

IST707 Applied Machine Learning

Naïve Bayes Classifiers

Classification Techniques

Many classification algorithms have been developed to date.

This class will introduce the details of several most popular algorithms

Decision Tree

Bayesian method (naïve Bayes)



Instance-based learning (k-Nearest Neighbor)

Support Vector Machines (SVMs)

- In this week, we illustrate classification tasks using **Bayesian method (naïve Bayes)** methods

Bayes Theorem

The Naive Bayes Classifier technique is based on the Bayesian theorem.

Before we can understand how these classifiers work, we must understand:

- How does the Bayes Theorem work mathematically?

Then we can study:

- How we can use these Naive Bayes classifiers techniques in Data Mining.

Review of probability concepts

An **event** is an outcome of an experiment, e.g.

- Experiment: toss a fair coin
- Possible outcomes: “head” or “tail”.
- The **probability** that either event (“head” or “tail”) occurs is 50%, i.e.:
 - $P(\text{head})=0.5$, $P(\text{tail})=0.5$
 - $P(\text{head}) + P(\text{tail}) = 1$

Two Additional Concepts.

- **Joint Probability**
- **Conditional Probability**

Joint probability

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

- Experiment: toss two fair coins
 - Event A: coin1="head"
 - Event B: coin2="head"
- $P(A, B) = P(\text{coin1} = \text{"head"}, \text{coin2} = \text{"head"})$

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

Independent events

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

E.g, toss 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,

$$P(\text{card1}=\text{"queen"}, \text{card2}=\text{"queen"}) \neq P(\text{card1}=\text{"queen"}) * P(\text{card2}=\text{"queen"})$$

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"
- $P(B | A) = P(\text{card2}=\text{"queen"} \mid \text{card1}=\text{"queen"}) = 3/51$,
because there are 51 remaining cards and only 3 queens
among them.

Relationship between conditional and joint probabilities

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

$$P(\text{card1}=\text{"queen"}, \text{card2}=\text{"queen"})$$

$$= P(\text{card1}=\text{"queen"}) * P(\text{card2}=\text{"queen"} | \text{card1}=\text{"queen"})$$

$$= 4/52 * 3/51$$

$$= 0.5\%$$

Probability is useful in daily life

Which one is greater?

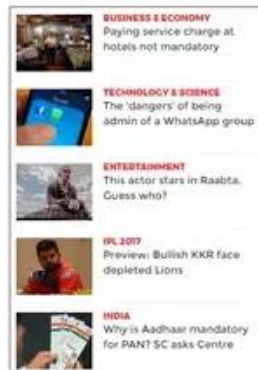
- $P(\text{"will"} \mid \text{"I"})$
- $P(\text{"has"} \mid \text{"I"})$

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.

Real World Applications using Naïve Bayes

Categorizing News



Email Spam Detection



Face Recognition



Sentiment Analysis



Medical Diagnosis



Digit Recognition



Weather Prediction



Exercise: calculate probability

Choose a card from a standard deck of 52 cards.
Without replacement, choose another card. What is the probability $P(\text{card1}=\text{"queen"}, \text{card2}=\text{"king"})$?

$$4/52 * 4/51$$

Bayes Theorem

Bayes Theorem lets us swap the order of the dependence between events

- Two events A and B
- We know that $P(A, B) = P(B | A)P(A)$
- We also know that $P(A, B) = P(A | B)P(B)$
- Since $P(A, B) = P(B, A)$,

$$P(A | B) \times P(B) = P(B | A) \times P(A)$$

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

The diagram illustrates the components of Bayes' Theorem using orange arrows and labels:

- An arrow points from the label "Likelihood" to the term $P(B | A)$ in the numerator.
- An arrow points from the label "Prior" to the term $P(A)$ in the numerator.
- An arrow points from the label "Evidence" to the term $P(B)$ in the denominator.

Chances are...

Read Professor Steven Strogatz's blog article on New York Times "Chances Are"

- <http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/>
- This article explains Bayes Theorem in lay-person's language

Professor Steven Strogatz's blog article on New York Times "Chances Are"

In one study, Gigerenzer and his colleagues asked doctors to estimate the probability that a woman with a positive mammogram actually has breast cancer

The probability that one of these women has breast cancer is 0.8 percent.

- *If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram.*
- *If a woman does not have breast cancer, the probability is 7 percent that she will still have a positive mammogram.*

Class Exercise

Estimate the probability that a woman with a positive mammogram actually has breast cancer

What do each of you believe is the probability that she actually has breast cancer?

Please do not look ahead !! 😊

Professor Steven Strogatz's blog article on New York Times "Chances Are"

When Gigerenzer asked 24 German doctors about the likelihood of cancer given a positive mammogram, their estimates whipsawed from 1 percent to 90 percent. Eight of them thought the chances were 10 percent or less, 8 more said 90 percent, and the remaining 8 guessed somewhere between 50 and 80 percent.

As for the American doctors, 95 out of 100 estimated the woman's probability of having breast cancer to be somewhere around 75 percent.

Imagine how upsetting it would be as a patient to hear such divergent opinions.

Professor Steven Strogatz's blog article on New York Times "Chances Are"

The right answer is 9 percent.

How can it be so low?

Gigerenzer's point is that the analysis becomes almost transparent if we translate the original information from percentages and probabilities into natural frequencies:

How This Works

Eight out of every 1,000 women have breast cancer. Of these 8 women with breast cancer, 7 will have a positive mammogram.

Of the remaining 992 women who don't have breast cancer, some 70 will still have a positive mammogram.

Since a total of $7 + 70 = 77$ women have positive mammograms, and only 7 of them truly have breast cancer, the probability of having breast cancer given a positive mammogram is 7 out of 77, which is 1 in 11, or about 9 percent.

Using Bayes Formula

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

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 result?

- We can calculate it from past data

$P(A | B)$ = For a patient with 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

$P(\text{cancer})$: The (prior) probability that a woman has breast cancer is ?

$P(\text{positive} | \text{cancer})$: The likelihood that a woman with breast cancer will have a positive mammogram is ?

$P(\text{positive} | \text{no cancer})$: The probability that she have a positive mammogram even without cancer?

The mammogram example

$P(\text{cancer})$: The (prior) probability that a woman has breast cancer is ? .008

$P(\text{positive} | \text{cancer})$: The likelihood that a woman with breast cancer will have a positive mammogram is ? .9

$P(\text{positive} | \text{no cancer})$: The probability that she have a positive mammogram even without cancer? .07

Posterior probability

The posterior probability is one of the quantities involved in Bayes' rule.

It is the conditional probability of a given event, computed after observing a second event whose conditional and unconditional probabilities were known in advance.

It is computed by revising the prior probability, that is, the probability assigned to the first event before observing the second event.

<https://www.statlect.com/glossary/posterior-probability>

Translate them into probability notations

All **prior probabilities** we have calculated

- $P(\text{cancer}) = 0.008$
- $P(\text{no cancer}) = 0.992$

All **conditional probabilities** we have calculated

- $P(\text{positive} \mid \text{cancer}) = 0.9$
- $P(\text{positive} \mid \text{no cancer}) = 0.07$

The **posterior probabilities** to be calculated

- $P(\text{cancer} \mid \text{positive}) = ?$
- $P(\text{no cancer} \mid \text{positive}) = ?$

So, our prediction is ...

$$\begin{aligned} &P(\text{cancer} \mid \text{positive}) \\ &= \frac{P(\text{positive} \mid \text{cancer}) \times P(\text{cancer})}{P(\text{positive})} = \frac{0.9 \times 0.008}{P(\text{positive})} = \frac{0.0072}{P(\text{positive})} \end{aligned}$$

unknown

$$\begin{aligned} &P(\text{no_cancer} \mid \text{positive}) \\ &= \frac{P(\text{positive} \mid \text{no_cancer}) \times P(\text{no_cancer})}{P(\text{positive})} = \frac{0.07 \times 0.992}{P(\text{positive})} = \frac{0.069}{P(\text{positive})} \end{aligned}$$

unknown

No cancer!

$$\begin{aligned} &P(\text{cancer} \mid \text{positive}) \\ &= \frac{P(\text{positive} \mid \text{cancer}) \times P(\text{cancer})}{P(\text{positive})} = \frac{0.9 \times 0.008}{P(\text{positive})} = \frac{0.0072}{P(\text{positive})} \end{aligned}$$

$$\begin{aligned} &P(\text{no_cancer} \mid \text{positive}) \\ &= \frac{P(\text{positive} \mid \text{no_cancer}) \times P(\text{no_cancer})}{P(\text{positive})} = \frac{0.07 \times 0.992}{P(\text{positive})} = \frac{0.069}{P(\text{positive})} \end{aligned}$$

Although we don't know $P(\text{positive})$, it does not matter. For predictive purposes, we just need to know which posterior probability is greater.

CHALLENGES WITH MULTIPLE VARIABLES

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, $2^* 2^3 = 16$

Challenge of Bayesian classifier

$P(\text{positive, yes, yes} \mid \text{cancer})$

$P(\text{positive, yes, no} \mid \text{cancer})$

$P(\text{positive, no, yes} \mid \text{cancer})$

$P(\text{positive, no, no} \mid \text{cancer})$

$P(\text{negative, yes, yes} \mid \text{cancer})$

$P(\text{negative, yes, no} \mid \text{cancer})$

$P(\text{negative, no, yes} \mid \text{cancer})$

$P(\text{negative, no, no} \mid \text{cancer})$

$P(\text{positive, yes, yes} \mid \text{cancer})$

$P(\text{positive, yes, no} \mid \text{no_cancer})$

$P(\text{positive, no, yes} \mid \text{no_cancer})$

$P(\text{positive, no, no} \mid \text{no_cancer})$

$P(\text{negative, yes, yes} \mid \text{no_cancer})$

$P(\text{negative, yes, no} \mid \text{no_cancer})$

$P(\text{negative, no, yes} \mid \text{no_cancer})$

$P(\text{negative, no, no} \mid \text{no_cancer})$

2ⁿ possible combinations of values for n binary attributes

Naïve Bayes Classifier

Assume independence among attributes A_i when class is given:

$$\begin{aligned} - P(A_1, A_2, \dots, A_n | C) &= P(A_1 | C_j) * P(A_2 | C_j) * \dots * P(A_n | C_j) \\ &= \prod P(A_i | C_j) \end{aligned}$$

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

For example, for people with cancer, the assumption is 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:

$$\begin{aligned} - P(A_1, A_2, \dots, A_n | C) &= P(A_1 | C_j) P(A_2 | C_j) \dots P(A_n | C_j) \\ &= \prod P(A_i | C_j) \end{aligned}$$

To greatly reduce the number of probabilities to estimate

Why is the independence assumption needed?

$$P(\text{positive}, \text{yes}_{\text{family}}, \text{yes}_{\text{race}} | \text{cancer})$$
$$= P(\text{positive} | \text{cancer}) \times P(\text{yes}_{\text{family}} | \text{cancer}) \times P(\text{yes}_{\text{race}} | \text{cancer})$$

$P(\text{positive}, \text{yes}, \text{yes} | \text{cancer})$
 $P(\text{positive}, \text{yes}, \text{no} | \text{cancer})$
 $P(\text{positive}, \text{no}, \text{yes} | \text{cancer})$
 $P(\text{positive}, \text{no}, \text{no} | \text{cancer})$
 $P(\text{negative}, \text{yes}, \text{yes} | \text{cancer})$
 $P(\text{negative}, \text{yes}, \text{no} | \text{cancer})$
 $P(\text{negative}, \text{no}, \text{yes} | \text{cancer})$
 $P(\text{negative}, \text{no}, \text{no} | \text{cancer})$

Reduce
calculation
 $2^3 \rightarrow 2*3$

$P(\text{positive} | \text{cancer})$
 $P(\text{yes}_{\text{family}} | \text{cancer})$
 $P(\text{yes}_{\text{race}} | \text{cancer})$
 $P(\text{negative} | \text{cancer})$
 $P(\text{no}_{\text{family}} | \text{cancer})$
 $P(\text{no}_{\text{race}} | \text{cancer})$

Similarly,

$P(\text{positive}, \text{yes}, \text{yes} \mid \text{no_cancer})$
 $P(\text{positive}, \text{yes}, \text{no} \mid \text{no_cancer})$
 $P(\text{positive}, \text{no}, \text{yes} \mid \text{no_cancer})$
 $P(\text{positive}, \text{no}, \text{no} \mid \text{no_cancer})$
 $P(\text{negative}, \text{yes}, \text{yes} \mid \text{no_cancer})$
 $P(\text{negative}, \text{yes}, \text{no} \mid \text{no_cancer})$
 $P(\text{negative}, \text{no}, \text{yes} \mid \text{no_cancer})$
 $P(\text{negative}, \text{no}, \text{no} \mid \text{no_cancer})$

Reduce
calculation
 $2^3 \rightarrow 2*3$

$P(\text{positive} \mid \text{no_cancer})$
 $P(\text{yes}_{\text{family}} \mid \text{no_cancer})$
 $P(\text{yes}_{\text{race}} \mid \text{no_cancer})$
 $P(\text{negative} \mid \text{no_cancer})$
 $P(\text{no}_{\text{family}} \mid \text{no_cancer})$
 $P(\text{no}_{\text{race}} \mid \text{no_cancer})$

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

If $n=10$, 2048 probabilities reduced to 40!

Why does “naïve” Bayes work?

This assumption is often violated in real-world problems

- E.g. Family history of cancer may be correlated with race

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

Domingos, P. and 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)

Class Exercise

Exercise: estimate prior probabilities from training data

Example	Positive mammogram	Family history	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

Given the above training data set:

Calculate prior probability for each class: $P(C) = N_c/N$

$P(\text{cancer}=\text{yes})=?$

$P(\text{cancer}=\text{no})=?$

Please do not look ahead !! 😊

Exercise: calculate posterior probabilities

Given a test case (pos_mammo=yes, fam_hist=yes, alcohol=yes), what is the prediction, cancer or no cancer?

$P(\text{cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

vs.

$P(\text{no cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

Exercise: estimate conditional probabilities from training data

Then calculate the conditional probabilities for each attribute

$P(A_i | C_k) = |A_{ik}| / N_c$ E.g.

$P(\text{pos_mammo}=\text{yes} | \text{cancer}=\text{yes}) = ?$

$P(\text{pos_mammo}=\text{no} | \text{cancer}=\text{yes}) = ?$

$P(\text{pos_mammo}=\text{yes} | \text{cancer}=\text{no}) = ?$

$P(\text{pos_mammo}=\text{no} | \text{cancer}=\text{no}) = ?$

Exercise: estimate conditional probabilities from training data

Repeat for the “family_history” attribute:

$P(\text{family_history}=\text{yes} \mid \text{cancer}=\text{yes}) = ?$

$P(\text{family_history}=\text{no} \mid \text{cancer}=\text{yes}) = ?$

$P(\text{family_history}=\text{yes} \mid \text{cancer}=\text{no}) = ?$

$P(\text{family_history}=\text{no} \mid \text{cancer}=\text{no}) = ?$

Exercise: estimate conditional probabilities from training data

Repeat for the “Alcohol” attribute:

$P(\text{Alcohol} = \text{yes} \mid \text{cancer} = \text{yes}) = ?$

$P(\text{Alcohol} = \text{no} \mid \text{cancer} = \text{yes}) = ?$

$P(\text{Alcohol} = \text{yes} \mid \text{cancer} = \text{no}) = ?$

$P(\text{Alcohol} = \text{no} \mid \text{cancer} = \text{no}) = ?$

Exercise: calculate posterior probabilities

Given a test case (pos_mammo=yes, fam_hist=yes, alcohol=yes), what is the prediction, cancer or no cancer?

$P(\text{cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

vs.

$P(\text{no cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

Exercise: calculate posterior probabilities

$$\begin{aligned} & P(\text{cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes}) \\ &= P(\text{pos_mammo}=\text{yes} \mid \text{cancer}) * \\ &\quad P(\text{fam_hist}=\text{yes} \mid \text{cancer}) * \\ &\quad P(\text{alcohol}=\text{yes} \mid \text{cancer}) * \\ &\quad P(\text{cancer}) / \\ &\quad P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes}) \\ &= (2/3 * 3/3 * 2/3 * 3/7) / \\ &\quad P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes}) \\ &= (\mathbf{12/63}) / \\ &\quad P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes}) \end{aligned}$$

Exercise: calculate posterior probabilities

Similarly,

$P(\text{no cancer} \mid \text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

$= P(\text{pos_mammo}=\text{yes} \mid \text{no cancer}) *$

$P(\text{fam_hist}=\text{yes} \mid \text{no cancer}) *$

$P(\text{alcohol}=\text{yes} \mid \text{no cancer}) *$

$P(\text{no cancer}) /$

$P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

$= (2/4 * 0/4 * 2/4 * 4/7) /$

$P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

$= (0) /$

$P(\text{pos_mammo}=\text{yes}, \text{fam_hist}=\text{yes}, \text{alcohol}=\text{yes})$

Exercise: calculate posterior probabilities

Prediction: **cancer!**

- **12/63 : 0**
- Choose the decision with max posterior probability

SMOOTHING

Problem of zero probabilities

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

$$P(\text{fam_hist}=\text{yes} \mid \text{cancer}=\text{yes}) = 1$$

$$P(\text{fam_hist}=\text{no} \mid \text{cancer}=\text{yes}) = 0$$

$$P(\text{fam_hist}=\text{yes} \mid \text{cancer}=\text{no}) = 0$$

$$P(\text{fam_hist}=\text{no} \mid \text{cancer}=\text{no}) = 1$$

But the zero prob may just be caused by lack of data

Solution to zero prob

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

Since all probabilities should add up to one, the other non-zero probabilities need to “shrink” a little bit, in order to “lend” the small amount to the zero probabilities.

Such technique is called “smoothing”

Smoothing for zero probabilities

using a **smoothing** algorithm, such as Laplace smoothing, or called “add-one” smoothing

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

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:

$$P(\text{fam_hist=yes} \mid \text{cancer=yes}) = 3/3$$

$$P(\text{fam_hist=no} \mid \text{cancer=yes}) = 0/3 = 0$$

$$P(\text{fam_hist=yes} \mid \text{cancer=no}) = 0/4 = 0$$

$$P(\text{fam_hist=no} \mid \text{cancer=no}) = 4/4 = 1$$

smoothed probabilities:

$$P(\text{fam_hist=yes} \mid \text{cancer=yes}) = 4/5$$

$$P(\text{fam_hist=no} \mid \text{cancer=yes}) = 1/5$$

$$P(\text{fam_hist=yes} \mid \text{cancer=no}) = 1/6$$

$$P(\text{fam_hist=no} \mid \text{cancer=no}) = 5/6$$

The non-zero probs become smaller, “lending” values to zero probs

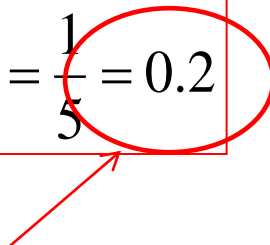
Limitations with add-one smoothing

Small numbers might mean that smoothing has a bit impact

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

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

Well this is not a very small probability??

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


Smoothing for zero probabilities

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

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

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

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

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

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 like in 0.000222.

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

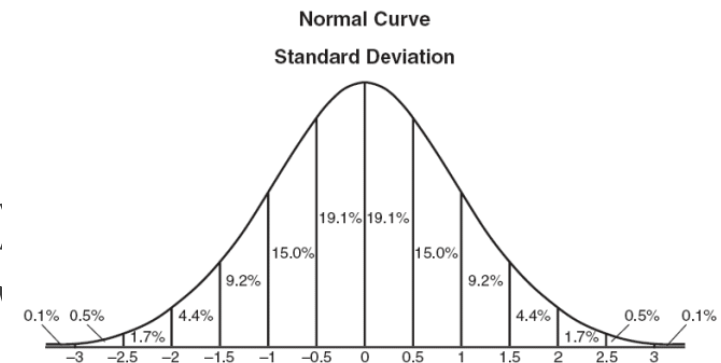
$$\log\left(\frac{1+1}{7000+2}\right) = \log\left(\frac{2}{7002}\right) = -8.161$$

PROBABILITY FOR A CONTINUOUS VARIABLE

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:

$$P(A_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- One for each (A_i, c_i) pair

For (Income, Class=No):

- If Class=No
 - sample mean = 110
 - sample variance = 2975

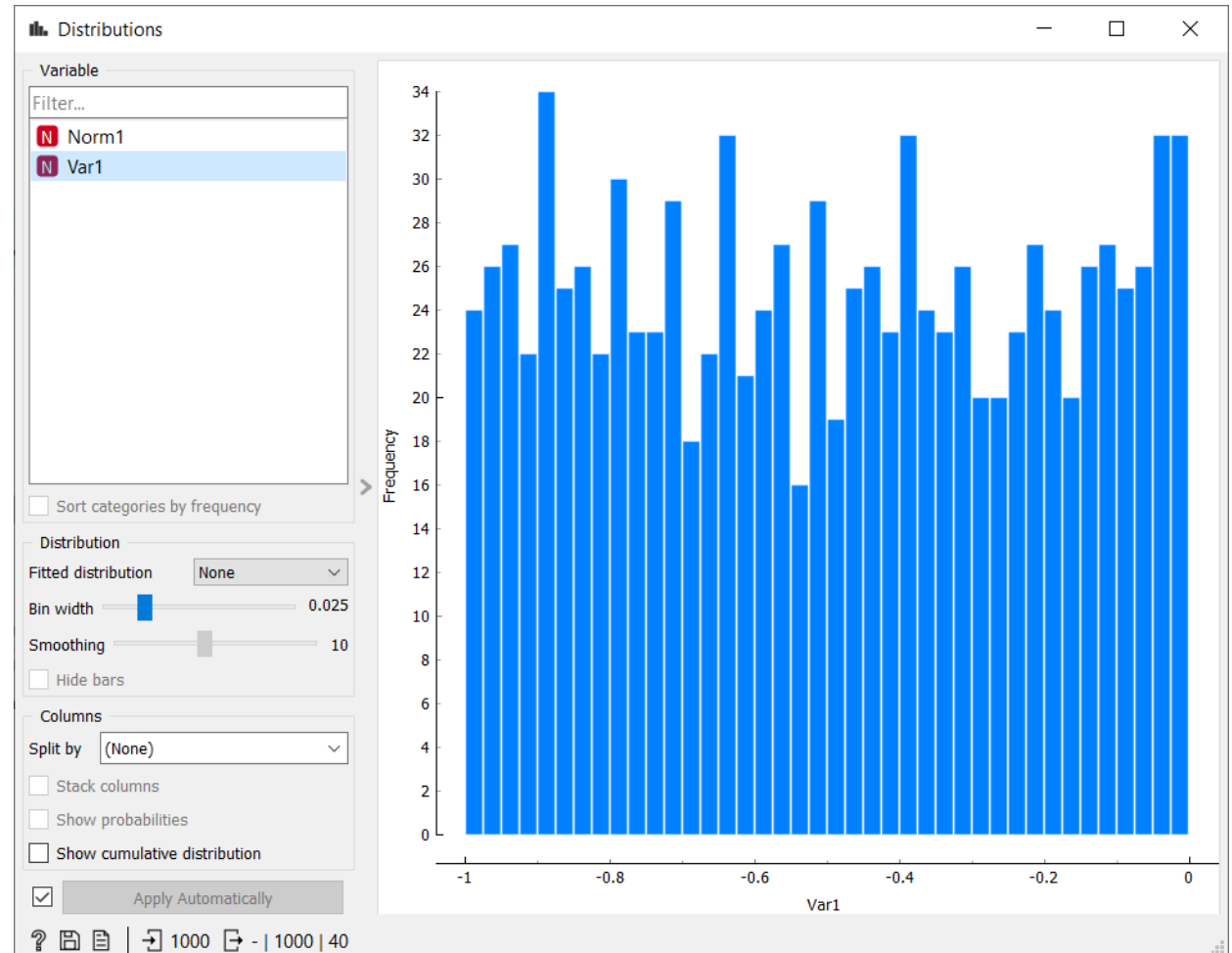
$$P(\text{Income} = 120 | \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

How do I know if a variable follows normal distribution?

