Chapter 6
Probability-based Learning
Part A

Prof. Chang-Chieh Cheng

Dept. Computer Science

National Chiao Tung University, Taiwan

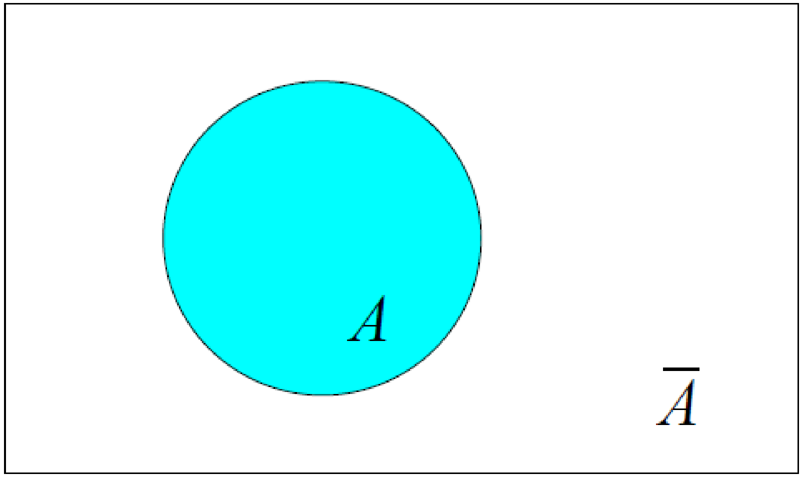
Probability

- Probability is the measure of the likelihood that an event will occur.
- The probability of an event A in a finite sample spaces
  - P(A) = the number of event A occurred / the number of total samples
  - What is the probability of headache in the following ten patients
    - P(Headache) = 7 / 10 = 0.7

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Probability

- The complement of an event
  - What is the probability of non-headache in the ten patients
    - P(non-Headache) = 3 / 10 = 0.3 = 1.0 – P(Headache)

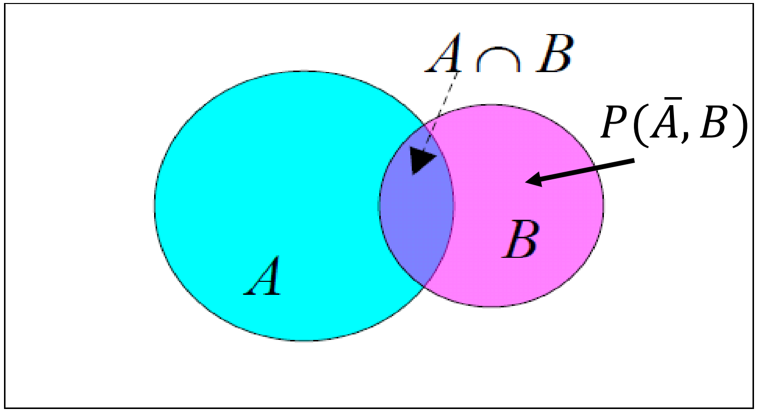


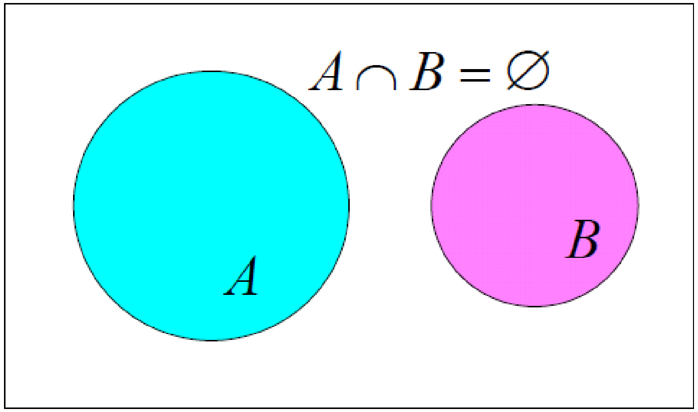
Probability

- $0.0 \leq P(A) \leq 1.0$
- Probability distribution
  - For all *n* random variables X_1, X_2, …, X_n,
    - $\sum_{i=1}^{n} P(X_i) = 1.0$
- EX:
  - Given four weather types: sunny, cloudy, shower, and rain
  - The probabilities for all weather in July 2017 are *P*(sunny), *P*(cloudy), *P*(shower), and *P*(rain) respectively.
  - *P*(sunny) + *P*(cloudy) + *P*(shower) + *P*(rain) = 1.0

