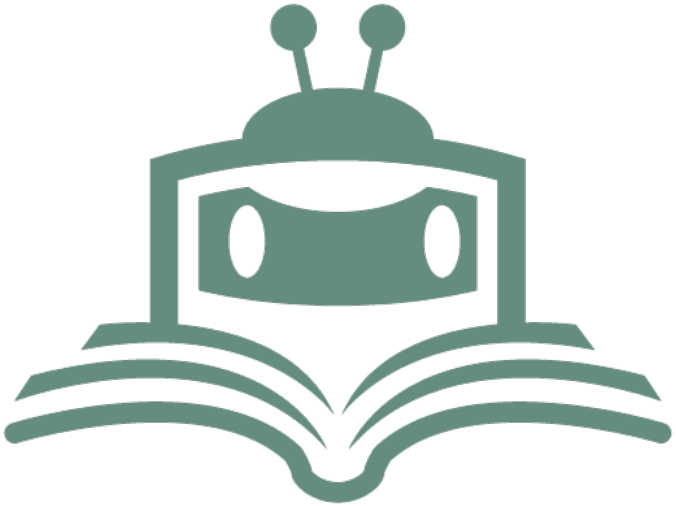

MATHEMATICS FOR MACHINE LEARNING

Nagoya University, Fall 2022

Lecture 6

Naive Bayes I

https://www.henrikbachmann.com/mml2022.html

Email spam filter

Let's assume we want to create an email spam filter by using supervised learning.

Features: Email

$\mathcal{X} =$ Emails

Labels: No Spam & Spam

$\mathcal{Y} = \{0, 1\}$

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

According to this training set: Is the email “Buy book today” spam?

Email spam filter: From emails to vectors

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

Create a dictionary and assign to an email a vector

1	book
2	buy
3	do
4	drugs
5	math
6	today

"Buy book today" $\longleftrightarrow$ $\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \in \mathcal{X} = \{0,1\}^6$

We will ignore the order and the number of appearances of words

Email spam filter

Recall: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

For a given email (feature) x we want to calculate the probability that this email is spam, i.e.

$$P(y = 1 \mid x)$$

Email spam filter

Recall: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

For a given email (feature) x we want to calculate the probability that this email is spam, i.e.

$$P(y = 1 \mid x)$$

We will do this by using **Bayes rule**

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x)}$$

Will be explained in more detail later!

together with the **Naive Bayes assumption** that

$$P(x \mid y) = P(x_1, \ldots, x_d|y) = \prod_{i=1}^{d} P(x_i \mid y)$$

and the **law of total probability**

$$P(x) = P(x \mid y = 1)P(y = 1) + P(x \mid y = 0)P(y = 0)$$

Email spam filter

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

The probabilities according to our training set:

i		appears in non-spam: $P(x_i = 1 \mid y = 0)$	appears in spam: $P(x_i = 1 \mid y = 1)$
1	book		
2	buy		
3	do		
4	drugs		
5	math		
6	today		

$P(y = 0) =$ $P(y = 1) =$

Email spam filter

Training set

	Email text	No spam / Spam
1	Do math today	0
2	Buy	1
3	Buy book	0
4	Today do math drugs	1
5	Buy drugs book today	1

The probabilities according to our training set:

i		appears in non-spam: $P(x_i = 1 \mid y = 0)$	appears in spam: $P(x_i = 1 \mid y = 1)$
1	book	$\frac{1}{2}$	$\frac{1}{3}$
2	buy	$\frac{1}{2}$	$\frac{2}{3}$
3	do	$\frac{1}{2}$	$\frac{1}{3}$
4	drugs	0	$\frac{2}{3}$
5	math	$\frac{1}{2}$	$\frac{1}{3}$
6	today	$\frac{1}{2}$	$\frac{2}{3}$

