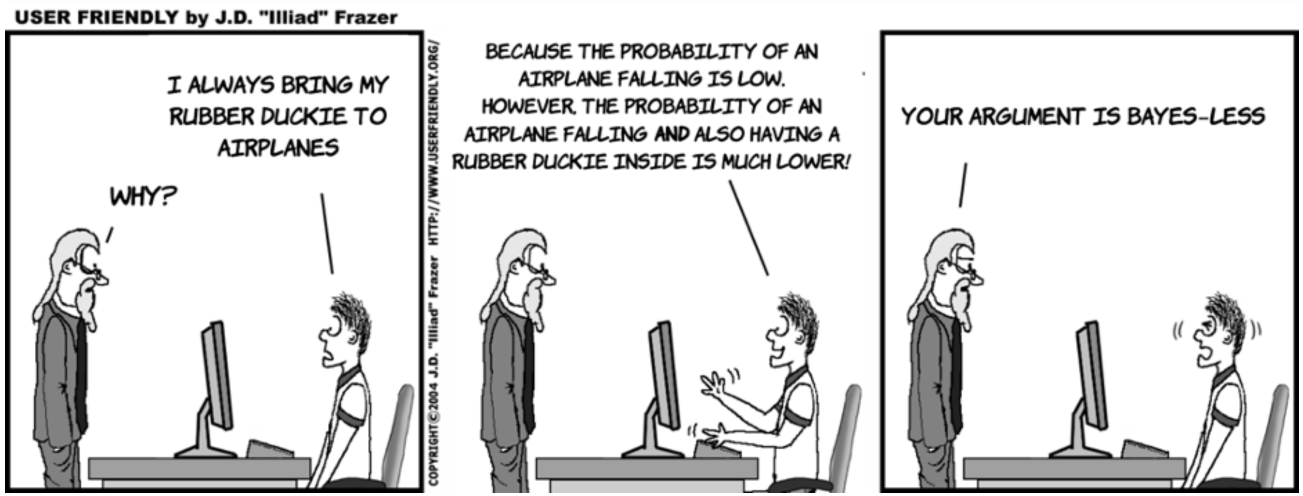


Naïve Bayes Classifier: refinements

Lecture 18

by Marina Barsky

Naïve Bayes classifier

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

$$\mathrm{P}(\mathrm{class} = \mathrm{A} | \mathrm{evidence1}, \mathrm{evidence2}, \dots, \mathrm{evidenceN})$$

