REVIEW OF PROBABILITY CONCEPTS

SYRACUSE UNIVERSITY
School of Information Studies

BAYES' THEOREM

What is Bayes' theorem mathematically?

How can computers use Bayes' theorem?

REVIEW OF PROBABILITY CONCEPTS

An **event** is an outcome of an experiment.

Experiment: Toss a fair coin

Possible outcomes: “head” or “tail”

The **probability** that either event (“head” or “tail”) occurs is 50%, i.e.:

$P(\text{head}) = 0.5$, $P(\text{tail}) = 0.5$

$P(\text{head}) + P(\text{tail}) = 1$

If the coin is not balanced, then $P(\text{head}) \neq P(\text{tail})$.

JOINT PROBABILITY

Joint probability P(A, B) is the chance that two events A and B co-occurred.

Experiment: Toss two fair coins

Event A: coin1 = “head”

Event B: coin2 = “head”

P(A, B) = P(coin1 = “head,” coin2 = “head”)

P(A, B) and P(B, A) are the same thing.

SYRACUSE UNIVERSITY
School of Information Studies

INDEPENDENT EVENTS

The occurrence of one event does not depend on the occurrence of the other event.

E.g., tossed two fair coins; both landed head up

Event A: coin1 = “head”

Event B: coin2 = “head”

P(A, B) = P(A) * P(B) = 0.5 * 0.5 = 0.25

DEPENDENT EVENTS

The occurrence of one event is dependent on the occurrence of another event.

E.g., choose a card from a standard deck of 52 cards. Without replacement, choose another card. What is the probability of choosing two queens?

Because the outcome of the second card is dependent on the outcome of the first card, P(card1 = “queen,” card2 = “queen”) **!=** P(card1 = “queen”) * P(card2 = “queen”)

CONDITIONAL PROBABILITY

Conditional probability P(B|A)

The probability that event B occurs if event A occurs

Event A: card1 = “queen”

Event B: card2 = “queen”

P(B|A) = P(card2 =“queen” | card1 = “queen”) = 3/51, because there are 51 remaining cards and only three queens among them.

RELATIONSHIP BETWEEN CONDITIONAL AND JOINT PROBABILITIES

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

P(card1 = “queen,” card2 =“queen”)
= P(card1 = “queen”) * P(card2 = “queen” | card1= “queen”)
= 4/52 * 3/51
= 0.5%

PROBABILITY IS USEFUL IN DAILY LIFE

Which one is greater?

P("will" | "I")

P("has" | "I")

This is how your smartphone knows what words to recommend when you are texting!

All words that people typed become the training corpus, in which the conditional probabilities are calculated.

BAYES RULES IN MAMMOGRAM EXAMPLE

SYRACUSE UNIVERSITY
School of Information Studies

BAYES' THEOREM

Bayes' theorem lets us swap the order of the dependence between events.

Two events A and B

We know that P(A,B) = P(B|A)P(A)

Since P(A,B) = P(B,A), we also know that P(B,A) = P(B|A)P(A)

Therefore:

$$P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

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

BACK TO THE MAMMOGRAM EXAMPLE IN PROFESSOR STROGATZ'S ARTICLE

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

Event A: A patient has cancer

Event B: A patient's mammogram test result is positive

P(B|A) = Among patients who have cancer, how many have positive test results?

We can calculate it from past data.

P(A|B) = For a patient with a positive test result, what is the chance of cancer?

This is **prediction**!

Bayes' theorem provides an approach to predict the probability of future events based on prior experience.

THE MAMMOGRAM EXAMPLE

The probability that a woman has breast cancer is … ?

If a woman has breast cancer, the probability that she will have a positive mammogram is … ?

If a woman does not have breast cancer, the probability that she will still have a positive mammogram is … ?

BACK TO THE MAMMOGRAM EXAMPLE

The probability that a woman has breast cancer is
P(cancer) = 0.008

If a woman has breast cancer, the probability that she will have a positive mammogram is
P(positive|cancer) = 0.9

If a woman does not have breast cancer, the probability that she will still have a positive mammogram is
P(positive|no cancer) = 0.07

TRANSLATE THEM INTO PROBABILITY NOTATIONS

Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?

Which probability is greater:
P(cancer | positive) or P(no cancer | positive)?

To calculate P(cancer | positive), need to calculate:

- P(positive)
- P(cancer)
- P(positive | cancer)

To calculate P(no cancer | positive), need to calculate:

- P(positive)
- P(no cancer)
- P(positive | no cancer)

TRANSLATE THEM INTO PROBABILITY NOTATIONS

The probability that a woman has positive mammogram is:

P(positive)=Unknown

The probability that a woman has breast cancer is 0.008.

P(cancer) = 0.008, P(no cancer) = 0.992, (add up to 1)

If a woman has breast cancer, the probability that she will have a positive mammogram is 0.9.

P(positive | cancer) = 0.9

If a woman does not have breast cancer, the probability that she will still have a positive mammogram is 0.07.

P(positive | no cancer) = 0.07

TRANSLATE THEM INTO PROBABILITY NOTATIONS

All **prior probabilities** we have calculated:

P(cancer) = 0.008

P(no cancer) = 0.992

All **conditional probabilities** we have calculated:

P(positive | cancer) = 0.9

P(positive | no cancer) = 0.07

The **posterior probabilities** to be calculated:

P(cancer | positive) = ? P(no cancer | positive) = ?

SO, OUR PREDICTION IS …

$$P(cancer \mid positive)$$

$$= \frac{P(positive \mid cancer) \cdot P(cancer)}{P(positive)} = \frac{0.9 \cdot 0.008}{P(positive)} = \frac{0.0072}{P(positive)}$$

$$P(no\_cancer \mid positive)$$

$$= \frac{P(positive \mid no\_cancer) \cdot P(no\_cancer)}{P(positive)} = \frac{0.07 \cdot 0.992}{P(positive)} = \frac{0.069}{P(positive)}$$

NO CANCER!

$$P(cancer \mid positive)$$

$$= \frac{P(positive \mid cancer) \cdot P(cancer)}{P(positive)} = \frac{0.9 \cdot 0.008}{P(positive)} = \frac{0.0072}{P(positive)}$$

$$P(no\_cancer \mid positive)$$

$$= \frac{P(positive \mid no\_cancer) \cdot P(no\_cancer)}{P(positive)} = \frac{0.07 \cdot 0.992}{P(positive)} = \frac{0.069}{P(positive)}$$

Although we don't know P(positive), it does not matter. We just need to know which posterior probability is greater.

THIS DIAGNOSIS IS DETERMINED BY THE MAMMOGRAM RESULT ONLY

$$P(cancer \mid positive)$$

$$= \frac{P(positive \mid cancer) \cdot P(cancer)}{P(positive)} = \frac{0.9 \cdot 0.008}{P(positive)} = \frac{0.0072}{P(positive)}$$

$$P(no\_cancer \mid positive)$$

$$= \frac{P(positive \mid no\_cancer) \cdot P(no\_cancer)}{P(positive)} = \frac{0.07 \cdot 0.992}{P(positive)} = \frac{0.069}{P(positive)}$$

WHAT IF THE DIAGNOSIS IS DETERMINED BY MORE FACTORS THAN JUST THE MAMMOGRAM RESULT?

Attribute 1: Positive mammogram? Yes or no
Attribute 2: Family history? Yes or no
Attribute 3: Alcohol? Yes or no

How many posteriors to calculate?
Two posteriors for each possible combination of the attributes.
In this case, $2 * 2^3 = 16$

ALL POSTERIOR PROBABILITIES FOR THREE BINARY ATTRIBUTES

$$P(positive, yes, yes \mid cancer)$$
$$P(positive, yes, no \mid cancer)$$
$$P(positive, no, yes \mid cancer)$$
$$P(positive, no, no \mid cancer)$$
$$P(negative, yes, yes \mid cancer)$$
$$P(negative, yes, no \mid cancer)$$
$$P(negative, no, yes \mid cancer)$$
$$P(negative, no, no \mid cancer)$$

$$P(positive, yes, yes \mid cancer)$$
$$P(positive, yes, no \mid no\_cancer)$$
$$P(positive, no, yes \mid no\_cancer)$$
$$P(positive, no, no \mid no\_cancer)$$
$$P(negative, yes, yes \mid no\_cancer)$$
$$P(negative, yes, no \mid no\_cancer)$$
$$P(negative, no, yes \mid no\_cancer)$$
$$P(negative, no, no \mid no\_cancer)$$

NAIVE BAYES

SYRACUSE UNIVERSITY
School of Information Studies

CHALLENGE OF BAYESIAN CLASSIFIER

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid cancer)$
$P(positive, no, yes \mid cancer)$
$P(positive, no, no \mid cancer)$
$P(negative, yes, yes \mid cancer)$
$P(negative, yes, no \mid cancer)$
$P(negative, no, yes \mid cancer)$
$P(negative, no, no \mid cancer)$

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid no\_cancer)$
$P(positive, no, yes \mid no\_cancer)$
$P(positive, no, no \mid no\_cancer)$
$P(negative, yes, yes \mid no\_cancer)$
$P(negative, yes, no \mid no\_cancer)$
$P(negative, no, yes \mid no\_cancer)$
$P(negative, no, no \mid no\_cancer)$

CHALLENGE OF BAYESIAN CLASSIFIER

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid cancer)$
$P(positive, no, yes \mid cancer)$
$P(positive, no, no \mid cancer)$
$P(negative, yes, yes \mid cancer)$
$P(negative, yes, no \mid cancer)$
$P(negative, no, yes \mid cancer)$
$P(negative, no, no \mid cancer)$

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid no\_cancer)$
$P(positive, no, yes \mid no\_cancer)$
$P(positive, no, no \mid no\_cancer)$
$P(negative, yes, yes \mid no\_cancer)$
$P(negative, yes, no \mid no\_cancer)$
$P(negative, no, yes \mid no\_cancer)$
$P(negative, no, no \mid no\_cancer)$

2^{n+1} possible combinations of values for n binary attributes

CHALLENGE OF BAYESIAN CLASSIFIER

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid cancer)$
$P(positive, no, yes \mid cancer)$
$P(positive, no, no \mid cancer)$
$P(negative, yes, yes \mid cancer)$
$P(negative, yes, no \mid cancer)$
$P(negative, no, yes \mid cancer)$
$P(negative, no, no \mid cancer)$

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid no\_cancer)$
$P(positive, no, yes \mid no\_cancer)$
$P(positive, no, no \mid no\_cancer)$
$P(negative, yes, yes \mid no\_cancer)$
$P(negative, yes, no \mid no\_cancer)$
$P(negative, no, yes \mid no\_cancer)$
$P(negative, no, no \mid no\_cancer)$

2^n possible combinations of values for n binary attributes
What if n = 10,000? Quite common for text classification when each word is an attribute.

CHALLENGE OF BAYESIAN CLASSIFIER

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid cancer)$
$P(positive, no, yes \mid cancer)$
$P(positive, no, no \mid cancer)$
$P(negative, yes, yes \mid cancer)$
$P(negative, yes, no \mid cancer)$
$P(negative, no, yes \mid cancer)$
$P(negative, no, no \mid cancer)$

$P(positive, yes, yes \mid cancer)$
$P(positive, yes, no \mid no\_cancer)$
$P(positive, no, yes \mid no\_cancer)$
$P(positive, no, no \mid no\_cancer)$
$P(negative, yes, yes \mid no\_cancer)$
$P(negative, yes, no \mid no\_cancer)$
$P(negative, no, yes \mid no\_cancer)$
$P(negative, no, no \mid no\_cancer)$

Too many probabilities to calculate!

CHALLENGE OF BAYESIAN CLASSIFIER

$$P(positive, yes, yes \mid cancer) \quad P(positive, yes, yes \mid no\_cancer)$$
$$P(positive, yes, no \mid cancer) \quad P(positive, yes, no \mid no\_cancer)$$
$$P(positive, no, yes \mid cancer) \quad P(positive, no, yes \mid no\_cancer)$$
$$P(positive, no, no \mid cancer) \quad P(positive, no, no \mid no\_cancer)$$
$$P(negative, yes, yes \mid cancer) \quad P(negative, yes, yes \mid no\_cancer)$$
$$P(negative, yes, no \mid cancer) \quad P(negative, yes, no \mid no\_cancer)$$
$$P(negative, no, yes \mid cancer) \quad P(negative, no, yes \mid no\_cancer)$$
$$P(negative, no, no \mid cancer) \quad P(negative, no, no \mid no\_cancer)$$

There may not be enough data for accurately estimating each probability.

NAIVE BAYES CLASSIFIER

Assume independence among attributes A_i when class is given:

$$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j) * P(A_2 \mid C_j) * \ldots * P(A_n \mid C_j)$$
$$= \Pi\, P(A_i \mid C_j)$$

This assumption means that given the target class C, the probability of observing the conjunction $A_1, A_2, A_3, \ldots A_n$ is just the product of the probabilities for the individual attributes.

For example, for people with cancer, the assumption is that whether there is family history of cancer has nothing to do with race, which may or may not be scientifically true.

WHY IS THE INDEPENDENCE ASSUMPTION NEEDED?

Assume independence among attributes A_i when class is given:

$$P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$$
$$= \Pi\, P(A_i \mid C_j)$$

To greatly reduce the number of probabilities to estimate

WHY IS THE INDEPENDENCE ASSUMPTION NEEDED?

$$P(positive, yes_{family}, yes_{race} \mid cancer)$$

$$= P(positive \mid cancer) \cdot P(yes_{family} \mid cancer) \cdot P(yes_{race} \mid cancer)$$

$$P(positive, yes, yes \mid cancer)$$
$$P(positive, yes, no \mid cancer)$$
$$P(positive, no, yes \mid cancer)$$
$$P(positive, no, no \mid cancer)$$
$$P(negative, yes, yes \mid cancer)$$
$$P(negative, yes, no \mid cancer)$$
$$P(negative, no, yes \mid cancer)$$
$$P(negative, no, no \mid cancer)$$

Reduce calculation

2^3 -> 2*3

$$P(positive \mid cancer)$$
$$P(yes_{family} \mid cancer)$$
$$P(yes_{race} \mid cancer)$$
$$P(negative \mid cancer)$$
$$P(no_{family} \mid cancer)$$
$$P(no_{race} \mid cancer)$$

SIMILARLY, …

$$P(positive, yes, yes \mid no\_cancer)$$
$$P(positive, yes, no \mid no\_cancer)$$
$$P(positive, no, yes \mid no\_cancer)$$
$$P(positive, no, no \mid no\_cancer)$$
$$P(negative, yes, yes \mid no\_cancer)$$
$$P(negative, yes, no \mid no\_cancer)$$
$$P(negative, no, yes \mid no\_cancer)$$
$$P(negative, no, no \mid no\_cancer)$$

Reduce calculation

2^3 -> 2*3

$$P(positive \mid no\_cancer)$$
$$P(yes_{family} \mid no\_cancer)$$
$$P(yes_{race} \mid no\_cancer)$$
$$P(negative \mid no\_cancer)$$
$$P(no_{family} \mid no\_cancer)$$
$$P(no_{race} \mid no\_cancer)$$

SIMILARLY, …

$$P(positive, yes, yes \mid no\_cancer)$$
$$P(positive, yes, no \mid no\_cancer)$$
$$P(positive, no, yes \mid no\_cancer)$$
$$P(positive, no, no \mid no\_cancer)$$
$$P(negative, yes, yes \mid no\_cancer)$$
$$P(negative, yes, no \mid no\_cancer)$$
$$P(negative, no, yes \mid no\_cancer)$$
$$P(negative, no, no \mid no\_cancer)$$

Reduce calculation

2^3 -> 2*3

$$P(positive \mid no\_cancer)$$
$$P(yes_{family} \mid no\_cancer)$$
$$P(yes_{race} \mid no\_cancer)$$
$$P(negative \mid no\_cancer)$$
$$P(no_{family} \mid no\_cancer)$$
$$P(no_{race} \mid no\_cancer)$$

Total calculation reduction $2*2^n \rightarrow 2*2*n$

If n = 10, then 2,048 probabilities reduced to 40!

WHY DOES "NAIVE" BAYES WORK?

This assumption is often violated in real-world problems.

E.g., family history of cancer may be correlated with race.

However, naive Bayes algorithm works quite well in many problems, even when the assumption is violated. There are some mathematical explanation for this phenomenon.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29,* 103–30.

What if the correlations between attributes have to be addressed?

Use Bayesian belief network (BBN)

SMOOTHING

SYRACUSE UNIVERSITY
School of Information Studies

PROBLEM OF ZERO PROBABILITIES

$$
\begin{aligned}
&P(cancer = no \mid pos\_mammo = yes, fam\_hist = yes, caucasian = yes) \\
&= \frac{P(pos\_mammo = yes \mid cancer = no) \cdot P(fam\_hist = yes \mid cancer = no) \cdot P(caucasian = yes \mid cancer = no) \cdot P(cancer = no)}{P(pos\_mammo = yes, fam\_hist = yes, caucasian = yes)} \\
&= \frac{(2/4) \cdot (0/4) \cdot (2/4) \cdot (4/7)}{P(pos\_mammo = yes, fam\_hist = yes, caucasian = yes)} \\
&= 0
\end{aligned}
$$

Zero probability of no cancer?

PROBLEM OF ZERO PROBABILITIES

If one of the conditional probabilities is zero, then the entire product becomes zero.

But the zero probability may just be caused by lack of data.

P(fam_hist=yes | cancer=yes) = 1
P(fam_hist=no | cancer=yes) = 0
P(fam_hist=yes | cancer=no) = 0
P(fam_hist=no | cancer=no) = 1

SOLUTION TO ZERO PROBABILITY

Probability estimation should replace these zero probabilities with a very small probability, which means such events still occur in real world but are so rare that the training data did not include any of them.

Since all probabilities should add up to one, the other nonzero probabilities need to "shrink" a little bit, in order to "lend" the small amount to the zero probabilities.

This technique is called "smoothing."

SMOOTHING FOR ZERO PROBABILITIES

Using a smoothing algorithm, such as Laplacian smoothing, also called "add-one" smoothing

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

A binary classifier has two classes (c = 2).

ADD-ONE SMOOTHING

Example	Pos_mammo	Fam_hist	Alcohol	Cancer
1	Yes	Yes	Yes	Yes
2	Yes	Yes	No	Yes
3	No	Yes	Yes	Yes
4	Yes	No	No	No
5	Yes	No	Yes	No
6	No	No	Yes	No
7	No	No	No	No

	Cancer = Yes	Cancer = No
Fam_hist = Yes	3	0
Fam_hist = No	0	4

	Cancer = Yes	Cancer = No
Fam_hist = Yes	3 + 1	0 + 1
Fam_hist = No	0 + 1	4 + 1

Add-one smoothing means to add an example to each category.

ADD-ONE SMOOTHING

	Cancer = Yes	Cancer = No
Fam_hist = Yes	3	0
Fam_hist = No	0	4

→

	Cancer = Yes	Cancer = No
Fam_hist = Yes	3 + 1	0 + 1
Fam_hist = No	0 + 1	4 + 1

Add-one smoothing means to add an example to each category.

Original probabilities:

P(fam_hist=yes | cancer=yes) = 3/3
P(fam_hist=no | cancer=yes) = 0/3 = 0
P(fam_hist=yes | cancer=no) = 0/4 = 0
P(fam_hist=no | cancer=no) = 4/4 = 1

→

Smoothed probabilities:

P(fam_hist=yes | cancer=yes) = 4/5
P(fam_hist=no | cancer=yes) = 1/5
P(fam_hist=yes | cancer=no) = 1/6
P(fam_hist=no | cancer=no) = 5/6

The nonzero probabilities become smaller, "lending" values to zero probabilities.

SMOOTHING FOR ZERO PROBABILITIES

Using a smoothing algorithm, such as Laplacian smoothing, also called "add-one" smoothing

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

A binary classifier has two classes (c = 2).

$$\frac{0}{3} = 0$$

$$\frac{0+1}{3+2} = \frac{1}{5} = 0.2$$

Well, this is not a very small probability!

SYRACUSE UNIVERSITY
School of Information Studies

SMOOTHING FOR ZERO PROBABILITIES

Using a smoothing algorithm, such as Laplacian smoothing, also called "add-one" smoothing

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

The probabilities in real-world problems are usually very small; thus, add-one smoothing would change the probability just a little bit.

$$\frac{1}{7000} = 0.000143$$

$$\frac{1+1}{7000+2} = \frac{2}{7002} = 0.000285$$

LOG PROBABILITIES

The probabilities are so small that we usually use log(P) instead, so that we don't have to store that many preceding zeros, as in 0.000222.

$$\log\left(\frac{1}{7000}\right) = \quad 8.854$$

$$\log\left(\frac{1+1}{7000+2}\right) = \log\left(\frac{2}{7002}\right) = \quad 8.161$$

PROBABILITY OF CONTINUOUS VARIABLE

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO ESTIMATE PROBABILITIES OF CONTINUOUS ATTRIBUTES

Two approaches for continuous attributes:

Discretization

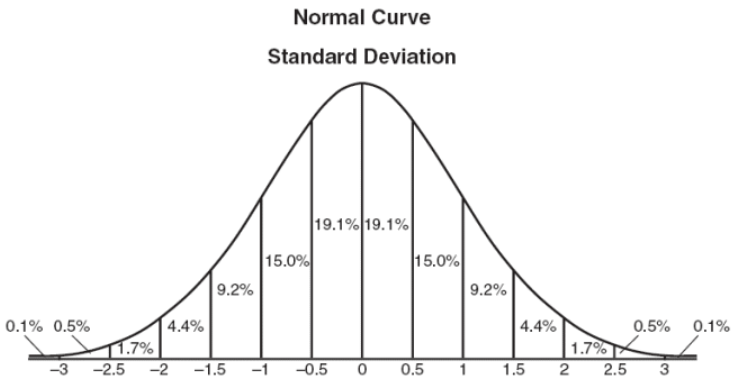
(0, 60k), (60k, 100k), (100k, ...)

Probability density estimation

Assume attribute follows a normal distribution.

Use data to estimate parameters of distribution (e.g., mean and standard deviation).

Once probability distribution is known, can use the probability density function to estimate the conditional probability $P(A_i|c)$.



HOW TO ESTIMATE PROBABILITIES OF CONTINUOUS ATTRIBUTES

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

Normal distribution:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2}\ _{ij}^{2}} e^{\frac{(A_i\ \ _{ij})^2}{2\ _{ij}^{2}}}$$

One for each (A_i, c_i) pair

For (Income, Class = No):

If Class = No

Sample mean = 110

Sample variance = 2,975

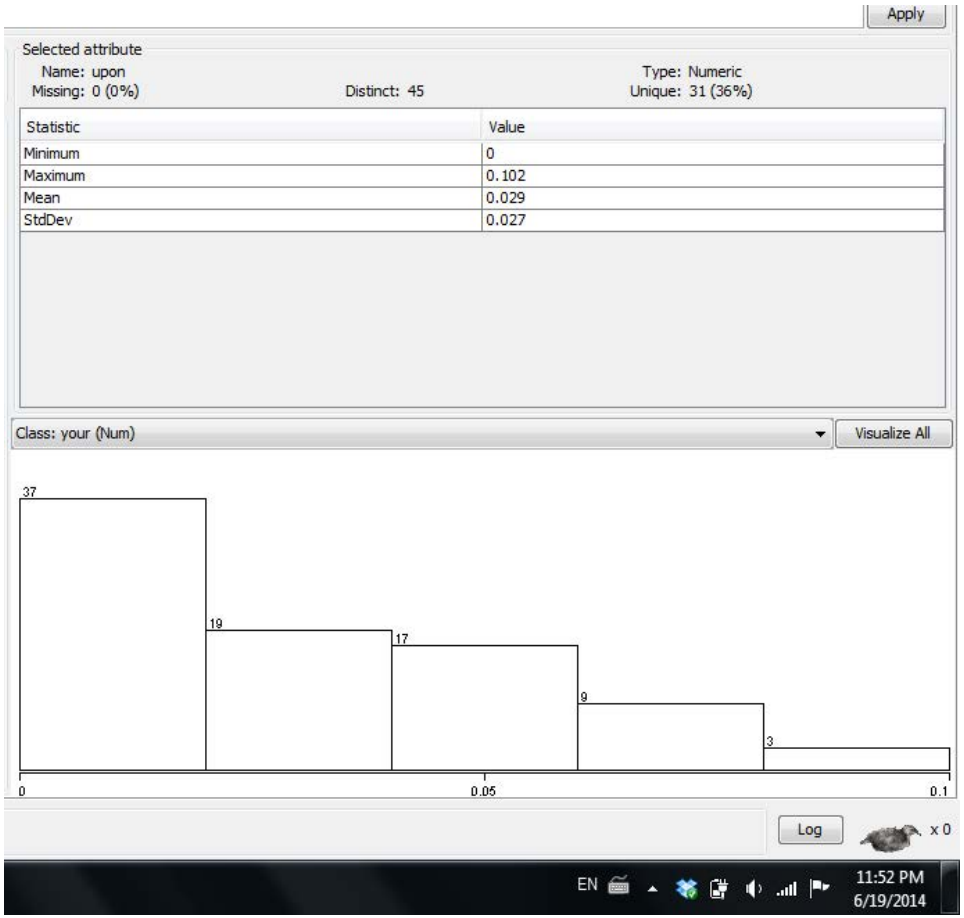
$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2}\ (54.54)} e^{\frac{(120\ \ 110)^2}{2(2975)}} = 0.0072$$

HOW DO I KNOW IF A VARIABLE FOLLOWS NORMAL DISTRIBUTION?

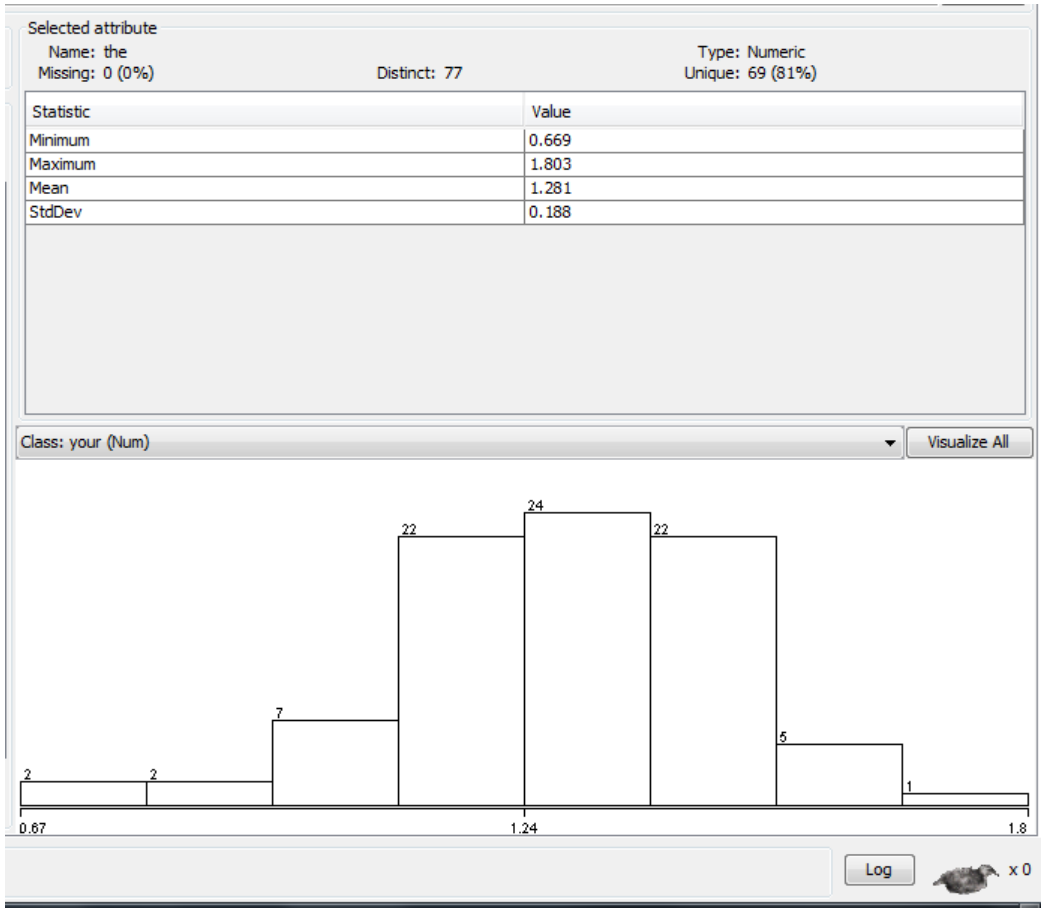
Approach 1: Use statistics test.

Approach 2 (recommended to our class): Use visualization. (Does it look like a bell curve?)

A VARIABLE THAT SEEMS NOT TO FOLLOW NORMAL DISTRIBUTION



A VARIABLE THAT SEEMS TO FOLLOW NORMAL DISTRIBUTION



NORMAL VS. KERNEL DENSITY ESTIMATION

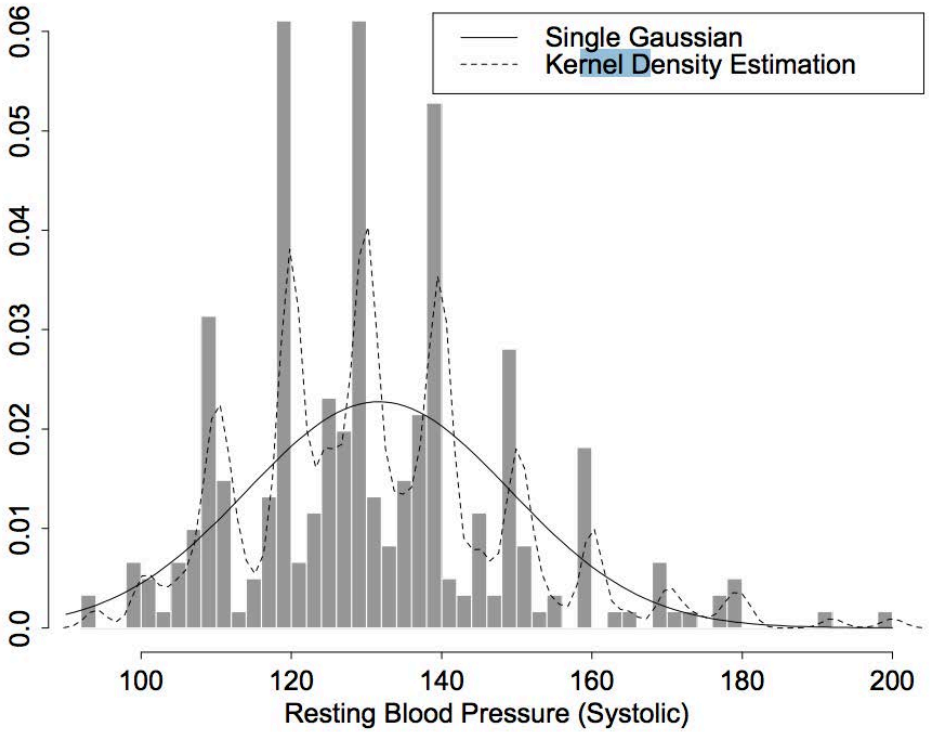


Figure 3: Systematic measurement errors in the Cleveland heart disease database.

Normal distribution and Gaussian distribution are two names for the same thing.

http://www.cs.iastate.edu/~honavar/bayes-continuous.pdf

SUMMARY OF NAIVE BAYES

SYRACUSE UNIVERSITY
School of Information Studies

BUILD A NAIVE BAYES CLASSIFIER

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

P(refund=no) = 7/10
P(refund=yes) = 3/10

Prior probability

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/3
P(Marital Status=Divorced|Yes)=1/3
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

Conditional probability

USE NAIVE BAYES CLASSIFIER FOR PREDICTION

Given a test record, calculate posterior probabilities, and choose decision with maximum posterior probabilities.

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

P(X | Class=No) = P(Refund=No | Class=No)
× P(Married | Class=No)
× P(Income=120K | Class=No)
= 4/7 × 4/7 × 0.0072 = 0.0024

P(Class=No | X) =P(X | Class=No) x P(Class=No) / P(X)

USE NAIVE BAYES CLASSIFIER FOR PREDICTION

Given a test record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

P(X|Class=No) = P(Refund=No|Class=No)
× P(Married|Class=No)
× P(Income=120K|Class=No)
= 4/7 × 4/7 × 0.0072 = 0.0024

P(Class=No|X) =P(X|Class=No) x P(Class=No) / P(X)

P(X|Class=Yes) = P(Refund=No|Class=Yes)
× P(Married|Class=Yes)
× P(Income=120K|Class=Yes)
= 1 × 0 × 1.2 × 10^{-9} = 0

P(Class=Yes|X) =P(X|Class=Yes) x P(Class=Yes) / P(X)

Don't forget smoothing!

FEATURES OF BAYESIAN LEARNING METHODS

Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct, and therefore the algorithm is robust to inconsistent examples.

Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.

E.g., an unbalanced coin has 60% chance heads, 40% tails.

Bayesian methods provide probabilistic predictions.

E.g., “This pneumonia patient has a 93% chance of complete recovery.”

CHALLENGE OF BAYESIAN METHODS

Practical difficulty

- Requires initial knowledge of many probabilities
- Estimates the probabilities when they are unknown
- May need to assume normal distribution for continuous variables

Significant cost to compute all probabilities

- Specialized assumptions to reduce the computational cost
  - E.g., naive Bayes is fast.
- Independence assumption may not hold for some attributes
  - Use other techniques, such as Bayesian belief networks (BBN).

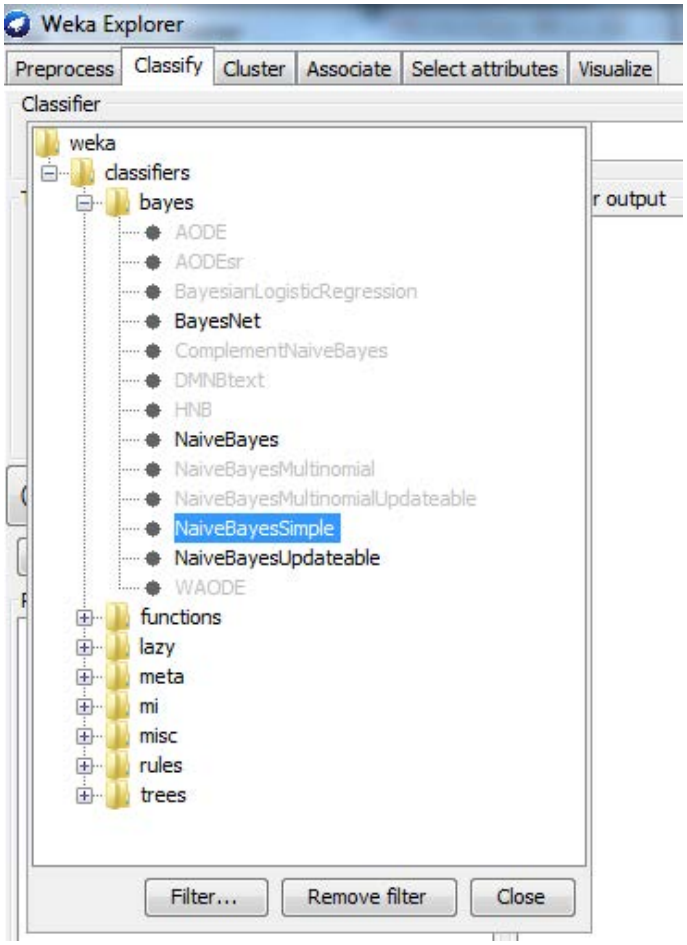
Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–30.

Mitchell, T. (1990). *Machine learning*. New York: McGraw-Hill.

VARIATIONS IN NAIVE BAYES IMPLEMENTATIONS

SYRACUSE UNIVERSITY
School of Information Studies

WEKA'S IMPLEMENTATION OF NAIVE BAYES ALGORITHM



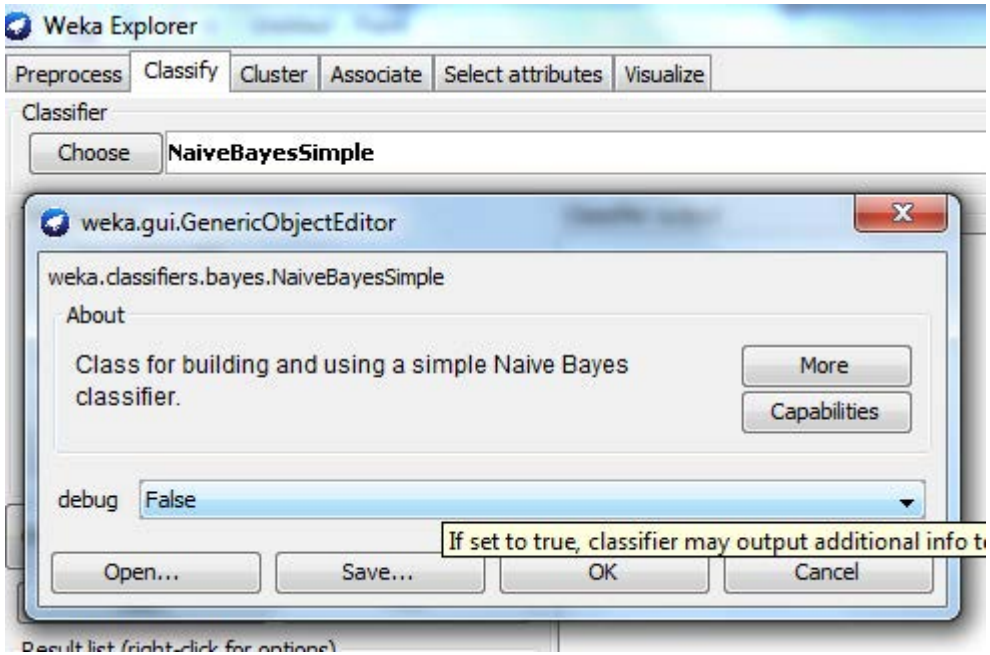
Many different versions of naive Bayes' algorithms.

- NaiveBayesSimple
- NaiveBayes
- NaiveBayesUpdateable
- NaiveBayesMultinomial

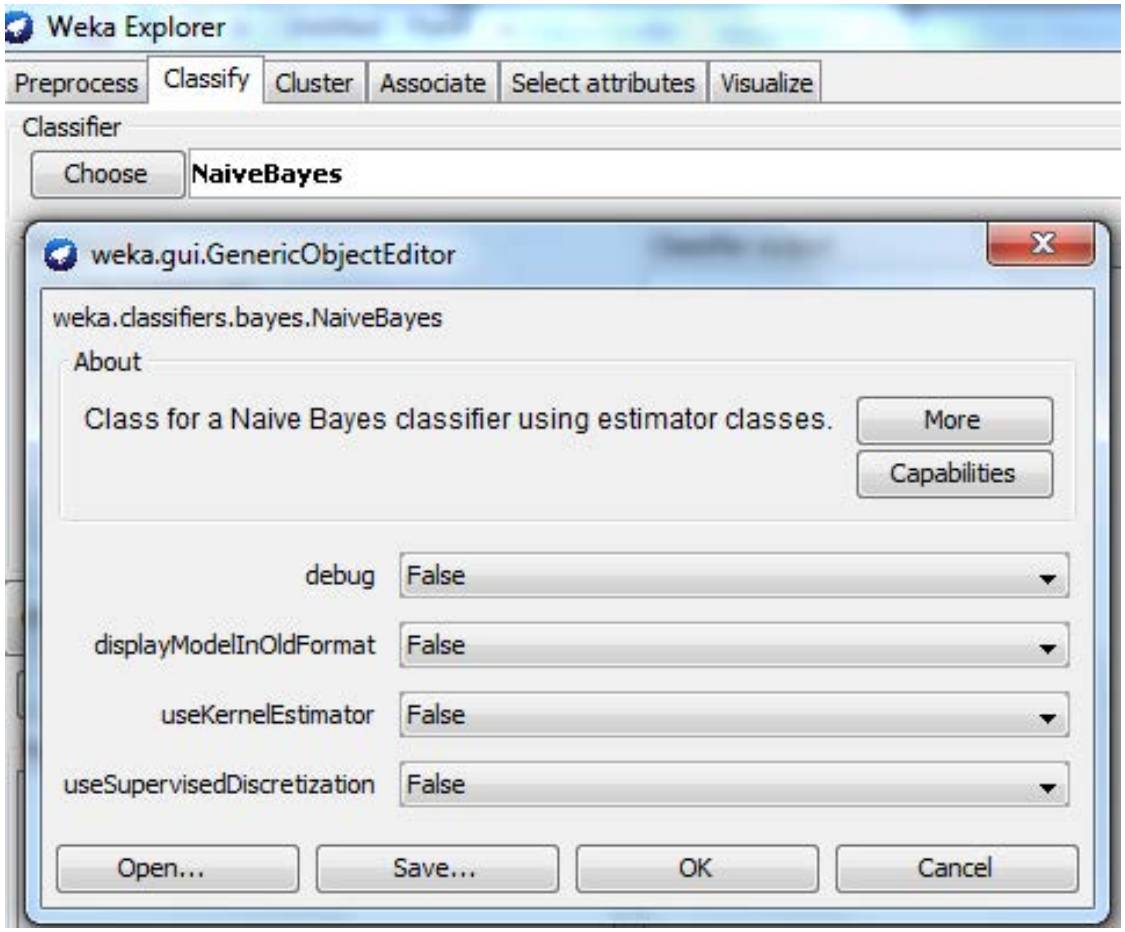
What are their differences?

NAIVE BAYES SIMPLE

NaiveBayesSimple assumes all numeric variables follow normal distribution, and thus does not allow discretization. There are no performance-related parameters to tune in this algorithm.



NAIVE BAYES


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
weka.gui.GenericObjectEditor
weka.classifiers.bayes.NaiveBayes
About
Class for a Naive Bayes classifier using estimator classes.
More
Capabilities
debug
False
displayModelInOldFormat
False
useKernelEstimator
False
useSupervisedDiscretization
False
Open...
Save...
OK
Cancel

NAIVE BAYES

NaiveBayes allows you to choose from three different methods to handle numeric variables.

- Method 1: Use default parameter setting; assume they follow normal distribution (same as NaiveBayesSimple).
- Method 2: Turn on the "useKernelEstimator" option to use kernel estimation for numeric variables. The kernel method does not assume normal distribution.
- Method 3: Turn on the "useSupervisedDiscretization" option to discretize numeric variables.

Which method to choose?

- If your numeric variables follow normal distribution, choose method 1 or NaiveBayesSimple.
- Otherwise, try the other two methods.

NAIVE BAYES UPDATEABLE

Sometimes training data may trickle in, instead of coming in a batch. It is computationally expensive to retrain the entire model whenever a new training example is added to the training set.

NaiveBayesUpdateable can take new training examples and “update” the existing model without complete retraining.

NAIVE BAYES MULTINOMIAL

This algorithm is particularly designed for text categorization. For implementation details, see the supplementary reading mitchell-nb-text-classifier.pdf, an excerpt from Tom Mitchell's textbook on machine learning.

THE E1071 PACKAGE IN R

Library(e1071)

https://cran.r-project.org/web/packages/e1071/e1071.pdf

BAYES' THEOREM IN THE NEWS

MIT Technology Review: "How Statisticians Found Air France Flight 447 Two Years After It Crashed Into Atlantic"

http://www.technologyreview.com/view/527506/how-statisticians-found-air-france-flight-447-two-years-after-it-crashed-into-atlantic/

NPR News: "Can a 250-Year-Old Mathematical Theorem Find a Missing Plane?"

http://www.npr.org/blogs/thetwo-way/2014/03/25/294390476/can-a-250-year-old-mathematical-theorem-find-a-missing-plane