$$P(y = 0) = \frac{2}{5} \qquad P(y = 1) = \frac{3}{5}$$

Email spam filter

$$P(y=0) = \frac{2}{5} \qquad P(y=1) = \frac{3}{5}$$

$$x = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

i		appears in non-spam: $P(x_i = 1 \mid y = 0)$	appears in spam: $P(x_i = 1 \mid y = 1)$
1	book	$\frac{1}{2}$	$\frac{1}{3}$
2	buy	$\frac{1}{2}$	$\frac{2}{3}$
3	do	$\frac{1}{2}$	$\frac{1}{3}$
4	drugs	0	$\frac{2}{3}$
5	math	$\frac{1}{2}$	$\frac{1}{3}$
6	today	$\frac{1}{2}$	$\frac{2}{3}$

"Buy book today"

$$P(x \mid y = 1)$$

$$P(x \mid y = 0)$$

Email spam filter

x = “Buy book today”

$$\longleftrightarrow \quad \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \in \mathcal{X} = \{0,1\}^6$$

i		appears in non-spam: $P(x_i = 1 \mid y = 0)$	appears in spam: $P(x_i = 1 \mid y = 1)$
1	book	$\frac{1}{2}$	$\frac{1}{3}$
2	buy	$\frac{1}{2}$	$\frac{2}{3}$
3	do	$\frac{1}{2}$	$\frac{1}{3}$
4	drugs	0	$\frac{2}{3}$
5	math	$\frac{1}{2}$	$\frac{1}{3}$
6	today	$\frac{1}{2}$	$\frac{2}{3}$

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x)} \qquad P(y = 0) = \frac{2}{5} \qquad P(y = 1) = \frac{3}{5}$$

$$P(x) = P(x \mid y = 1)P(y = 1) + P(x \mid y = 0)P(y = 0)$$

Generative vs Discriminative learning algorithm

Logistic regression

x: feature (e.g. hours of studying)
y: label (e.g. passing or failing exam)

Want to find a hypothesis which describes P(y|x)

$$P(y = 1 \mid x; \theta) = h_\theta(x) \, .$$

Learning P(y|x) is an example of a **discriminative learning algorithm.**

Generative learning algorithm: Learn P(x|y) and P(y).

Generative vs Discriminative learning algorithm

Notation: P(A|B) refers to the conditional probability that event A occurs, given that event B has occurred.

Generative learning algorithm: Learn P(x | y) and P(y).

Question: But we want P(y | x)… don't we??

Yes, but we can use:

Bayes rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Bayes rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Naive Bayes

An example for a generative learning algorithm: Naive Bayes

Example: Spam filter

- Feature: Email
- Label: Spam & No Spam

$\mathcal{X} =$

$\mathcal{Y} = \{0, 1\}$

Naive Bayes assumption

Naive Bayes assumption:

The features are “conditionally independent” given the label.

conditionally independent

A and B are conditionally independent given C if and only if, given knowledge that C occurs, knowledge of whether A occurs provides no information on the likelihood of B occurring, and knowledge of whether B occurs provides no information on the likelihood of A occurring.

Conditionally independent: Example

A: Ability to do math of a person

B: Foot size of a person

A and B are **not independent**, since if I tell you someone's foot size, it hints at their age, which in turn hints at their ability to do math.

C: Age of a person

A and B are **conditionally independent given C**, since if tell you someone's age (C), then the ability of doing math (A) will not change whether you know the foot size (B) or not, i.e.

$$P(A \mid C) = P(A \mid B,C)$$

Naive Bayes assumption

Naive Bayes assumption:

The features are "conditionally independent" given the label.

If x_1, x_2 are conditionally independent given y, then we have

$$P(x_1 \mid y, x_2) = P(x_1|y)\,.$$

We want to calculate $P(x|y) = P(x_1, \ldots, x_d|y)$.

Chain rule of probabilities: $P(A, B) = P(A|B)P(B)$

Naive Bayes assumption

Naive Bayes assumption:

$$P(x_1 \mid y, x_2) = P(x_1|y)\,.$$

Chain rule of probabilities:

$$P(A, B) = P(A|B)P(B)$$

We want to calculate $P(x|y) = P(x_1, \ldots, x_d|y)$.

Naive Bayes classifier

$$P(x \mid y) = P(x_1, \ldots, x_d|y) = \prod_{i=1}^{d} P(x_i \mid y)$$

Our model is parametrized (the stuff we need to remember after training) by

$$\begin{aligned}
\phi_{i|y=1} &= P(x_i = 1 \mid y = 1)\,, \\
\phi_{i|y=0} &= P(x_i = 1 \mid y = 0)\,, \\
\phi_{y=1} &= P(y = 1)\,.
\end{aligned}$$

By Bayes rule we get for a feature $x \in \mathcal{X}$

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x)}$$

$P(x) = P(x \mid y = 1)P(y = 1) + P(x \mid y = 1)P(y = 1)$

Naive Bayes classifier: Training

$$\phi_{i|y=1} = P(x_i = 1 \mid y = 1)\,,$$
$$\phi_{i|y=0} = P(x_i = 1 \mid y = 0)\,,$$
$$\phi_{y=1} = P(y = 1)\,.$$

Indicator function

$$I(S) = \begin{cases} 1\,, & S \text{ is true} \\ 0\,, & S \text{ is false} \end{cases}.$$

Given a training set $\mathcal{T} = \left((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$ we can calculate them by

$$\phi_{i|y=1} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 1)}{\sum_{j=1}^{n} I(y^{(j)} = 1)}$$

$$\phi_{i|y=0} = \frac{\sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 0)}{\sum_{j=1}^{n} I(y^{(j)} = 0)}$$

$$\phi_{y=1} = \frac{1}{n} \sum_{j=1}^{n} I(y^{(j)} = 1)$$

Naive Bayes classifier: Judging a new email

Training: Determine the values

$$\phi_{i|y=1} = P(x_i = 1 \mid y = 1)\,,$$
$$\phi_{i|y=0} = P(x_i = 1 \mid y = 0)\,,$$
$$\phi_{y=1} = P(y = 1)\,.$$

The probability of a new email $x = (x_1, \ldots, x_d)^T$ being spam is then

$$P(y = 1 \mid x) = \frac{P(x \mid y = 1)P(y = 1)}{P(x)}$$
$$= \frac{\prod_{i=1}^{d} P(x_i \mid y = 1) \cdot \phi_{y=1}}{\prod_{i=1}^{d} P(x_i \mid y = 1) \cdot \phi_{y=1} + \prod_{i=1}^{d} P(x_i \mid y = 0)(1 - \phi_{y=1})}\,.$$

$$P(x_i = 1 \mid y = 1) = \phi_{i|y=1}\,,$$
$$P(x_i = 0 \mid y = 1) = 1 - \phi_{i|y=1}\,,$$

$$P(x_i = 1 \mid y = 0) = \phi_{i|y=0}\,,$$
$$P(x_i = 0 \mid y = 0) = 1 - \phi_{i|y=0}\,.$$

Naive Bayes classifier: Problem

Here are some problems:

What is the probability of “buy anti drug book today” to be spam?

Naive Bayes classifier: Laplace smoothing

We should never assume an event to have probability 0 or 1.

For this we will use **Laplace smoothing** and change our parameters as follows:

$$\widetilde{\phi}_{i|y=1} = \frac{1 + \sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 1)}{2 + \sum_{j=1}^{n} I(y^{(j)} = 1)}$$

$$\widetilde{\phi}_{i|y=0} = \frac{1 + \sum_{j=1}^{n} I(x_i^{(j)} = 1 \wedge y^{(j)} = 0)}{2 + \sum_{j=1}^{n} I(y^{(j)} = 0)}$$

Possible interpretation: Assume that every word appeared & did not appear in a spam and non-spam email.