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

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

what is our best guess for prostate cancer?

A very simple dataset – one field / one class

It’s useful to know:
P(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(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

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

It's useful to know:
P(cancer = Y)

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

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

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

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

P(cancer=Y | P34 = H)

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

A very simple dataset – one field / one class

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

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

- this seems to be $2/3 = \sim 0.67$

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 | **P34 = H**) = 0.67
P (c =N | **P34 = 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 | **P34 = H**) = 0.5

P (c=N | **P34 = H**) = 0.25

P(c = Maybe | **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:

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

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:

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

It is very useful when it is hard to get $\mathrm{P}(\mathrm{A} \mid \mathrm{B})$ directly, but easier to get the things on the right

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

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

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

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

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

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

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:

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

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

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*(Fieldval = 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(class = C | 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(class = C | F1,F2,F3,...,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* ***naive*** *assumption that all of the fields are independent of each other*

(*e.g.* P(F1| F2) = P(F1), etc ...) ... then

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

$$= P(\text{ F1 and F2 and} \ldots \text{and Fn} \mid \text{C}) \times P\,(\text{C})$$

$$= \boxed{P(\text{F1} \mid \text{C}) \times P\,(\text{F2} \mid \text{C}) \times \ldots \times P(\text{Fn} \mid \text{C}) \times P(\text{C})}$$

... which is what we calculate in NB

Nave-Bayes -- in general

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

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

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

P(F1=v1 | c1) × P(F2=v2 | c1) × ... × P(Fn=vn | c1) × P(c1)

P(F1=v1 | c2) × P(F2=v2 | c2) × ... × P(Fn=vn | c2) × P(c2)

...

P(F1=v1 | cq) × P(F2=v2 | cq) × ... × P(Fn=vn | cq) × P(cq)

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

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

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

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

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

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

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

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

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

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

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

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

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**

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

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

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

$0.4 \times 0 \times 0.4 \times 0.5 = 0$

$0.2 \times 0.4 \times 0.2 \times 0.5 = 0.008$

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

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:

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

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

and choose class based on highest of these.

Because … ?

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

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
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