

Naive Bayes

A very simple dataset – one field / one class

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

A new patient has
a blood test – his P34
level is HIGH.

what is our best guess
for prostate cancer?

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

It's useful to know:
 $P(\text{cancer} = Y)$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

It's useful to know:

$P(\text{cancer} = Y)$

- on basis of this tiny dataset, $P(c = Y)$ is $5/10 = 0.5$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

It's useful to know:

$P(\text{cancer} = Y)$

- on basis of this tiny dataset, $P(c = Y)$ is $5/10 = 0.5$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

So, with **no other info** you'd expect $P(\text{cancer}=Y)$ to be 0.5

A very simple dataset – one field / one class

But we know that $P34 = H$,
so actually we want:

$$P(\text{cancer}=Y \mid P34 = H)$$

- the prob that cancer is Y,
given that P34 is high

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

$$P(\text{cancer}=\text{Y} \mid \text{P34} = \text{H})$$

- the prob that cancer is Y,
given that P34 is high

- this seems to be
 $2/3 \approx 0.67$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

A very simple dataset – one field / one class

So we have:

$$P(c=Y \mid \mathbf{P34} = \mathbf{H}) = 0.67$$

$$P(c=N \mid \mathbf{P34} = \mathbf{H}) = 0.33$$

The class value with the
highest probability is our
best guess

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

In general we may have any number of class values

suppose again we know that
P34 is High;
here we have:

$$P(c=Y \mid \mathbf{P34} = \mathbf{H}) = 0.5$$

$$P(c=N \mid \mathbf{P34} = \mathbf{H}) = 0.25$$

$$P(c = \text{Maybe} \mid \mathbf{H}) = 0.25$$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
High	Maybe
Medium	Y

... and again, Y is the winner

That is the essence
of Naive Bayes,
but:

the probability calculations are much
trickier when there are >1 fields

so we make a 'Naive' assumption that
makes it simpler

Bayes' theorem

As we saw, on the right
we are illustrating:

$$P(\text{cancer} = Y \mid \text{P34} = H)$$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

Bayes' theorem

And now we are illustrating

$$P(\text{P34} = \text{H} \mid \text{cancer} = \text{Y})$$

This is a different thing,
that turns out as $2/5 = 0.4$

P34 level	Prostate cancer
High	Y
Medium	Y
Low	Y
Low	N
Low	N
Medium	N
High	Y
High	N
Low	N
Medium	Y

Bayes' theorem is this:

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

It is very useful when it is hard to get $P(A | B)$ directly, but easier to get the things on the right

Bayes' theorem in 1-non-class-field DMML context:

$$P(\text{Class}=\text{X} \mid \text{Fieldval} = \text{F}) =$$

$$\frac{P(\text{Fieldval} = \text{F} \mid \text{Class} = \text{X}) \times P(\text{Class} = \text{X})}{P(\text{Fieldval} = \text{F})}$$

Bayes' theorem in 1-non-class-field DMML context:

$$P(\text{Class}=\text{X} \mid \text{Fieldval} = \text{F}) =$$

$$\frac{P(\text{Fieldval} = \text{F} \mid \text{Class} = \text{X}) \times P(\text{Class} = \text{X})}{P(\text{Fieldval} = \text{F})}$$

We want to check this for each class and choose the class that gives the highest value.

Bayes' theorem in 1-non-class-field DMML context:

$$P(\text{Class}=\text{X} \mid \text{Fieldval} = \text{F}) =$$

$$P(\text{Fieldval} = \text{F} \mid \text{Class} = \text{X}) \times \frac{P(\text{Class} = \text{X})}{P(\text{Fieldval} = \text{F})}$$

E.g. We compare:

$$P(\text{Fieldval} \mid \text{Yes}) \times P(\text{Yes})$$
$$P(\text{Fieldval} \mid \text{No}) \times P(\text{No})$$
$$P(\text{Fieldval} \mid \text{Maybe}) \times P(\text{Maybe})$$

... we can ignore “ $P(\text{Fieldval} = \text{F})$ ” ... why ?

and that was
Exactly how we do
Naive Bayes for a
1-field dataset

Deriving NB

Essence of Naive Bayes, with 1 non-class field, is to calc *this* for each class value, given some new instance with fieldval = F:

$$P(\text{class} = C \mid \text{Fieldval} = F)$$

For many fields, our new instance is (e.g.) (F1, F2, ...Fn), and the 'essence of Naive Bayes' is to calculate *this* for each class:

$$P(\text{class} = C \mid F1, F2, F3, \dots, Fn)$$

i.e. What is prob of class C, given all these field vals together?

Apply magic dust and Bayes theorem, and ...

