \\ & \cdot \theta^1_{S=0}\theta^2_{S=1} \end{aligned}$$

and now

$$\begin{aligned} f((1,1),(1,1),(1,0),(0,1)|\theta) = & \theta^2_{W=1|S=1}\theta^1_{W=0|S=1} \\ & \cdot \theta^1_{W=1|S=0}\theta^0_{W=0|S=0} \\ & \cdot \theta^1_{S=0}\theta^3_{S=1} \end{aligned}$$

Likelihood $p(D|\theta)$

If we consider the whole dataset D, we have

$$\begin{aligned} f(D|\theta) = & \theta^3_{W=1|S=1}\theta^3_{W=0|S=1} \\ & \cdot \theta^1_{W=1|S=0}\theta^1_{W=0|S=0} \\ & \cdot \theta^2_{S=0}\theta^6_{S=1} \end{aligned}$$

so we need just to sum up the counts (example: number of emails that include "Won" that are spam that is 3, that is the exponent of $\theta_{W=1|S=1}$.

Prior $f(\theta)$

We treat the thetas in the rows of the likelihood as three independent coins whose we aim to infer the bias

$$\begin{aligned} f(D|\theta) = & \theta^3_{W=1|S=1}\theta^3_{W=0|S=1} \\ & \cdot \theta^1_{W=1|S=0}\theta^1_{W=0|S=0} \\ & \cdot \theta^2_{S=0}\theta^6_{S=1} \end{aligned}$$

We therefore assume the following prior

$$\begin{aligned} f(\theta) = & \frac{1}{B(\alpha_1,\beta_1)}\theta^{\alpha_1-1}_{W=1|S=1}\theta^{\beta_1-1}_{W=0|S=1} \\ & \cdot \frac{1}{B(\alpha_2,\beta_2)}\theta^{\alpha_2-1}_{W=1|S=0}\theta^{\beta_2-1}_{W=0|S=0} \\ & \cdot \frac{1}{B(\alpha_3,\beta_3)}\theta^{\alpha_3-1}_{S=1}\theta^{\beta_3-1}_{S=0} \end{aligned}$$

Posterior $f(\theta|D)$

The posterior can be easily computed using the same formulas we used for the coin

$$f(\theta|D) = \frac{1}{B(\alpha_1+3, \beta_1+3)} \theta_{W=1|S=1}^{\alpha_1+3-1} \theta_{W=0|S=1}^{\beta_1+3-1}$$
$$\cdot \frac{1}{B(\alpha_2+1, \beta_2+1)} \theta_{W=1|S=0}^{\alpha_2+1-1} \theta_{W=0|S=10}^{\beta_2+1-1}$$
$$\cdot \frac{1}{B(\alpha_3+2, \beta_3+6)} \theta_{S=0}^{\alpha_3+2-1} \theta_{S=1}^{\beta_3+6-1}$$

Done!

Posterior $p(\theta|D)$

If we assume a uniform prior, because we assume we do not have any prior information, that is $\alpha_1 = \beta_1 = 1$, $\alpha_2 = \beta_2 = 1$, $\alpha_3 = \beta_3 = 1$, the posterior reduces to

$$\begin{aligned} f(\theta|D) = & \frac{1}{B(4,4)} \theta^3_{W=1|S=1} \theta^3_{W=0|S=1} \\ & \cdot \frac{1}{B(2,2)} \theta^1_{W=1|S=0} \theta^1_{W=0|S=0} \\ & \cdot \frac{1}{B(3,7)} \theta^2_{S=0} \theta^6_{S=1} \end{aligned}$$

This is the way we do learning in probabilistic ML.

General Recipe ML

What is the **general recipe ML** estimate of θ?
It is the Maximum Likelihood Estimator: $\hat{\theta} = \arg\max_\theta f(D|\theta)$

The solution is

$$\hat{\theta}_{S=1} = \frac{\text{number of spam emails}}{\text{number of emails}} = \frac{6}{8}$$

$$\hat{\theta}_{W=1|S=1} = \frac{\text{number of spam emails that include Won}}{\text{number of spam emails}} = \frac{3}{6}$$

$$\hat{\theta}_{W=1|S=0} = \frac{\text{number of not-spam emails that include Won}}{\text{number of not-spam emails}} = \frac{1}{2}$$

This is what the Multinomial Naive-Bayes classifier (MultinomialNB in sklearn) learns on our simplified dataset of 8 emails.

Is it bad?

Yes it is bad because

$$0.75 = \frac{6}{8} = \frac{12}{16} = \frac{60}{80} = \frac{600}{800}$$

it does not take into account the size of the dataset.

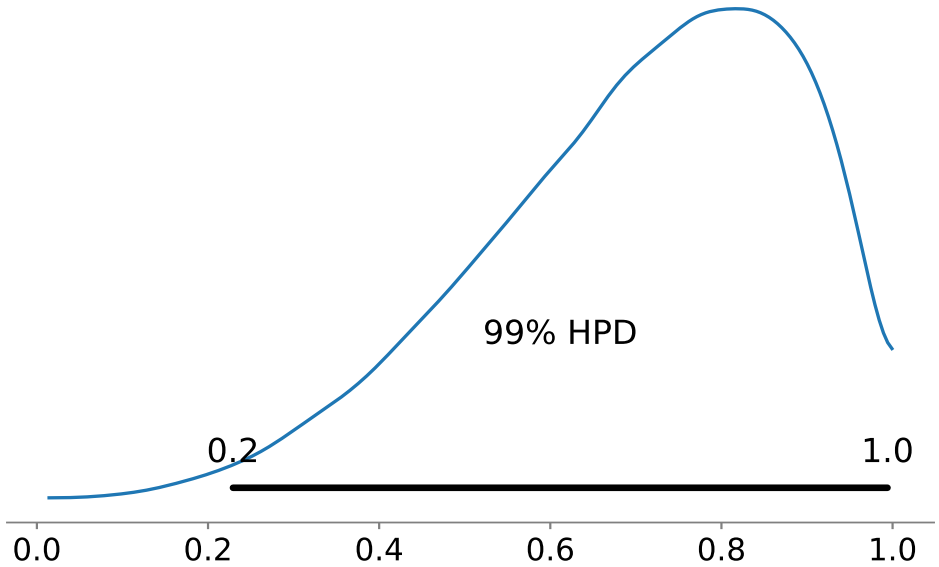
General Recipe ML

What is $p(S = 1|W = 1) =?$, that is the probability that an email is spam given it includes "Won"?

$$p(S = 1|W = 1) = \frac{\hat{\theta}_{W=1|S=1}\hat{\theta}_{S=1}}{\hat{\theta}_{W=1|S=1}\hat{\theta}_{S=1} + \hat{\theta}_{W=1|S=0}\hat{\theta}_{S=0}} = \frac{\frac{3}{6}\frac{6}{8}}{\frac{3}{6}\frac{6}{8} + \frac{1}{2}\frac{1}{2}} = 0.6$$

this means that if we use MultinomialNB to compute $p(S = 1|W = 1)$ and make the decision "move the email to spam folder when $p(S = 1|W = 1) \geq 0.6$", then any email that includes "Won" will be classified as Spam.

Probabilistic ML answer

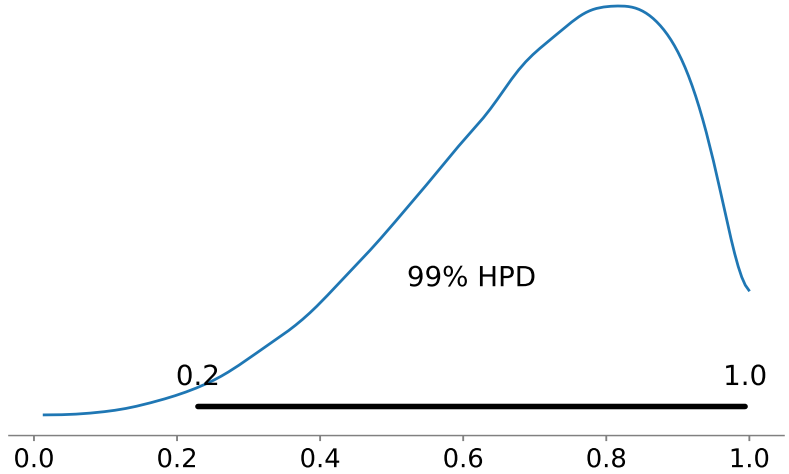


This is the posterior PDF of $p(S = 1|W = 1)$ that is:

$$f(p(S = 1|W = 1)|D) = f\left(\frac{\theta_{W=1|S=1}\theta_{S=1}}{\theta_{W=1|S=1}\theta_{S=1}+\theta_{W=1|S=0}\theta_{S=0}}\middle|D\right)$$

We can compute this probability density function by sampling the parameters $\theta_S, \theta_{W|S}$ from the posterior $f(\theta|D)$ and computing the above fraction for each sample.

Probabilistic ML answer



The posterior in the figure tells us that is more probable that $p(S = 1|W = 1)$ is greater than 0.5 (the majority of the mass is at the right of 0.5), but there is also a significant probability that $p(S = 1|W = 1)$ is less than 0.5.

We can quantify

$$P(p(S = 1|W = 1) < 0.6) = \int_0^{0.6} f\left(\frac{\theta_{W=1|S=1}\theta_{S=1}}{\theta_{W=1|S=1}\theta_{S=1}+\theta_{W=1|S=0}\theta_{S=0}}\middle|D\right) d\theta \approx 0.28$$

this is the area under the curve between 0 and 0.6. This means that $p(S = 1|W = 1)$ could be less than 0.6 with probability 0.28, while MultinomialNB tells us that $p(S = 1|W = 1) = 0.6$ with certainty.
We must take into account this uncertainty when we make decisions.

Conclusions

In a real spam filter, we will use the uncertainty to separate

- emails that are easy to classify (small uncertainty)
- emails that are difficult to classify (large uncertainty).

and we will see that MultinomialNB has a much lower accuracy in the emails we classify as "large uncertainty" (although it can return $p(S = 1|Words) \geq 0.9$ in those cases).