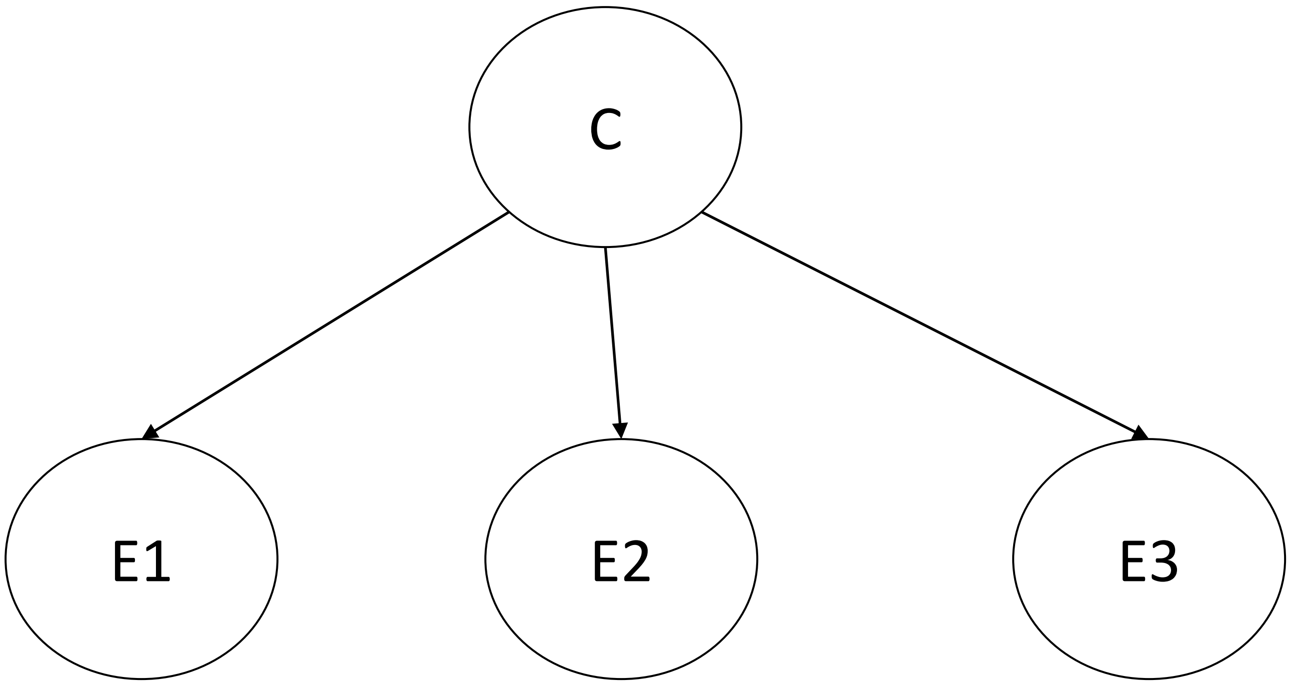
$$= \frac{\mathrm{P}(\mathrm{evidence1}|\mathrm{class{=}A}) * \cdots * \mathrm{P}(\mathrm{evidence}N|\mathrm{class{=}A}) * \mathrm{P}(\mathrm{class{=}A})}{\mathrm{P}(\mathrm{evidence1}) * \cdots * \mathrm{P}(\mathrm{evidence}N)}$$