... *If we make the **naïve assumption** that all of the fields are **independent** of each other (e.g. $P(F1 | F2) = P(F1)$, etc ...) ... then*

$$\begin{aligned} & P(\text{class} = C | F1 \text{ and } F2 \text{ and } F3 \text{ and } \dots F_n) \\ = & P(F1 \text{ and } F2 \text{ and } \dots \text{ and } F_n | C) \times P(C) \\ = & P(F1 | C) \times P(F2 | C) \times \dots \times P(F_n | C) \times P(C) \end{aligned}$$

... which is what we calculate in NB

Nave-Bayes -- in general

N fields, q possible class values, New unclassified instance: $F1 = v1, F2 = v2, \dots, Fn = vn$

what is the class value? i.e. Is it $c1, c2, \dots$ or cq ?

calculate each of these q things – biggest one gives the class:

$$\begin{aligned} &P(F1=v1 \mid c1) \times P(F2=v2 \mid c1) \times \dots \times P(Fn=vn \mid c1) \times P(c1) \\ &P(F1=v1 \mid c2) \times P(F2=v2 \mid c2) \times \dots \times P(Fn=vn \mid c2) \times P(c2) \\ &\dots \\ &P(F1=v1 \mid cq) \times P(F2=v2 \mid cq) \times \dots \times P(Fn=vn \mid cq) \times P(cq) \end{aligned}$$

Nave-Bayes with Many-fields

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$P(p34=M \mid Y) \times P(p61=M \mid Y) \times P(BMI=H \mid Y) \times P(\text{cancer} = Y)$$

$$P(p34=M \mid N) \times P(p61=M \mid N) \times P(BMI=H \mid N) \times P(\text{cancer} = N)$$

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$P(\mathbf{p34=M} \mid \mathbf{Y}) \times P(\mathbf{p61=M} \mid \mathbf{Y}) \times P(\mathbf{BMI=H} \mid \mathbf{Y}) \times P(\mathbf{cancer = Y})$$

$$P(\mathbf{p34=M} \mid \mathbf{N}) \times P(\mathbf{p61=M} \mid \mathbf{N}) \times P(\mathbf{BMI=H} \mid \mathbf{N}) \times P(\mathbf{cancer = N})$$

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$P(p34=M \mid Y) \times \textcolor{red}{P(p61=M \mid Y)} \times P(BMI=H \mid Y) \times P(\text{cancer} = Y)$$

$$P(p34=M \mid N) \times P(p61=M \mid N) \times P(BMI=H \mid N) \times P(\text{cancer} = N)$$

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$P(p34=M | Y) \times P(p61=M | Y) \times \mathbf{P(BMI=H | Y)} \times P(\text{cancer} = Y)$$

$$P(p34=M | N) \times P(p61=M | N) \times P(BMI=H | N) \times P(\text{cancer} = N)$$

Nave-Bayes with

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$P(p34=M \mid Y) \times P(p61=M \mid Y) \times P(BMI=H \mid Y) \times \textcolor{red}{P(\text{cancer} = Y)}$$

$$P(p34=M \mid N) \times P(p61=M \mid N) \times P(BMI=H \mid N) \times P(\text{cancer} = N)$$

Nave-Bayes with Many-fields

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$$\begin{array}{ccccccc}
 0.4 & \times 0 & \times 0.4 & \times 0.5 = 0 \\
 0.2 & \times 0.4 & \times 0.2 & \times 0.5 = 0.008
 \end{array}$$

In practice, we finesse the zeroes and use logs:
 (note: $\log(A \times B \times C \times D \times \dots) = \log(A) + \log(B) + \dots$)

New patient:

P34=M, P61=M, BMI = H

Best guess at cancer field ?

which of these gives the
highest value?

P34 level	P61 level	BMI	Prostate cancer
High	Low	Medium	Y
Medium	Low	Medium	Y
Low	Low	High	Y
Low	High	Low	N
Low	Low	Low	N
Medium	Medium	Low	N
High	Low	Medium	Y
High	Medium	Low	N
Low	Low	High	N
Medium	High	High	Y

$\log(0.4)$	$+ \log(0.001)$	$+ \log(0.4)$	$+ \log(0.5) = -4.09$
$\log(0.2)$	$+ \log(0.4)$	$+ \log(0.2)$	$+ \log(0.5) = -2.09$

Nave-Bayes -- in general

As indicated, what we normally do, when there are *more than a handful of fields*, is this

Calculate:

$$\log(P(F1=v1 \mid c1)) + \dots + \log(P(Fn=vn \mid c1)) + \log(P(c1))$$

$$\log(P(F1=v1 \mid c2)) + \dots + \log(P(Fn=vn \mid c2)) + \log(P(c2))$$

and choose class based on highest of these.

Because ... ?

Table 8.1 Class-Labeled Training Tuples from the *AllElectronics* Customer Database

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$X = (age = youth, income = medium, student = yes, credit_rating = fair)$

Predict if Bob will default his loan

Bob

Home owner: *No*

Marital status: *Married*

Job experience: *3*

Home owner	Marital Status	Job experience (1-5)	Defaulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes