Naïve Bayes classifier

Lecture 5

Mathematical predictions

- We can ‘predict’ where the spacecraft will be at noon in 2 months from now
- We cannot predict where you will be tomorrow at noon
- But, based on numerous observations, we can estimate the probability

Bayesian beliefs

- How do we judge that something is true?
- Can mathematics help make judgments more accurate?
- Bayes: our believes should be updated as new evidences become available

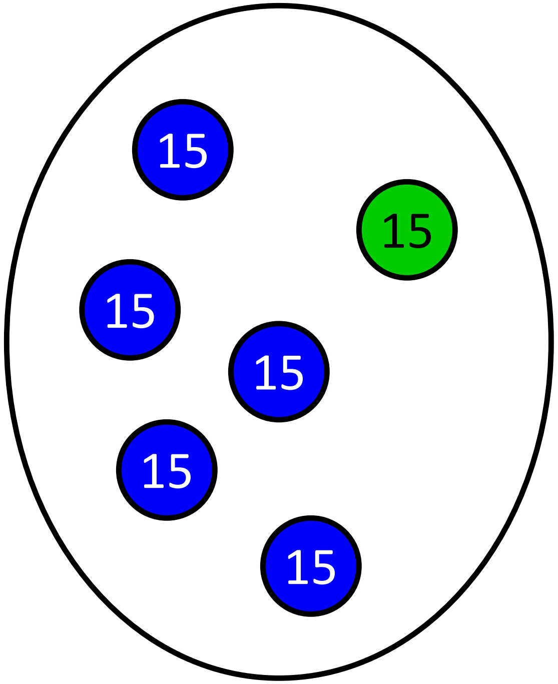
Bayes' method

- There are 2 events: **A** and not A (**B**) which you believe occur with probabilities P(**A**) and P(**B**). Estimation P(**A**):P(**B**) represents odds of A vs. B.
- Collect evidence data **E.**
- Re-estimate P(**A**|**E**):P(**B**|**E**) and update your beliefs.

Example (fictitious): hit-and-run

- 75 blue cabs (**B**) and 15 green cabs (**G**)
- P(**B**):P(**G**)=5:1

- At night: hit-and-run episode
- Witness: “I saw a green cab”: X_G
- Witness is tested at night conditions: identifies correct color 4 times out of 5



- Question: what is more probable:

B or **G**

?

Adopted from: **The numbers behind NUMB3RS:** solving crime with mathematics by Devlin and Lorden.

Probability

- Basic element: random variable

  e.g., *Car* is one of *<blue, ¬blue(green)>*

  Weather is one of *<sunny,rainy,cloudy,snow>*

- Both *Car* and *Weather* are discrete random variables
  - Domain values must be
    - exhaustive (blue and green – are all the cabs)
    - mutually exclusive (green is always not blue, there are no cars which are half green, half blue)

- Elementary propositions are constructed by the assignment of a value to a random variable:

  e.g., *Car = ¬blue*,

  Weather = sunny

Conditional probability

- $P(A|B)$ – probability of event A given that event B has happened
- In our case we want to compare:

the car was **G** given a witness testimony X_G: $P(\mathbf{G}|\mathbf{X_G})$

vs.

the car was **B** given a witness testimony X_G: $P(\mathbf{B}|\mathbf{X_G})$

Prior probability and distribution

- Prior or unconditional probability associated with a proposition is the degree of belief accorded to it in the absence of any other information.

  e.g.,

  P(Car = blue) = 0.83 (or abbrev. P(blue) = 0.83)

  P(Weather = sunny) = 0.7 (or abbrev. P(sunny) = 0.7)

- Probability distribution gives probabilities of all possible value assignments:

  P(Weather = sunny) = 0.7

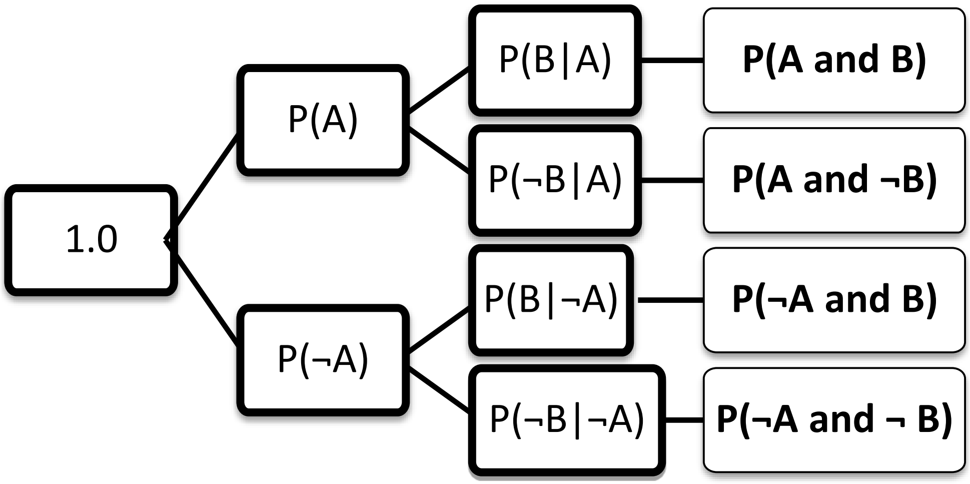
  P(Weather = rain) = 0.2

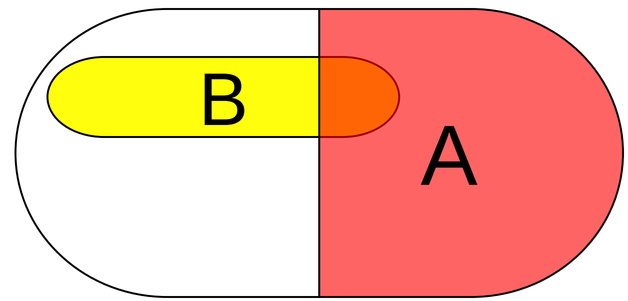
  P(Weather = cloudy) = 0.08

  P(Weather = snow) = 0.02

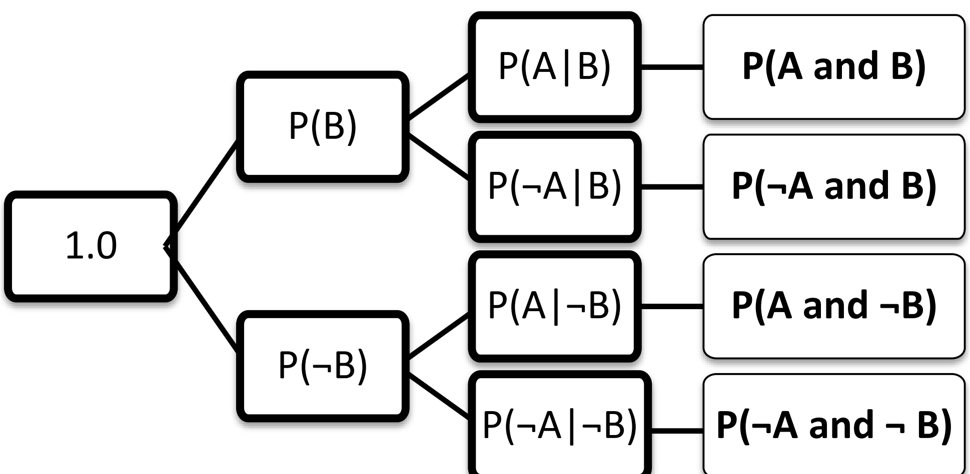
  \- Sums up to 1.0

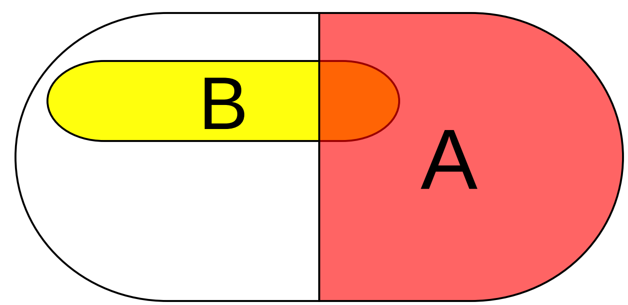
Two random events (not independent) happen at the same time – P(A and B)





Possible event combinations when we know the outcome of event A:
P(B|A)=1/12 and P(A)=1/2

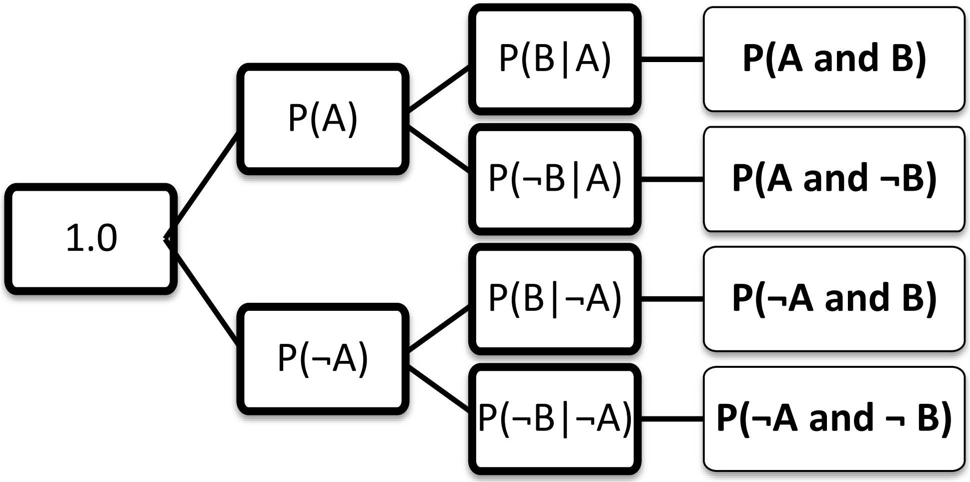




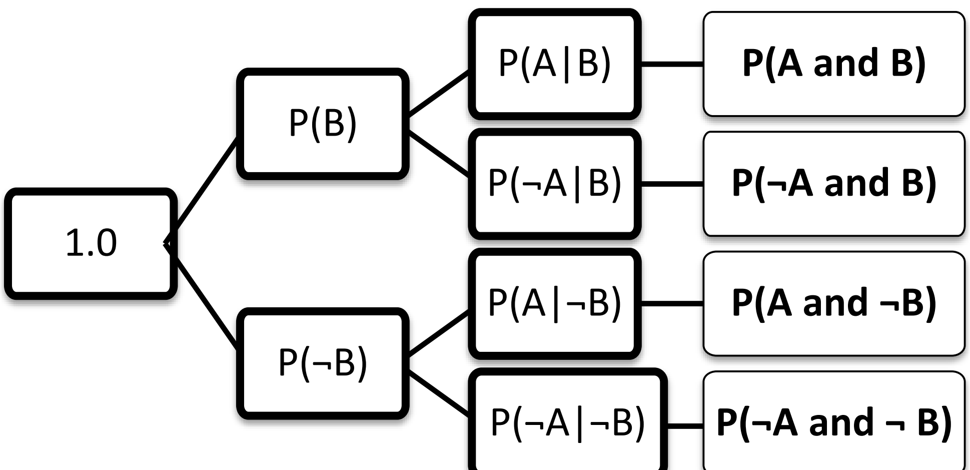
Possible event combinations when we know the outcome of event B:
P(A|B)=1/4 and P(B)=1/6

But in both cases P(A and B) is the same: orange area in the diagram

Intuition for Bayes's theorem



→ P(A and B)=P(A)*P(B|A)

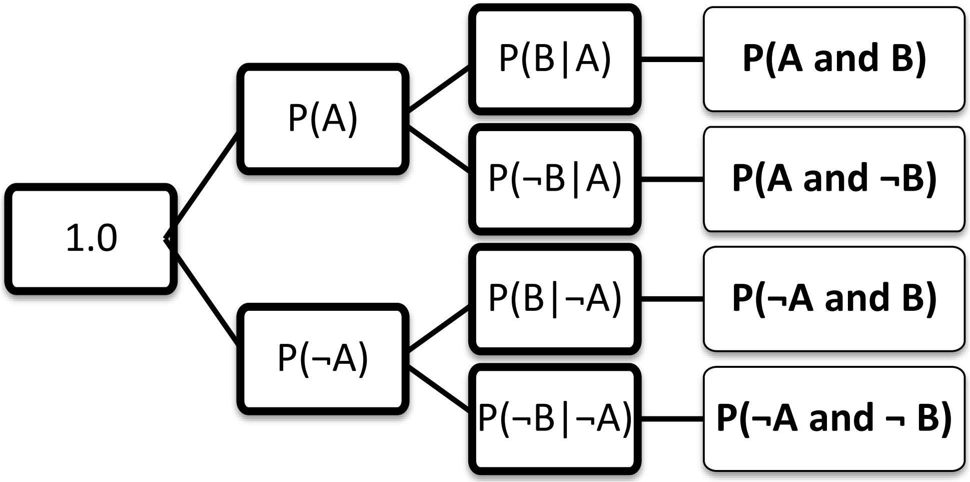


→ P(A and B)=P(B)*P(A|B)

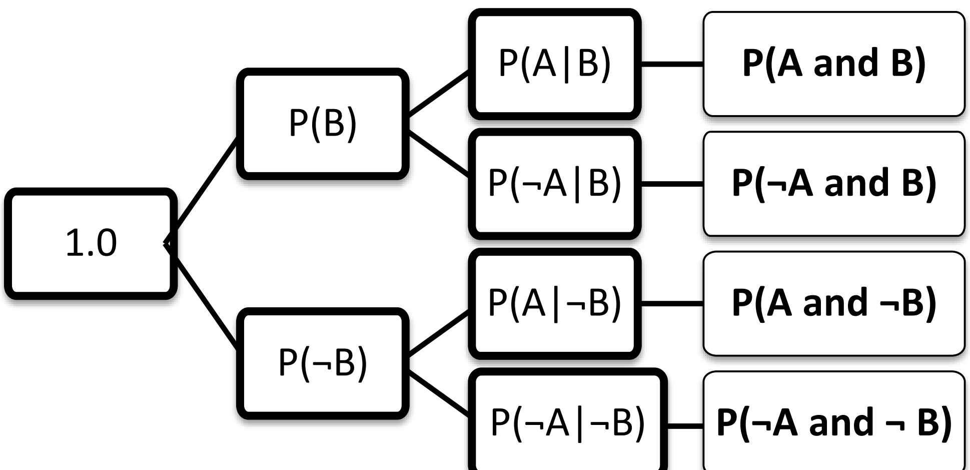
P(A and B)=P(A)*P(B|A)=P(B)*P(A|B)

P(¬A and B)=P(¬A)*P(B|¬A)=P(B)*P(¬A|B)

Bayes' theorem



→ P(A and B)=P(A)*P(B|A)

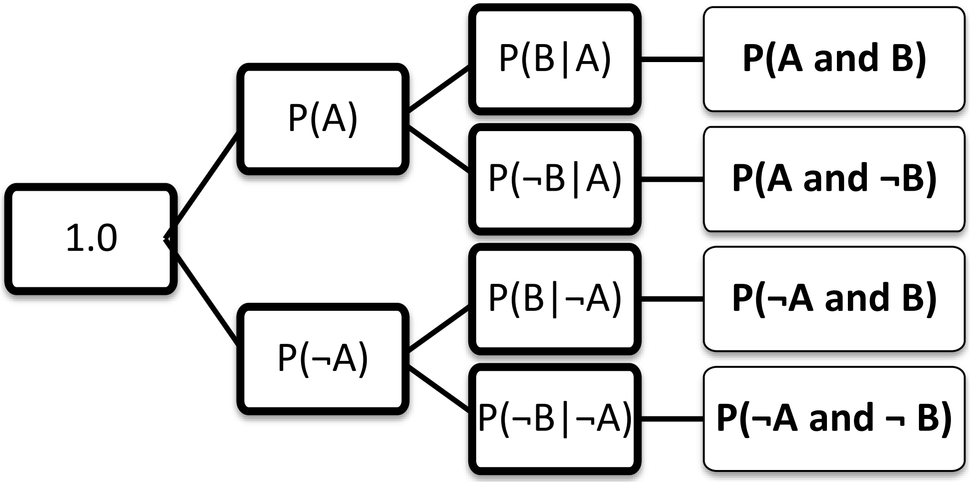


→ P(A and B)=P(B)*P(A|B)

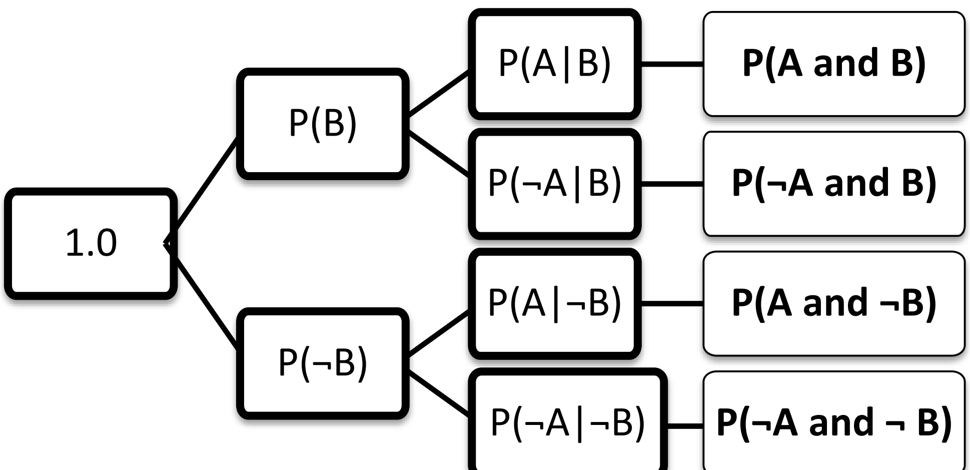
P(A)*P(B|A)=P(B)*P(A|B)

P(¬A)*P(B|¬A)=P(B)*P(¬A|B)

In other words:



→ P(A and B)=P(A)*P(B|A)



→ P(A and B)=P(B)*P(A|B)

P(A|B)=P(A)*P(B|A)/P(B)

P(¬A|B)=P(¬A)*P(B|¬A)/P(B)

Bayes' Rule for updating beliefs

$$\boxed{P(A|B)=P(A)*P(B|A)/P(B)}$$

$$\boxed{P(\neg A|B)=P(\neg A)*P(B|\neg A)/P(B)}$$

- We want to compare $P(A|B)$ and $P(\neg A|B)$, i.e. given evidence B what probability is higher: that A occurred or that ¬A occurred?

- We know $P(A)$ and $P(\neg A)$ – prior probabilities
- We know $P(B|A)$ and $P(B|\neg A)$

- From Bayes' theorem:

$$\mathbf{P(A|B)} = P(A)*P(B|A) / P(B)$$

$$\mathbf{P(\neg A|B)} = P(\neg A)*P(B|\neg A) / P(B)$$

Back to hit-and-run

What is more probable: **B** or **G** ?

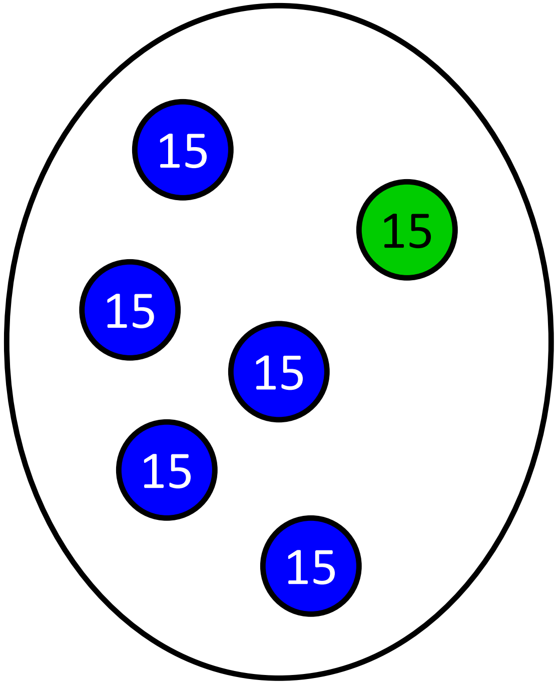
- All cabs were on the streets:

Prior probabilities: P(**B**) =5/6, P(**G**) = 1/6

- The eyewitness test has shown:

$P(X_G \mid G) = 4/5$ (correctly identified)

$P(X_G \mid B) = 1/5$ (incorrectly identified)



$\mathbf{P(G \mid X_G)} = P(G) * P(X_G \mid G) / P(X_G)$

$\mathbf{P(\neg G \mid X_G)} = P(\neg G) * P(X_G \mid \neg G) / P(X_G)$

Bayes rule

Hit-and-run: solution

$P(B) = 5/6$, $P(G) = 1/6$

$P(X_G \mid G) = 4/5$ $P(X_G \mid B) = 1/5$

- Probability that car was **green** given the evidence X_G:

$P(G|X_G) = P(G) * P(X_G|G) / P(X_G) = [1/6 * 4/5] / P(X_G) = 4/30P(X_G)$

//- 4 parts of $30P(X_G)$

- Probability that car was **blue** given the evidence X_G:

$P(B|X_G) = P(B) * P(X_G|B) / P(X_G) = [5/6 * 1/5] / P(X_G) = 6/30P(X_G)$

//- 6 parts of $30P(X_G)$

<u>6:4 odds that the car was **B**!</u>

Probabilistic classifier

- Given the evidence (data),

can we certainly derive

the diagnostic rule:

if Toothache=true then Cavity=true ?

Name	Toothache	…	Cavity
Smith	true	…	true
Mike	true	…	true
Mary	false	…	true
Quincy	true	…	false
…	…	…	…

- This rule isn't right always.
  - Not all patients with toothache have cavities; some of them have gum disease, an abscess, etc.

- We could try an inverted rule:

if Cavity=true then Toothache=true

- But this rule isn't necessarily right either; not all cavities cause pain.

Certainty and Probability

- The connection between toothaches and cavities is not a certain logical consequence in either direction.

- However, we can provide a **probability** that given an evidence (toothache) the patient has cavity.

- For this we need to know:
  - Prior probability of having cavity: how many times dentist patients had cavities: P(cavity)
  - The number of times that the evidence (toothache) was observed among all cavity patients: P(toothache|cavity)

Bayes' Rule
for diagnostic probability

Bayes' rule:

$$\boxed{P(A|B)=P(A)*P(B|A)/P(B)}$$

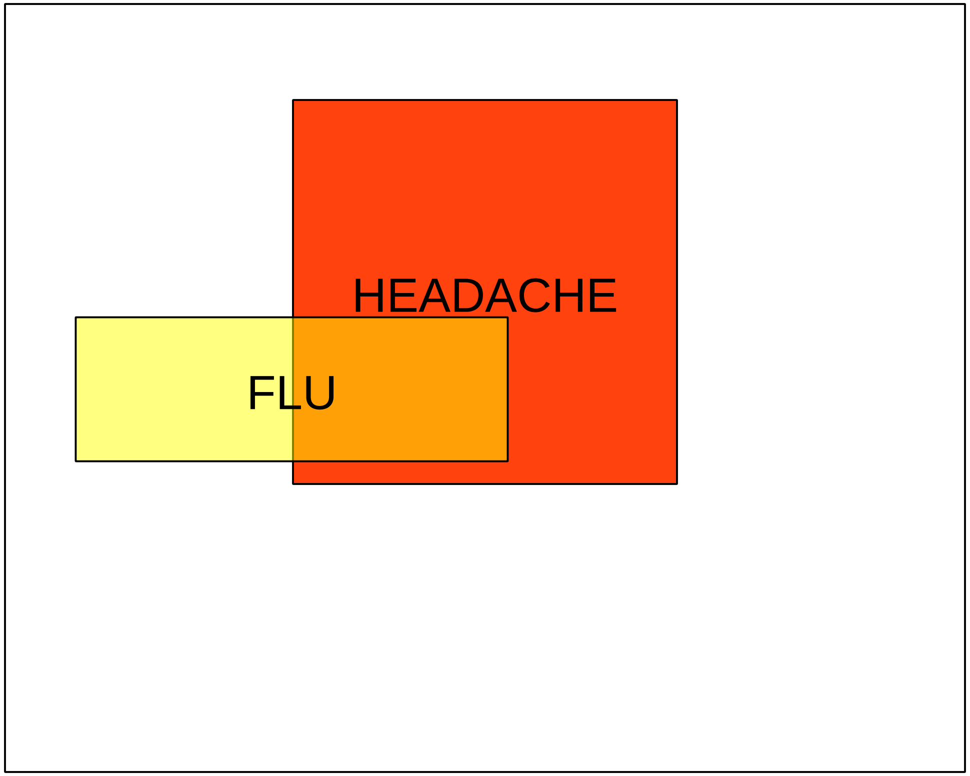
- Useful for assessing **diagnostic** probability from **symptomatic** probability as:

$$P(\text{Cause}|\text{Symptom}) = P(\text{Symptom}|\text{Cause})\, P(\text{Cause}) / P(\text{Symptom})$$

- Bayes's rule is useful in practice because there are many cases where we do have good probability estimates for these three numbers and need to compute the fourth.

Bayes rule application. Example 1

$P(H)=1/10$
$P(F)=1/40$
$P(H|F)=1/2$

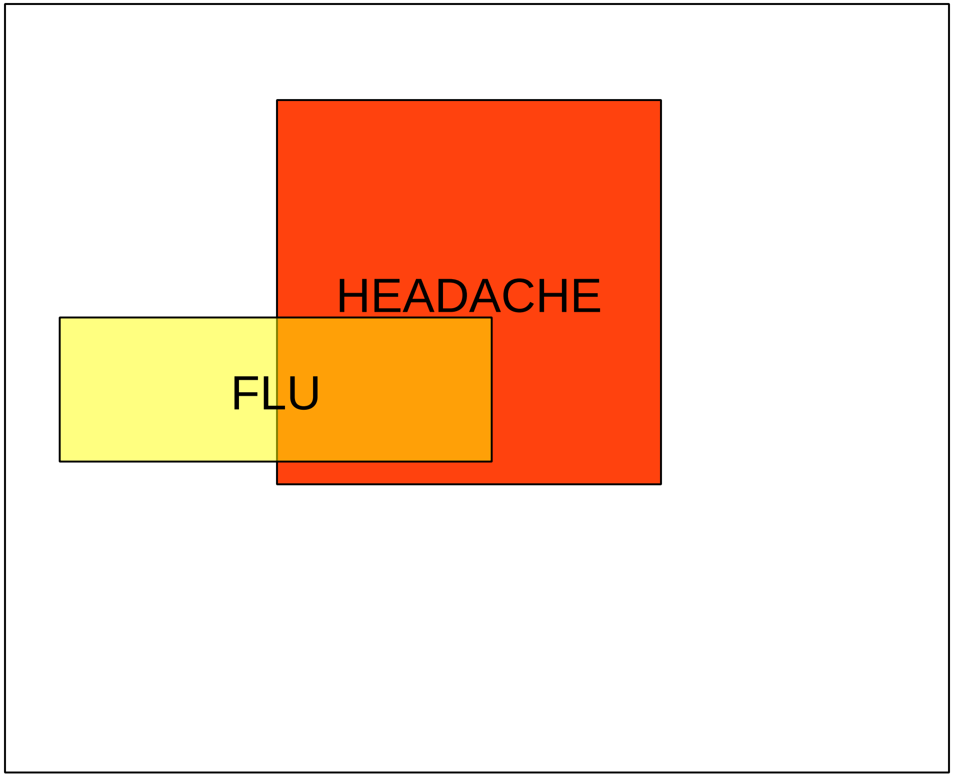


$P(F|H) = ?$

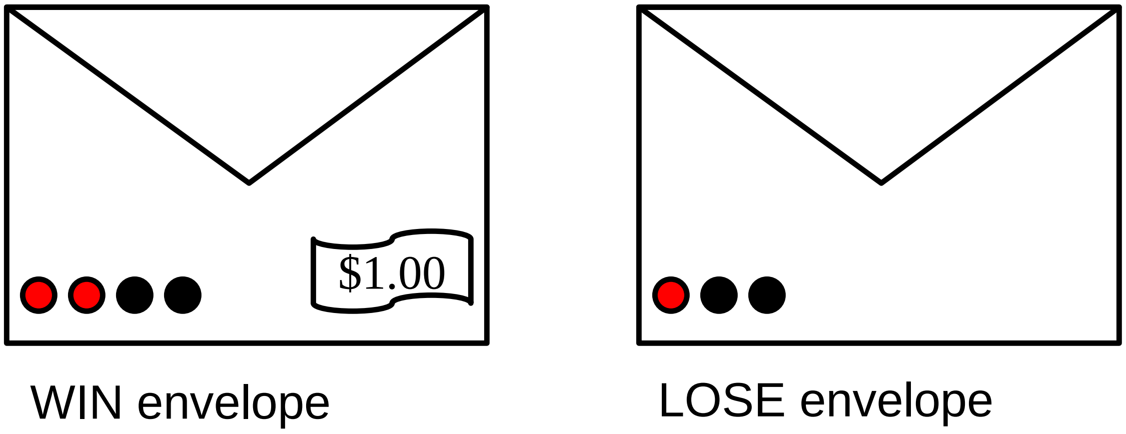
Bayes rule application. Example 1

$P(H)=1/10$
$P(F)=1/40$
$P(H|F)=1/2$

$P(F|H) = P(H|F)P(F)/P(H)$
$=1/2*1/40*10=1/8$



Bayes rule application. Example 2

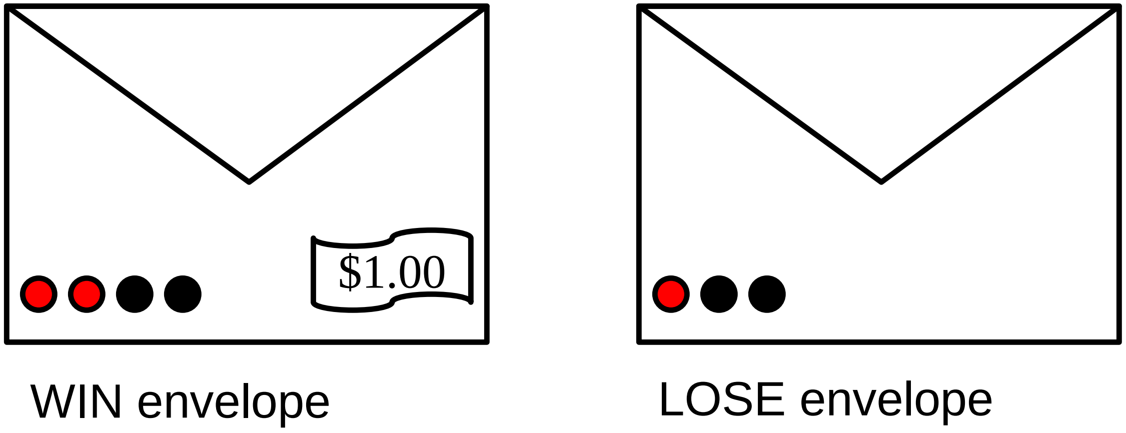


Someone draws an envelope at random and offers to sell it to you.

How much should you pay?

The probability to win is 1:1. Pay no more than 50c.

Bayes rule application. Example 2

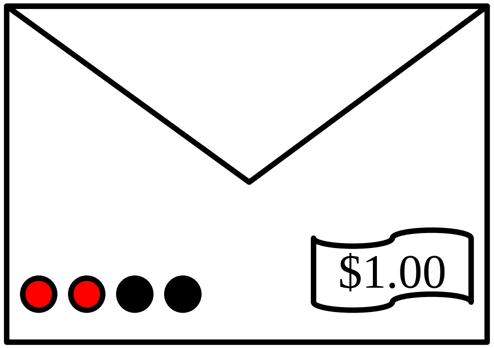


Variant: before deciding, you are allowed to see one bead drawn from the envelope.

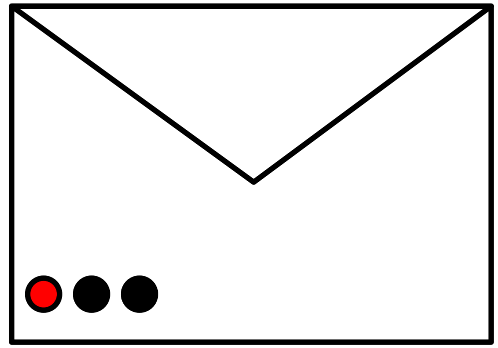
Suppose it's black: How much should you pay?

Suppose it's red: How much should you pay?

Bayes rule application. Example 2



WIN envelope



LOSE envelope

Variant: before deciding, you are allowed to see one bead drawn from the envelope.

Suppose it's black: How much should you pay?

P(W|b)=P(b|W)P(W)/P(b) =(1/2*1/2)/P(b)=1/4 *1/P(b)

P(L|b)=P(b|L)P(L)/P(b)=(2/3*1/2)/P(b) = 1/3 * 1/P(b)

Probability to win is now 3:4 – pay not more than $(3/7)

Suppose it's red: How much should you pay? – the same logic

Classifier based on Bayes rule

- We can build a classifier which will classify a new record as class C (yes or no) by comparing probabilities
- In this case all the attributes except C are evidences E
- The data-related task is to evaluate $P(E|C)$ from historical data and based on $P(E|C)$ and prior probabilities $P(C=Yes)$ and $P(C=No)$ compare $P(C=Yes|E)$ and $P(C=No|E)$ using Bayes rule.

Single-evidence classifier: priors

event (class)

Humidity	Play
High	No
High	No
High	Yes
High	Yes
Normal	Yes
Normal	No
Normal	Yes
High	No
Normal	Yes
Normal	Yes
Normal	Yes
High	Yes
Normal	Yes
High	No

- Prior probabilities:
  P(Play=yes)=9/14, P(play=no)=5/14
- From recording only 'play'/'not play' we have 5:9 odds for play to be canceled today

Single-evidence classifier: evidence

evidence	event (class)
Humidity	**Play**
High	No
High	No
High	Yes
High	Yes
Normal	Yes
Normal	No
Normal	Yes
High	No
Normal	Yes
Normal	Yes
Normal	Yes
High	Yes
Normal	Yes
High	No

- Priors: P(Play=yes)=9/14, P(play=no)=5/14
- After adding evidence about Humidity we have:

How many times Humidity=normal out of all 9 Yes's: 6

P(normal|yes)=6/9

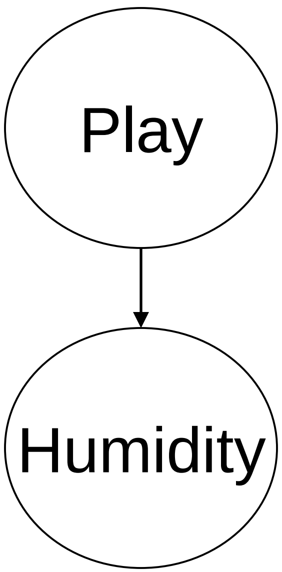
How many times Humidity=normal out of all 5 No's: 1

P(normal|no)=1/5

- Similarly:

P(high|yes)=3/9

P(high|no)=4/5



Single-evidence classifier: prediction

Humidity (evidence)	Play (event (class))
High	No
High	No
High	Yes
High	Yes
Normal	Yes
Normal	No
Normal	Yes
High	No
Normal	Yes
Normal	Yes
Normal	Yes
High	Yes
Normal	Yes
High	No

- P(yes)=9/14, P(no)=5/14
- P(high|yes)=3/9
- P(high|no)=4/5

Today is a high humidity day, what is the probability to play?

- P(yes|high)=P(yes)*P(high|yes)/P(high)
- P(no|high)=P(no)*P(high|no)/P(high)

Single-evidence classifier: prediction

evidence **Humidity**	event (class) **Play**
High	No
High	No
High	Yes
High	Yes
Normal	Yes
Normal	No
Normal	Yes
High	No
Normal	Yes
Normal	Yes
Normal	Yes
High	Yes
Normal	Yes
High	No

P(yes)=9/14, P(no)=5/14

P(high|yes)=3/9

P(high|no)=4/5

Today is a high humidity day, what is the probability to play?

P(yes|high)=P(yes)*P(high|yes)/P(high) = [9/14*3/9] * 1/P(high) = 3/14 α

P(no|high)=P(no)*P(high|no)/P(high) = [5/14*4/5] * 1/P(high) = 4/14 α

4:3 odds not to play given high humidity

(vs. 5:9 before evidence)

Bayes' rule – two evidences

Given that evidence1 is independent of evidence2:

$$P(\text{class} = A|\text{evidence1}, \text{evidence2})$$

$$= \frac{P(\text{evidence1}|\text{class=A}) * P(\text{evidence2}|\text{class=A}) * P(\text{class=A})}{P(\text{evidence1}) * P(\text{evidence2})}$$

$$= \propto P(\text{evidence1}|\text{class} = A) * P(\text{evidence2}|\text{class} = A) * P(\text{class} = A)$$

The same – let's call it 1/α

$$P(\text{class} = B|\text{evidence1}, \text{evidence2})$$

$$= \frac{P(\text{evidence1}|\text{class=B}) * P(\text{evidence2}|\text{class=B}) * P(\text{class=B})}{P(\text{evidence1}) * P(\text{evidence2})}$$

$$= \propto P(\text{evidence1}|\text{class} = B) * P(\text{evidence2}|\text{class} = B) * P(\text{class} = B)$$

Bayes' rule – multiple evidences

Generalized for N evidences

$$P(\text{class} = A \mid \text{evidence1}, \text{evidence2}, \ldots, \text{evidenceN})$$

$$= \frac{P(\text{evidence1} \mid \text{class=A}) * \cdots * P(\text{evidence}N \mid \text{class=A}) * P(\text{class=A})}{P(\text{evidence1}) * \cdots * P(\text{evidence}N)}$$

$$= \propto P(\text{evidence1} \mid \text{class} = A) * \cdots * (\text{evidenceN} \mid \text{class} = A) * P(\text{class} = A)$$

- Two assumptions:

  Attributes (evidences) are:

  – equally important

  – conditionally independent (given the class value)

- This means that knowledge about the value of a particular attribute doesn't tell us anything about the value of another attribute given the class value

Naïve Bayes classifier

To predict class value for a set of attribute values (evidences) for each class value compute and compare:

$$P(\text{class} = A|\text{evidence1}, \text{evidence2}, \ldots, \text{evidenceN})$$

$$= \frac{P(\text{evidence1}|\text{class=A}) * \cdots * P(\text{evidence}N|\text{class=A}) * P(\text{class=A})}{P(\text{evidence1}) * \cdots * P(\text{evidence}N)}$$

$$= \propto P(\text{evidence1}|\text{class} = A) * \cdots * (\text{evidenceN}|\text{class} = A) * P(\text{class} = A)$$

- Naïve – assumes independence of variables
- Although based on assumptions that are almost never correct, this scheme works well in practice!

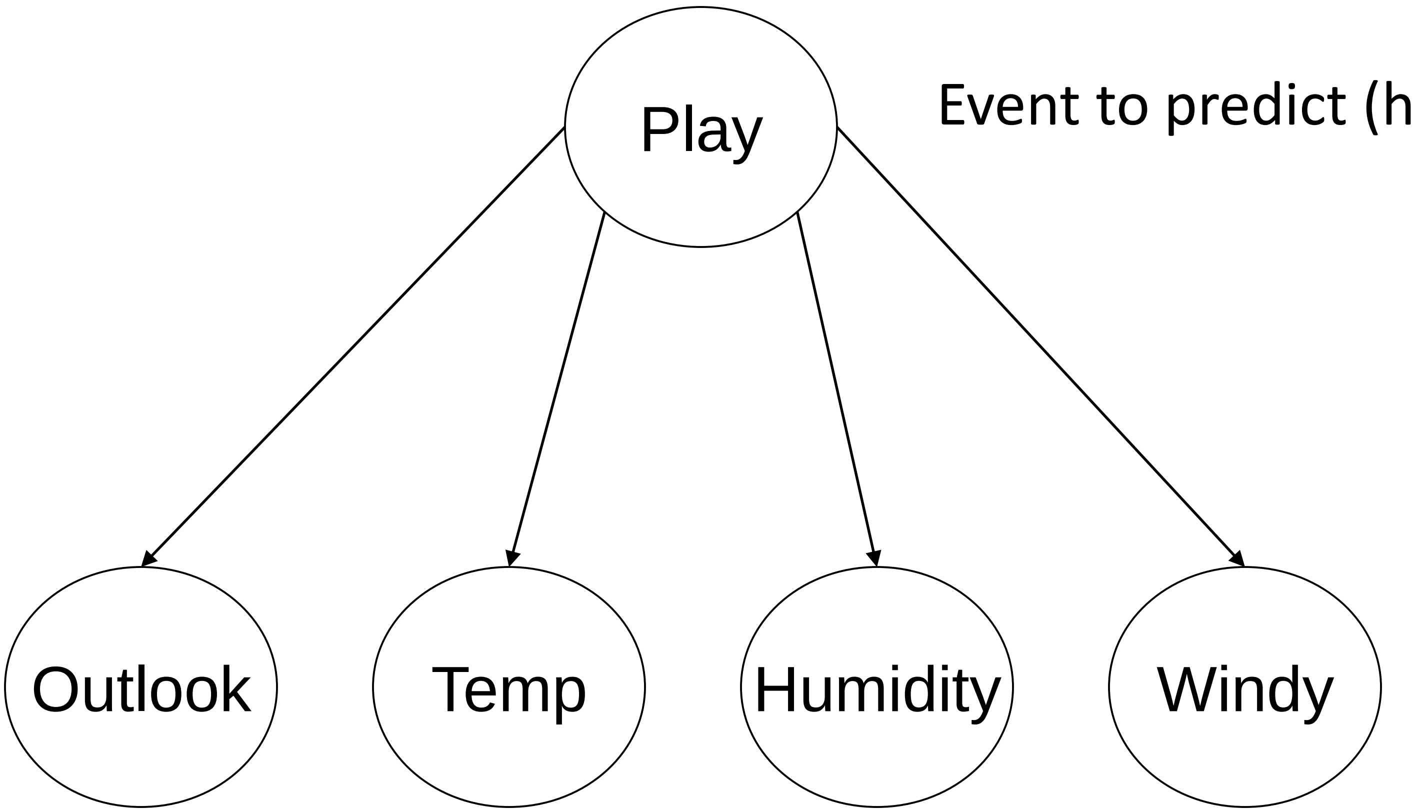
The weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- A new day:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Multi-evidence classifier



Event to predict (hidden)

Set of evidences (demonstrate themselves)

The weather data example: probabilities

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

Play	Sunny	Cool	High humidity	Windy= true
Yes: 9	**2**/9	**3**/9	**3**/9	**3**/9
No: 5	**3**/5	**1**/5	**4**/5	**3**/5
Total	**5**	**4**	**7**	**6**

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

The weather data example: yes

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

P(yes | E) =
P(Sunny | yes) *
P(Cool | yes) *
P(Humidity=High | yes) *
P(Windy=True | yes) *
P(yes) / P(E) =
= (2/9) *
(3/9) *
(3/9) *
(3/9) *
(9/14) / P(E) = 0.0053 / P(E)

Play	Sunny	Cool	High humidity	Windy= true
Yes: 9	2/9	3/9	3/9	3/9
No: 5	3/5	1/5	4/5	3/5
Total	5	4	7	6

Don’t worry for the 1/P(E); It’s alpha, the normalization constant.

The weather data example: no

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

P(no | E) =

P(Sunny | no) *

P(Cool | no) *

P(Humidity=High | no) *

P(Windy=True | no) *

P(no) / P(E) =

= (3/5) *

(1/5) *

(4/5) *

(3/5) *

(5/14) / P(E) = 0.0206 / P(E)

Play	Sunny	Cool	High humidity	Windy= true
Yes: 9	2/9	3/9	3/9	3/9
No: 5	3/5	1/5	4/5	3/5
Total	5	4	7	6

The weather data example: decision

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Evidence E*

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

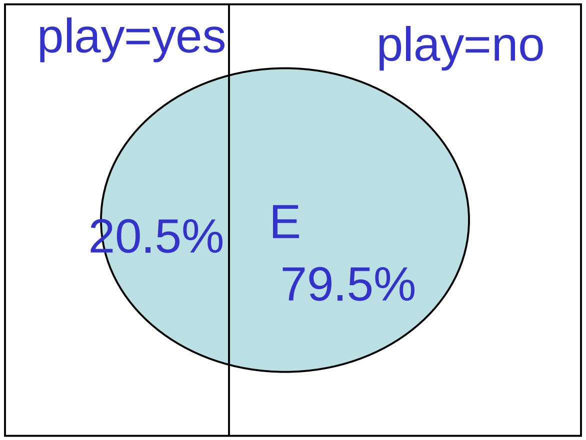
P(yes | E) = 0.0053 / P(E)

P(no | E) = 0.0206 / P(E)

More probable: no.

It would be nice to give the actual probability estimates

Normalization constant 1/P(E)



P(play=yes | E) + P(play=no | E) = 1 i.e.

0.0053 / P(E) + 0.0206 / P(E) = 1 i.e.

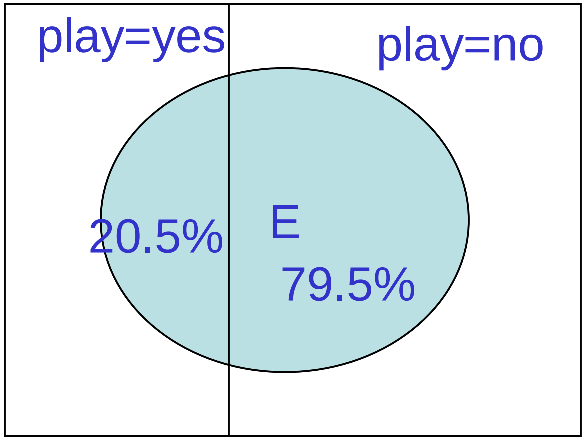
P(E) = 0.0053 + 0.0206

So,

P(play=yes | E) = 0.0053 / (0.0053 + 0.0206) = **20.5%**

P(play=no | E) = 0.0206 / (0.0053 + 0.0206) = **79.5%**

In other words:



P(play=yes | E) + P(play=no | E) = 1

P(play=yes |E) / P (play=no | E) = 0.0053 : 0.0206 = 0.26

0.26 * P (play=no | E) + P (play=no | E) = 1

P (play=no | E) = 1/1.26 = 79%

The remaining goes to yes: P(play=yes |E) = 21%

Naïve Bayes: issues

1. Zero frequency problem

2. Missing values

3. Numeric attributes

1. The "zero-frequency problem"

- What if an attribute value doesn't occur with every class value (e.g. "Humidity = High" for class "Play=Yes")?
  - Probability P(Humidity=High|play=yes) will be zero.

- P(Play="Yes"|E) will also be zero!
  - No matter how likely the other values are!

- Remedy – Laplace correction:
  - Add 1 to the count for every attribute value-class combination (Laplace estimator);
  - Add *k* (# of possible attribute values) to the denominator.

Laplace correction

Outlook	Play	Count
Sunny	No	0
Sunny	Yes	6
Overcast	No	2
Overcast	Yes	2
Rainy	No	3
Rainy	Yes	1

$\xrightarrow{+1}$

Outlook	Play	Count
Sunny	No	1
Sunny	Yes	7
Overcast	No	3
Overcast	Yes	3
Rainy	No	4
Rainy	Yes	2

It was: out of total 9 'Yes'

6 – Sunny, 2 – Overcast, 1 – Rainy

The probabilities were:

P(Sunny | yes)= 6/9; P(Overcast|yes) = 2/9; P(Rainy|yes)= 1/9

After correction:

7 – Sunny, 3 – Overcast, 2 – Rainy: Total 'Yes': 9+3=12

(hence add the cardinality of the attribute to the denominator)

Laplace correction

Outlook	**Play**	**Count**
Sunny	No	0
Sunny	Yes	6
Overcast	No	2
Overcast	Yes	2
Rainy	No	3
Rainy	Yes	1

$\xrightarrow{+1}$

Outlook	**Play**	**Count**
Sunny	No	1
Sunny	Yes	7
Overcast	No	3
Overcast	Yes	3
Rainy	No	4
Rainy	Yes	2

The probabilities were:

P(Sunny | yes)= 6/9; P(Overcast|yes) = 2/9; P(Rainy|yes)= 1/9

After correction the probabilities:

P(Sunny | yes)= 7/(9+3);

P(Overcast|yes) = 3/(9+3);

P(Rainy|yes)= 2/(9+3)

Needs to sum up to 1.0

Laplace correction example

P(yes | E) =

P(Outlook=Sunny | yes) *

P(Temp=Cool | yes) *

P(Humidity=High | yes) *

P(Windy=True | yes) *

P(yes) / P(E) =

= (2/9) * (3/9) * (3/9) * (3/9) *(9/14) / P(E) = 0.0053 / P(E)

With Laplace correction:

Number of possible values for 'Outlook'

= ((2+1)/(9+3)) * ((3+1)/(9+3)) * ((3+1)/(9+2)) * ((3+1)/(9+2)) *(9/14) / P(E)
= 0.007 / P(E)

Number of possible values for 'Windy'

2. Missing values: in the training set

- Missing values - not a problem for Naïve Bayes

- Suppose 1 value for outlook in the training set is missing. We count only existing values. For a large dataset, the probability P(outlook=sunny|yes) and P(outlook=sunny|no) will not change much. This is because we use probabilities rather than absolute counts.

2. Missing values: in the evidence set

- The same calculation without one fraction

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

P(yes | E) =

P(Temp=Cool | yes) *

P(Humidity=High | yes) *

P(Windy=True | yes) *

P(yes) / P(E) =

= (3/9) * (3/9) * (3/9) *(9/14) / P(E) =
0.0238 / P(E)

P(no | E) =

P(Temp=Cool | no) *

P(Humidity=High | no) *

P(Windy=True | no) *

P(play=no) / P(E) =

= (1/5) * (4/5) * (3/5) *(5/14) / P(E) =
0.0343 / P(E)

2. Missing values: in the evidence set

- With missing value:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

$P(yes \mid E) = 0.0238 / P(E)$ $P(no \mid E) = 0.0343 / P(E)$

- Without missing value:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

$P(yes \mid E) = 0.0053 / P(E)$ $P(no \mid E) = 0.0206 / P(E)$

The numbers are much higher for the case of missing values. But we care only about the ratio of yes and no.

2. Missing values: in the evidence set

- With missing value:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

P(yes | E) = 0.0238 / P(E)   P(no | E) = 0.0343 / P(E)

After normalization: P(yes | E) = **41%**,   P(no | E) = **59%**

- Without missing value:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

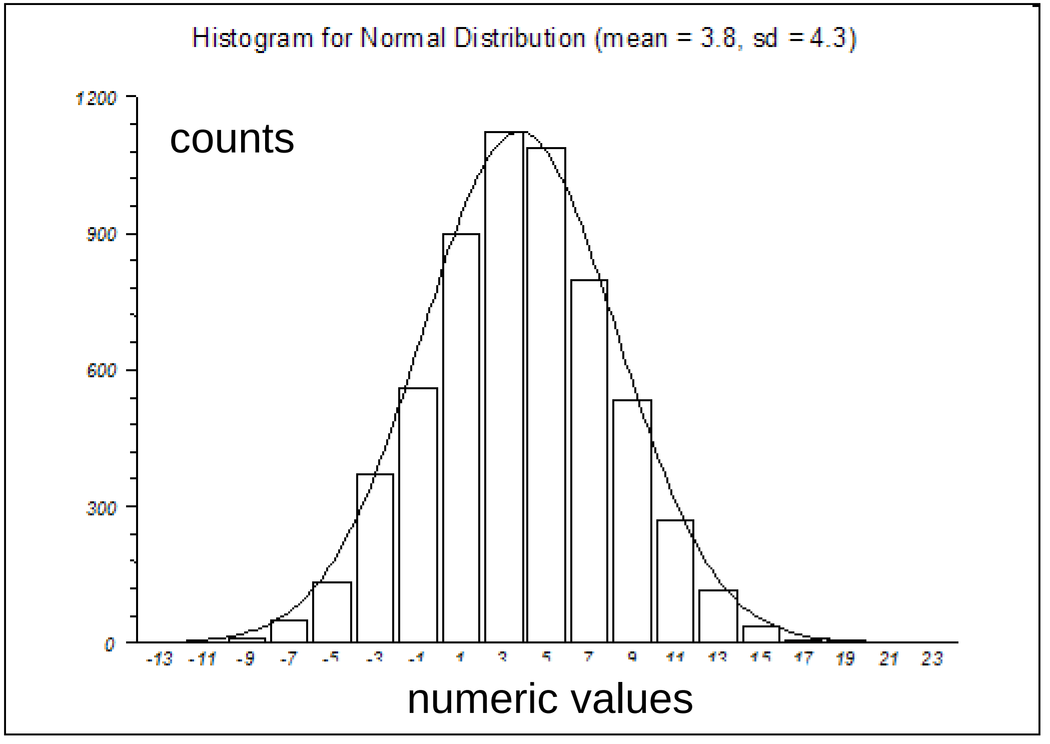
P(yes | E) = 0.0053 / P(E)   P(no | E) = 0.0206 / P(E)

After normalization: P(yes | E) = **21%**,   P(no | E) = **79%**

Of course, this is a very small dataset where each count matters, but the prediction is still the same: most probably – no play

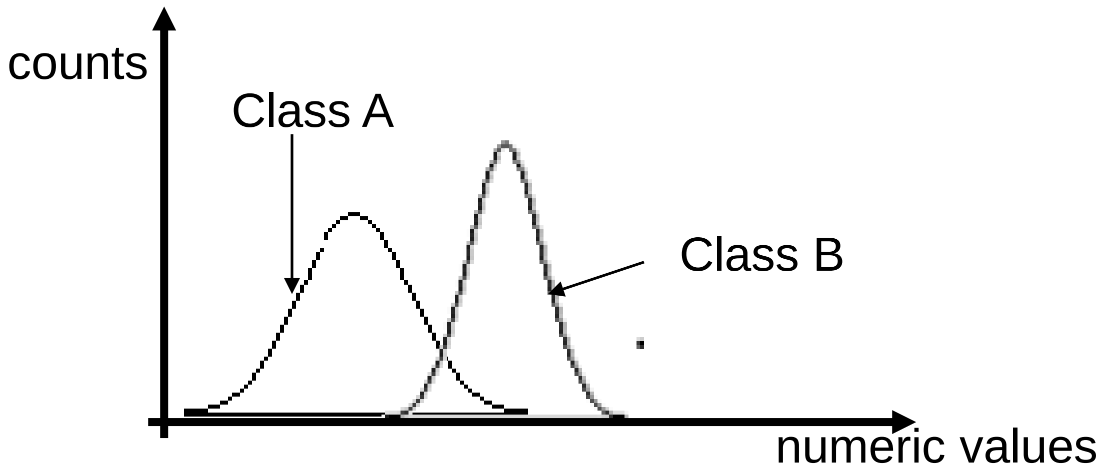
Normal distribution

- Usual assumption: attributes have a normal or Gaussian probability distribution.



Two classes have different distributions

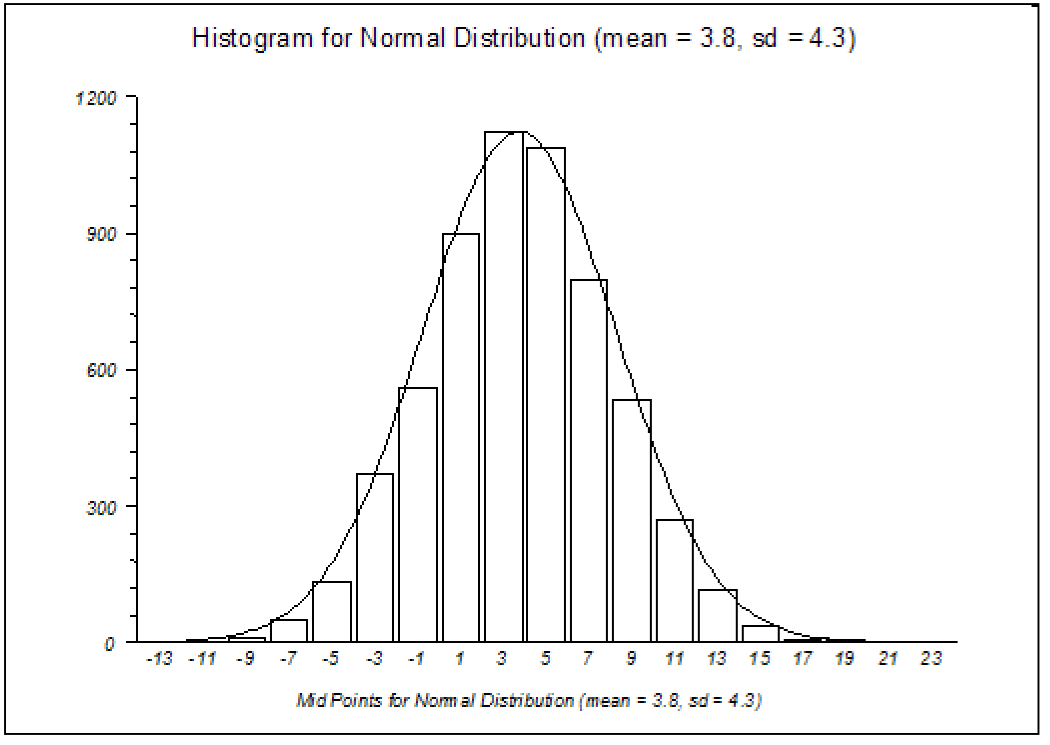
- Class A is normally distributed around its mean with its standard deviation. Class B is normally distributed around the different mean and with a different std



Probability density function

- Probability density function (PDF) for the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$



For a given x – evaluates its probability according to the distribution of probabilities in a given class

Probability and density

- Relationship between probability and density:

$$\Pr[c-\frac{\varepsilon}{2}<x<c+\frac{\varepsilon}{2}]\approx\varepsilon * f(c)$$

- But: to compare posteriori probabilities it is enough to calculate PDF, because ε cancels out
- Exact relationship:

$$\Pr[a\leq x\leq b]=\int_a^b f(t)dt$$

To compute probability P(X=V|class)

- Gives ≈ probability of X=V of belonging to class A:

$$f(x \mid class) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- We approximate μ by the sample mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- We approximate σ^2 by the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Numeric weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(x \mid yes) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Compute the probability of temp=66 for class Yes:

~μ (mean) = (83+70+68+64+69+75+75+72+81)/ 9 = 73

~σ^2 (variance) = ((83-73)^2 + (70-73)^2 + (68-73)^2 + (64-73)^2 + (69-73)^2 + (75-73)^2 + (75-73)^2 + (72-73)^2 + (81-73)^2)/ (9-1) = 38

$$f(x \mid yes) = \frac{1}{\sqrt{38*2*3.14}} 2.7^{-\frac{(x-73)^2}{2*38}}$$

Density function for temp in class Yes

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

Substitute x=66:

$$f(x=66 \mid yes) = \frac{1}{15.44} 2.7^{-\frac{(66-73)^2}{76}} = 0.034$$

P(temp=66|yes)=0.034

Numeric weather data example

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(x \mid yes) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Compute the probability of Humidity=90 for class Yes:

~μ (mean) = (86+96+80+65+70+80+70+90+75)/ 9 = 79

~σ^2 (variance) = ((86-79)^2 + (96-79)^2 + (80-79)^2 + (65-79)^2 + (70-79)^2 + (80-79)^2 + (70-79)^2 + (90-79)^2 + (75-79)^2)/ (9-1) = 104

outlook	temperature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

$$f(x \mid yes) = \frac{1}{\sqrt{104*2*3.14}} 2.7^{-\frac{(x-79)^2}{2*104}}$$

Density function for humidity in class Yes

Substitute x=90:

$$f(x=90 \mid yes) = \frac{1}{25.55} 2.7^{-\frac{(90-79)^2}{208}} = 0.022$$

P(humidity=90|yes)=0.022

Classifying a new day

- A new day E:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

P(play=yes | E) =
P(Outlook=Sunny | play=yes) *
P(Temp=66 | play=yes) *
P(Humidity=90 | play=yes) *
P(Windy=True | play=yes) *
P(play=yes) / P(E) =
= (2/9) * (0.034) * (0.022) * (3/9) *(9/14) / P(E) = 0.000036 / P(E)

P(play=no | E) =
P(Outlook=Sunny | play=no) *
P(Temp=66 | play=no) *
P(Humidity=90 | play=no) *
P(Windy=True | play=no) *
P(play=no) / P(E) =
= (3/5) * (0.0291) * (0.038) * (3/5) *(5/14) / P(E) = 0.000136 / P(E)

After normalization: P(play=yes | E) = **20.9%**, P(play=no | E) = **79.1%**

Practicality

- Naïve Bayes works surprisingly well (even if independence assumption is clearly violated)

- Because classification doesn't require accurate probability estimates as long as maximum probability is assigned to correct class

Applications of Naïve Bayes

The best classifier for:

- Document classification
- Diagnostics
- Clinical trials
- Assessing risks

Text Categorization

- Text categorization is the task of assigning a given document to one of a fixed set of categories, on the basis of the words it contains.

- The class is the document category, and the evidence variables are the presence or absence of each word in the document.

Text Categorization

- The model consists of the prior probability **P**(Category) and the conditional probabilities **P**(Word_i | Category).

- For each category *c*, P(Category=c) is estimated as the fraction of all the "training" documents that are of that category.

- Similarly, P(Word_i = true | Category = c) is estimated as the fraction of documents of category that contain this word.

- Also, P(Word_i = true | Category = $\neg$c) is estimated as the fraction of documents not of category that contain this word.

Text Categorization (cont'd)

- Now we can use naïve Bayes for classifying a new document with n words:

$P(Category = c \mid Word_1 = true, \ldots, Word_n = true) =$

$$\alpha * P(Category = c)\prod^{n}_{i=1} P(Word_i = true \mid Category = c)$$

$P(Category = \neg c \mid Word_1 = true, \ldots, Word_n = true) =$

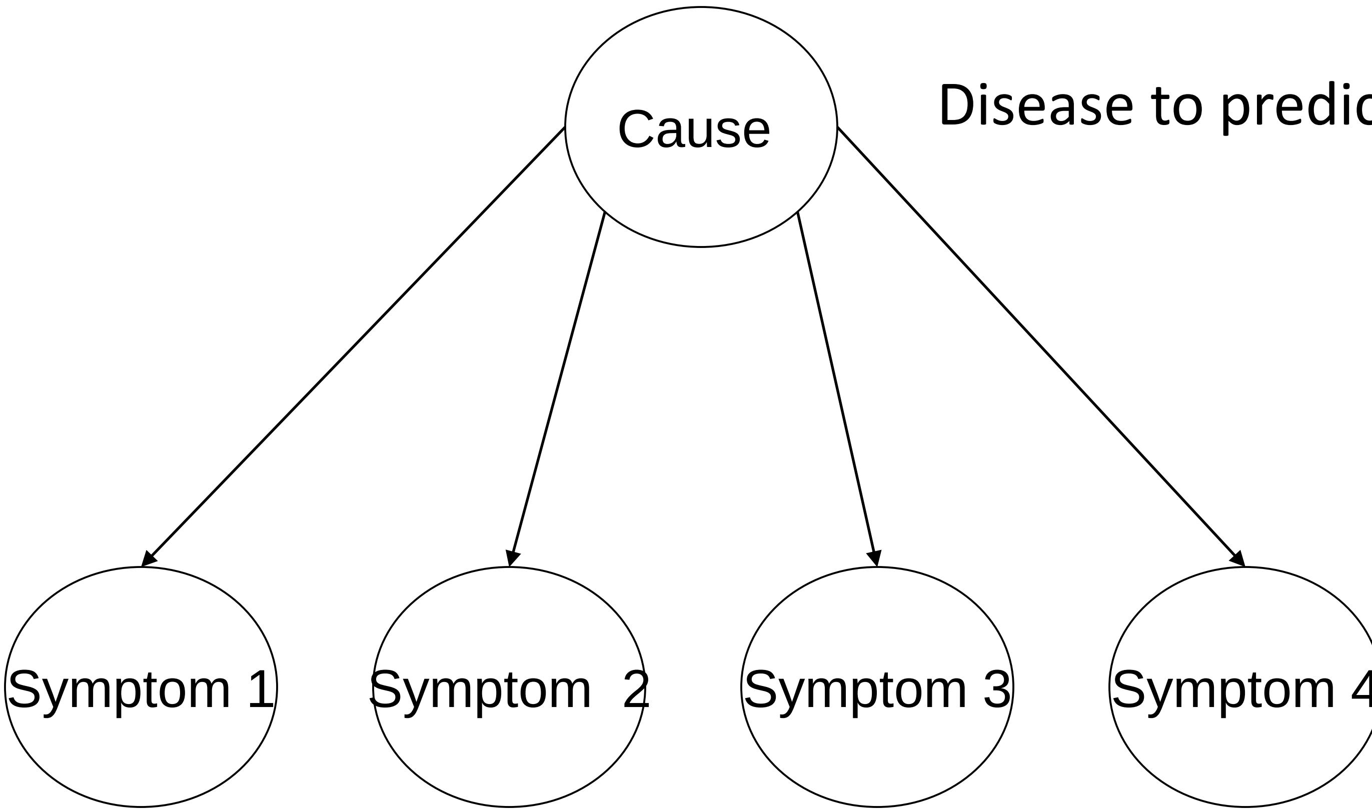
$$\alpha * P(Category = \neg c)\prod^{n}_{i=1} P(Word_i = true \mid Category = \neg c)$$

$Word_1, \ldots, Word_n$ are the words occurring in the new document

α is the normalization constant.

- Observe that similarly with the "missing values" the new document doesn't contain every word for which we computed the probabilities.

Diagnostics with Naïve Bayes



Disease to predict (hidden)

Set of effects (demonstrate themselves)

Example of diagnostic problem

- A doctor knows that 50% of patients with a stiff neck were diagnosed with meningitis.

- The **doctor also knows some unconditional** facts (prior probabilities):

  the prior probability that any patient has meningitis is 1/50,000

  the probability that he does not have a meningitis is 49,999/50,000

Diagnostic problem

P(StiffNeck=true | Meningitis=true) = 0.5

P(StiffNeck=true | Meningitis=false) = 0.5

P(Meningitis=true) = 1/50000

P(Meningitis=false) = 49999/50000

P(Meningitis=**true** | StiffNeck=true)

= P(StiffNeck=true | Meningitis=true) P(Meningitis=true) /
P(StiffNeck=true)

= (0.5) x (1/50000) / P(StiffNeck=true) =0.5 * 0.00002 / P(StiffNeck=true) =
0.00010 / P(StiffNeck=true)

P(Meningitis=**false** | StiffNeck=true)

= P(StiffNeck=true | Meningitis=false) P(Meningitis=false) /
P(StiffNeck=true)

= (0.5)*(49999/50000)/ P(StiffNeck=true) = 0.49999 / P(StiffNeck=true)

1/5000 chance that the patient with a stiff neck has meningitis (due to the very low prior probability)

Bayes' rule critics: prior probabilities

- The doctor has the above quantitative information in the diagnostic direction from symptoms (evidences, effects) to causes.
- The problem is that prior probabilities are hard to estimate and they may fluctuate. Imagine, there is sudden epidemic of meningitis. The prior probability, P(Meningitis=true), will go up.
- Clearly, P(StiffNeck=true|Meningitis=true) is unaffected by the epidemic. It simply reflects the way meningitis works.
- The estimation of P(Meningitis=true|StiffNeck=true) will be incorrect until new data about P(Meningitis=true) are collected

Tax Data – Naive Bayes

Classify: (_, No, Married, 95K, ?)

(Apply also the Laplace normalization)

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tax Data – Naive Bayes

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Classify: (_, No, Married, 95K, ?)

P(Yes) = 3/10 = 0.3

P(Refund=No|Yes) = (3+1)/(3+2) = 0.8

P(Status=Married|Yes) = (0+1)/(3+3) = 0.17

$$f(income \mid Yes) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Approximate μ with: (95+85+90)/3 =90

Approximate σ^2 with:

((95-90)^2+(85-90) ^2+(90-90) ^2)/ (3-1) = 25

f(income=95|Yes) =

e(- ((95-90)^2 / (2*25))) / sqrt(2*3.14*25) = .048

P(Yes | E) = α*.8*.17*.048*.3= α*.0019584

Tax Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Classify: (_, No, Married, 95K, ?)

P(No) = 7/10 = .7

P(Refund=No|No) = (4+1)/(7+2) = .556

P(Status=Married|No) = (4+1)/(7+3) = .5

$$f(income \mid No) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Approximate μ with:

(125+100+70+120+60+220+75)/7 =110

Approximate σ^2 with:

((125-110)^2 + (100-110)^2 + (70-110)^2 + (120-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2)/(7-1) = 2975

f(income=95|No) =

e(-((95-110)^2 / (2*2975))) /sqrt(2*3.14* 2975) = .00704

P(No | E) = α*.556*.5* .00704*0.7= α*.00137

Tax Data

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Classify: (_, No, Married, 95K, ?)

$P(Yes \mid E) = \alpha * .0019584$

$P(No \mid E) = \alpha * .00137$

$\alpha = 1/(.0019584 + .00137) = 300.44$

$P(Yes|E) = 300.44 * .0019584 = 0.59$

$P(No|E) = 300.44 * .00137 = 0.41$

We predict "Yes."