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

- **Naïve – because it assumes *conditional* independence of variables**

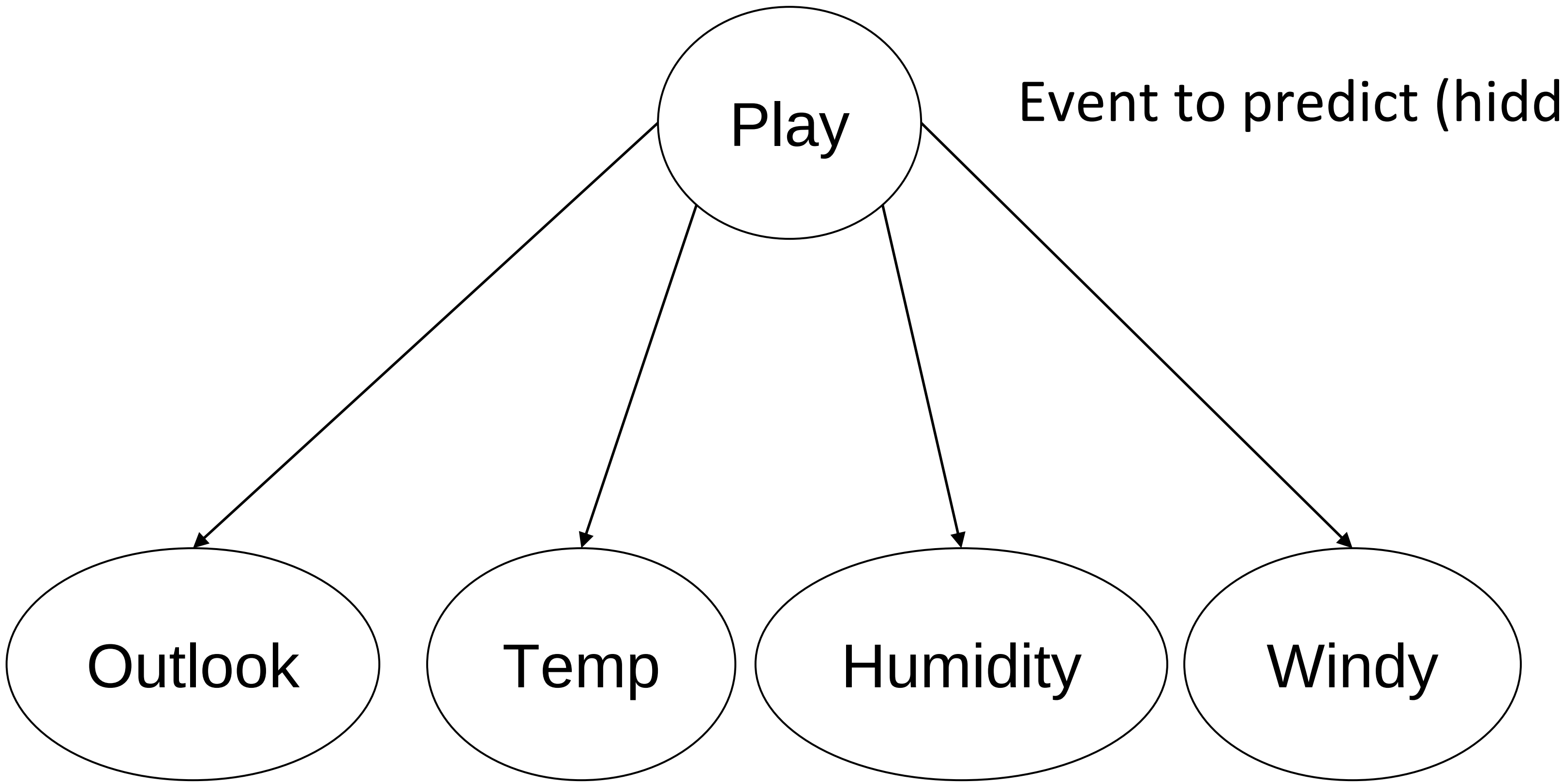
- Although based on assumptions that are almost never correct, this scheme works well in practice!

Naïve Bayes as a graph (network)



This graph states that there is a ***probabilistic dependence*** between C and each E_i. The probability of one of these variables (Class to predict) is influenced by the probabilities of the rest of the variables (set of evidences) and vice versa: $P(C|E) \neq P(C)$, and $P(E|C) \neq P(E)$

Multi-evidence classifier for Weather dataset



Event to predict (hidden)

Set of evidences (demonstrate themselves)

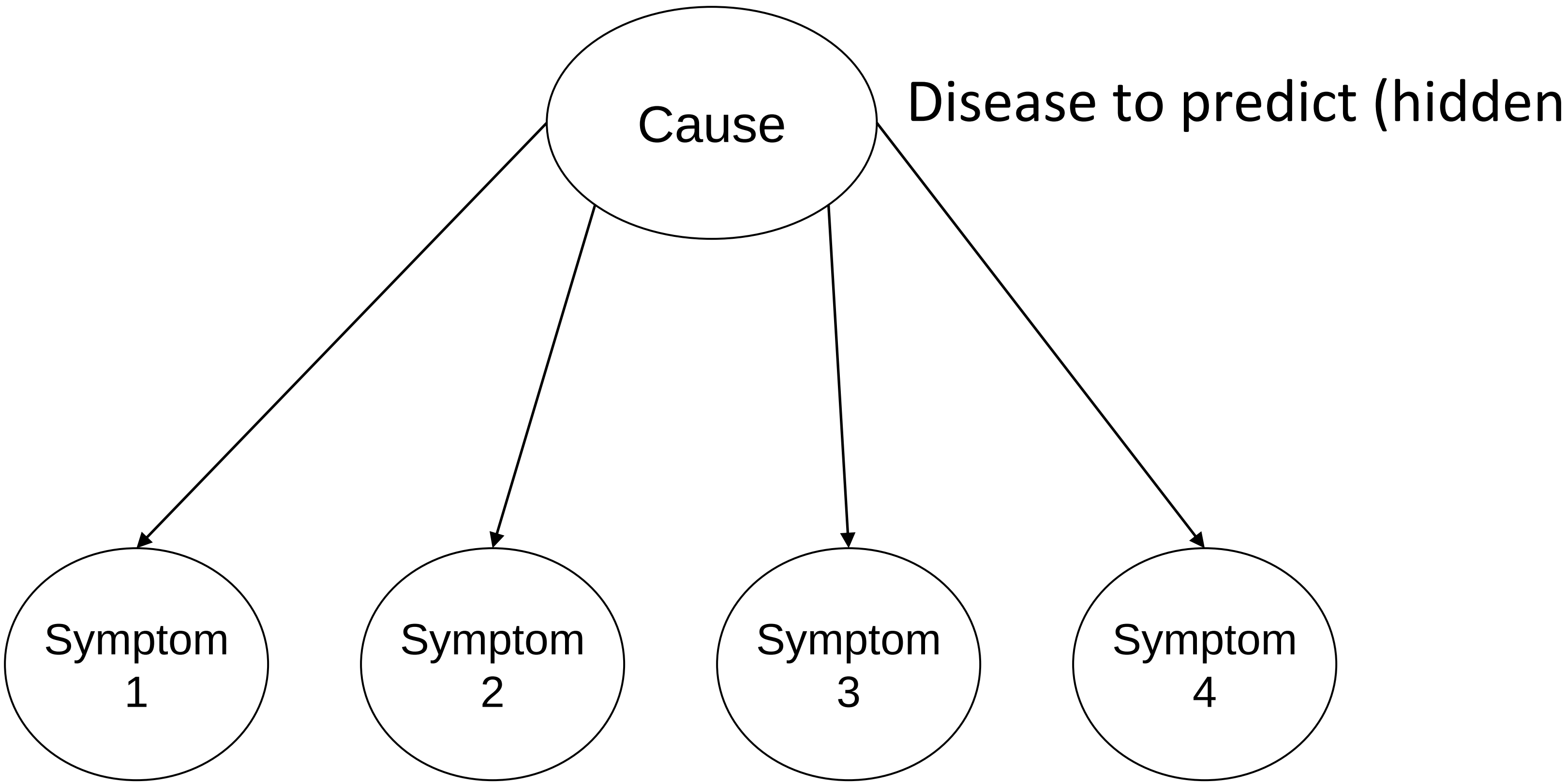
Naïve Bayes: issues

1. Prior probabilities may change
2. Zero frequency problem
3. Missing values
4. Numeric attributes

Issue 1

PRIOR PROBABILITIES

Diagnostics with Naïve Bayes



Set of effects (demonstrate themselves)

Example: diagnosing meningitis

- A doctor knows that 50% of patients with meningitis presented with a stiff neck syndrome.

- 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 a 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

Issue 2

ZERO FREQUENCY

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: example

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 5 'No'

0 – Sunny, 2 – Overcast, 3 – Rainy

The probabilities were:

P(Sunny | no)= 0/5; P(Overcast|no) = 2/5; P(Rainy|no)= 3/5

After correction:

1 – Sunny, 3 – Overcast, 4 – Rainy: Total 'No': 5+3=8

(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

After correction the probabilities:

$$\left.\begin{aligned} &P(\text{Sunny} \mid \text{no}) = 1/(5+3);\\ &P(\text{Overcast} \mid \text{no}) = 3/(5+3);\\ &P(\text{Rainy} \mid \text{no}) = 4/(5+3) \end{aligned}\right\} \text{Needs to sum up to 1.0}$$

You add this correction to all counts, **for both classes**

Laplace correction example

$P(\text{yes} \mid E) =$

$P(\text{Outlook=Sunny} \mid \text{yes}) *$

$P(\text{Temp=Cool} \mid \text{yes}) *$

$P(\text{Humidity=High} \mid \text{yes}) *$

$P(\text{Windy=True} \mid \text{yes}) *$

$P(\text{yes}) / P(E) =$

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

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’

Issue 3

MISSING VALUES

Missing values: in the **training** set

- Missing values - not a problem for Naïve Bayes

- Suppose that one 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 odds ratio rather than absolute counts.

Missing values: in the **query**

- 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)

Missing values: in the **query**

- 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)

- 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)

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

Missing values: in the **query**

- 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%**

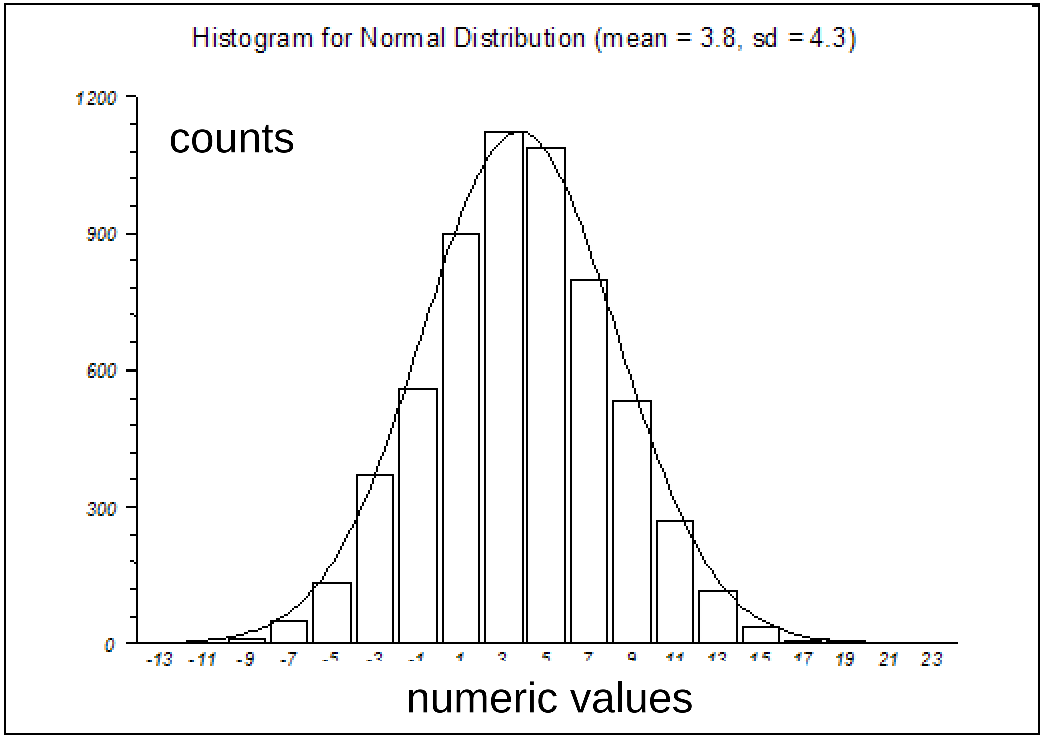
Of course, this is a very small dataset where each count matters, but the prediction is still the same: most probably – no play

Issue 4

NUMERIC ATTRIBUTES

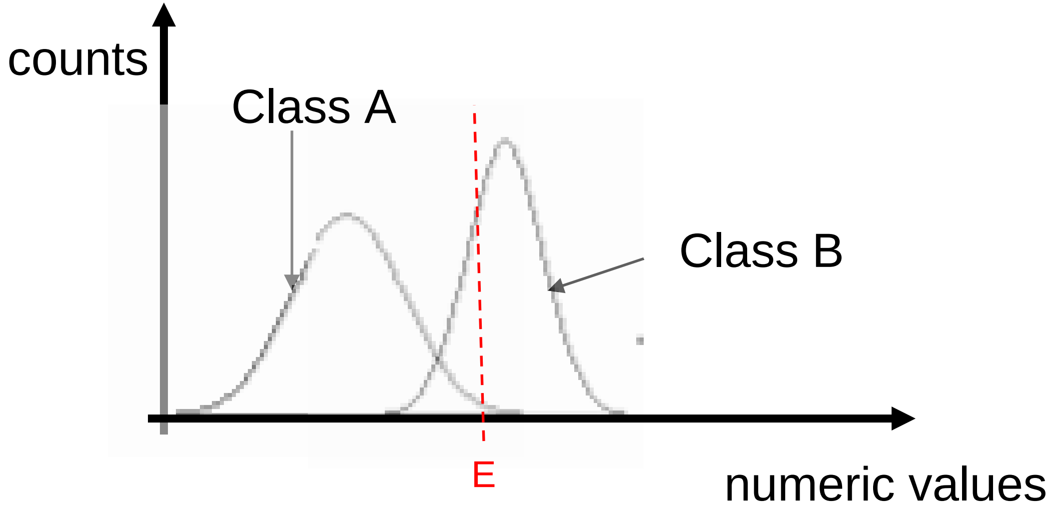
Normal distribution

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



Two classes have different distributions

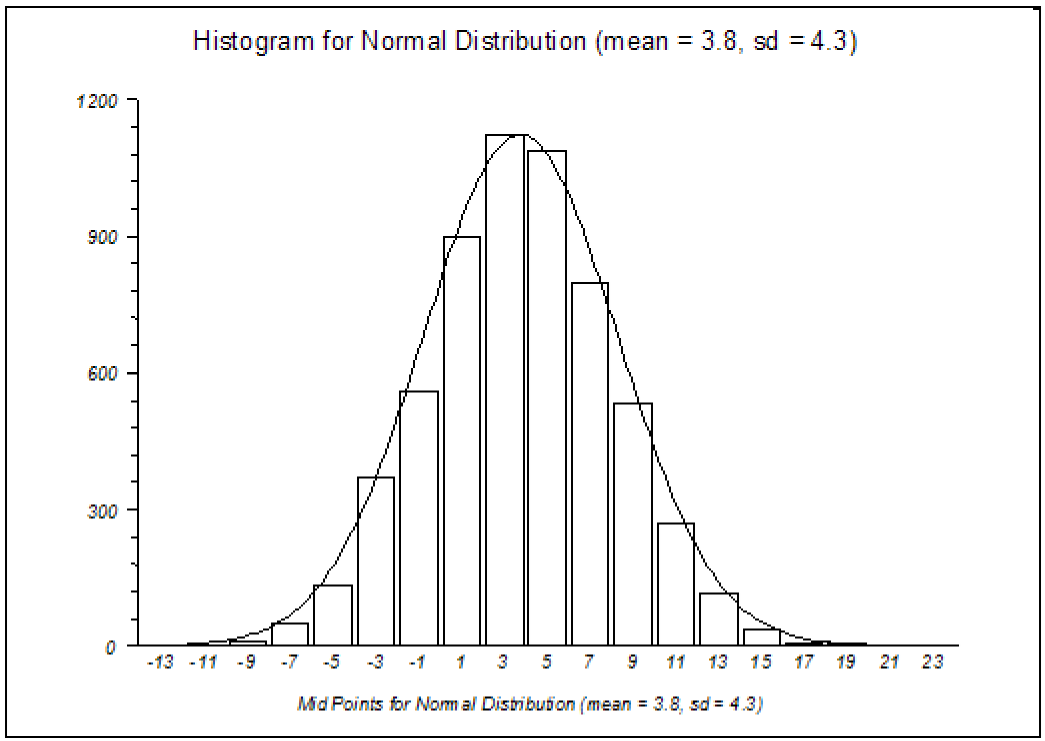
- Class A is normally distributed around its mean with its standard deviation.
- Class B is normally distributed around the different mean and with a different std



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 the probability of [x-$\boldsymbol{\varepsilon}$,x+$\boldsymbol{\varepsilon}$] 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:

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

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:

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

- We approximate σ^2 by the sample variance:

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

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

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

Density function for humidity in class Yes

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=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%**

Exercise: 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

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

Exercise: 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.”

Summary

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

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

Applications of Naïve Bayes

The best classifier for:

- Document classification (filtering)
- 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 *c* that contain this word.

- Also, P(Word_i = true | Category = ¬c) is estimated as the fraction of documents not of category *c* 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.