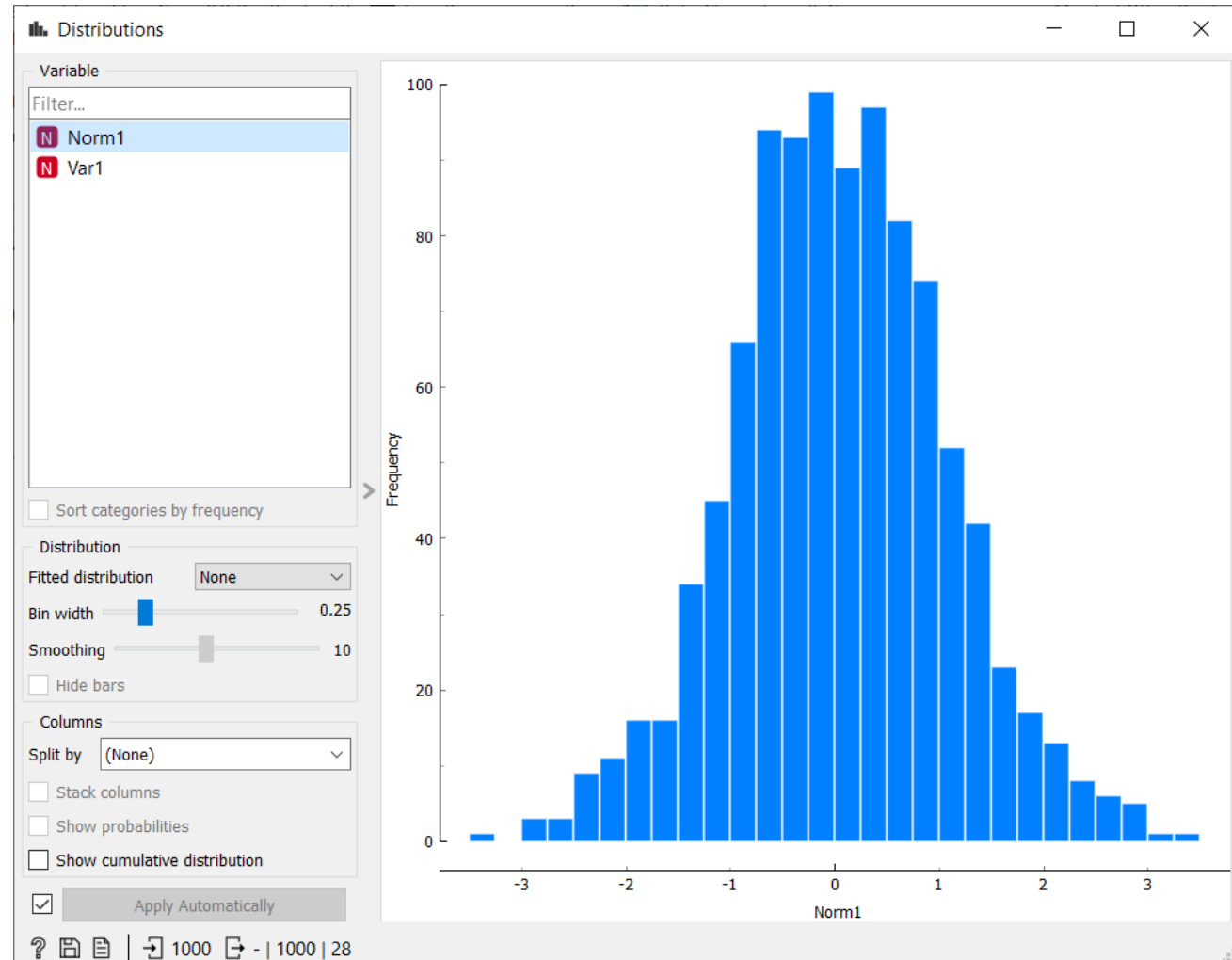
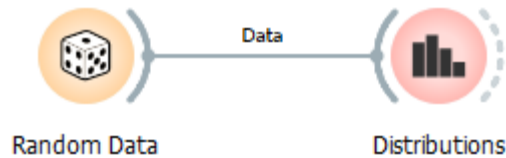
Approach 1: use statistics test

Approach 2 (recommended to our class) : use visualization (does it look at a bell curve?)

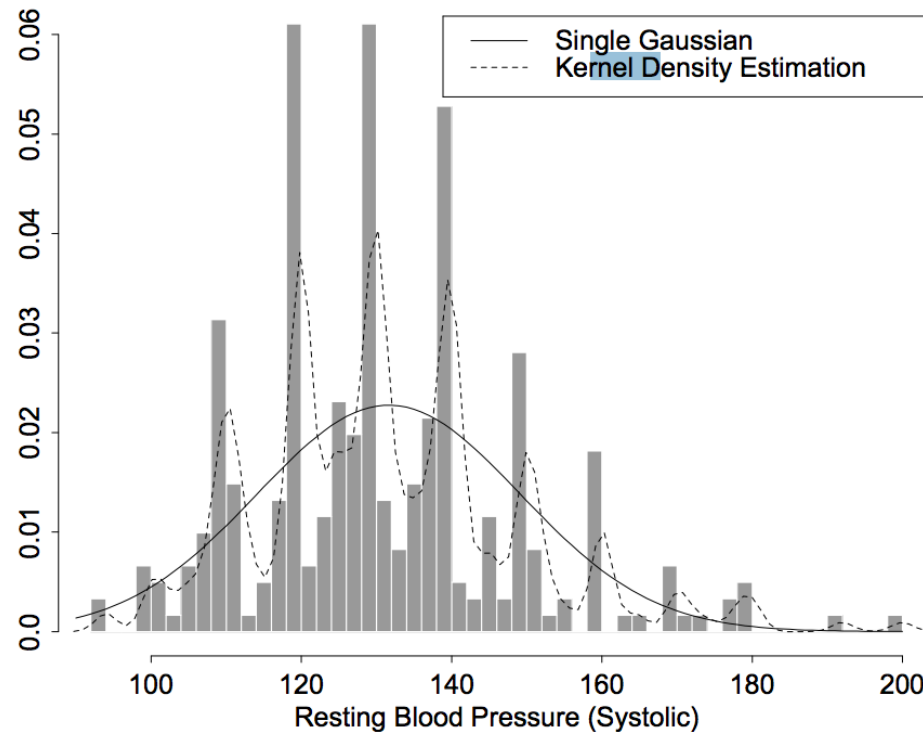
A variable that seems not normal distribution



A variable that seems follow normal distribution



Normal vs. Kernel density estimation



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

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

SUMMARY OF NAÏVE BAYES

Build a naïve Bayes classifier

<i>Tid</i>	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

$P(\text{refund}=\text{No})=7/10,$
 $P(\text{refund}=\text{yes})=3/10$

Prior probability

naive Bayes Classifier:

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

For taxable income:

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

conditional probability

Use naïve Bayes Classifier for prediction

Given a Test Record, calculate posterior probs, and choose decision with max posterior prob

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

$$\begin{aligned} P(X|\text{Class}=\text{No}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$P(\text{Class}=\text{No}|X) = P(X|\text{Class}=\text{No}) \times P(\text{Class}=\text{No}) / P(X)$$

Use naïve Bayes Classifier for prediction

Given a Test Record:

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

$$\begin{aligned} P(X|\text{Class}=\text{No}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{No}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$P(\text{Class}=\text{No}|X) = P(X|\text{Class}=\text{No}) \times P(\text{Class}=\text{No}) / P(X)$$

$$\begin{aligned} P(X|\text{Class}=\text{Yes}) &= P(\text{Refund}=\text{No}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Married}|\text{Class}=\text{Yes}) \\ &\quad \times P(\text{Income}=120\text{K}|\text{Class}=\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

$$P(\text{Class}=\text{Yes}|X) = P(X|\text{Class}=\text{Yes}) \times P(\text{Class}=\text{Yes}) / P(X)$$

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 60% chance head 40% tail

Bayesian methods provide probabilistic predictions

- E.g. “this pneumonia patient has a 93% chance of complete recovery”

Challenge of Bayesian methods

Practical difficulty

- require initial knowledge of many probabilities
- Estimate 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. naïve Bayes is fast
- Independence assumption may not hold for some attributes
 - Use other techniques such as Bayesian Belief Networks (BBN)

Domingos, P. and 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*. McGraw-Hill

Bayes Theorem in 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>

Exercise: Orange for Naïve Bayes and Model Evaluation

1. Using the Federalist Papers data, generate and evaluate a Naïve Bayes classifier
 1. Be sure to use the Data Sampler to split the data into training and test sets
2. Similarly, generate a decision tree classifier
3. Testing on the test data, **did the Naïve Bayes classifier perform *better or worse* than the decision tree? Why?**