Probability

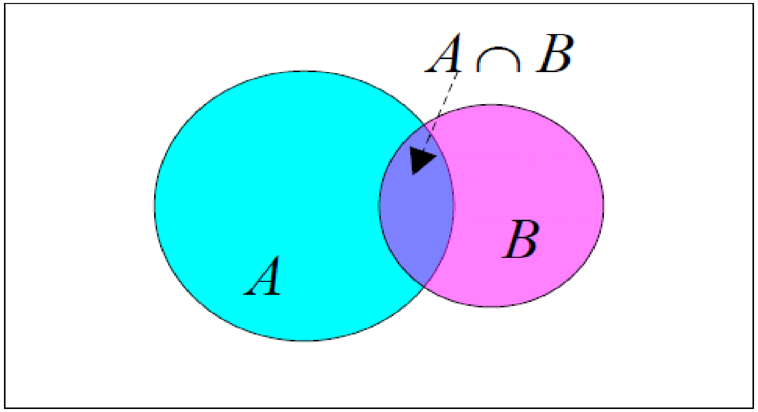
- Joint probability
  - $P(A, B)$ or $P(A \cap B)$ or $P(A \text{ and } B)$
  - if A and B are independent events: $P(A \cap B) = P(A)\,P(B)$
  - $P(A, B) = P(B, A)$
  - $P(B) = P(A, B) + P(\bar{A}, B)$

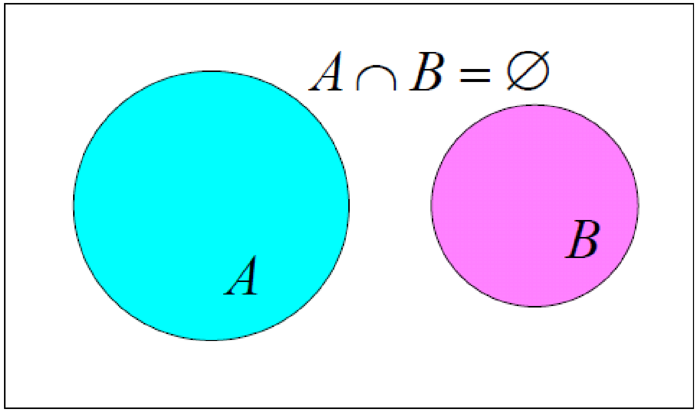




Probability

- $P(A \text{ or } B) = P(A \cup B) = \mathrm{P(A)} + \mathrm{P(B)} - P(A \cap B)$

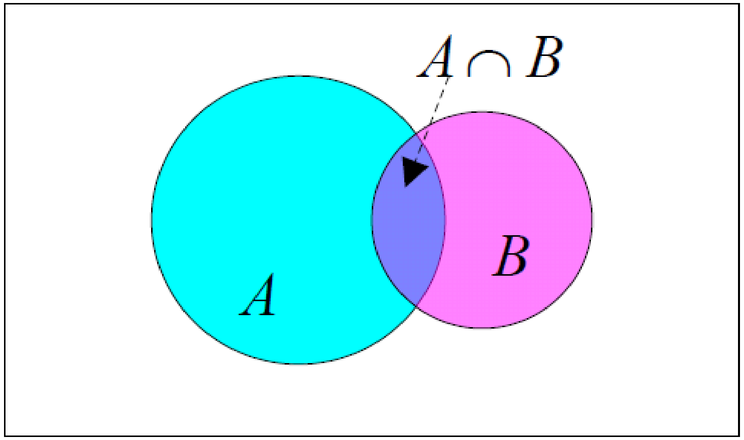




Probability

- Conditional probability
  - $P(A|B)$: the probability of event B under event A occurred

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



Probability

- Conditional probability
  - Bayes' theorem:

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

  - *P*(*A*), *P*(*B*), *P*(*B*|*A*) : prior probability, they are already known.
  - *P*(*A*|*B*): post probability, calculated form priors.
  - Proof:
    - $P(B|A) = \frac{P(B,A)}{P(A)}$
    - $\Rightarrow P(B|A)P(A) = P(B, A)$
    - $\Rightarrow \frac{P(B|A)P(A)}{P(B)} = \frac{P(B,A)}{P(B)} = \frac{P(A,B)}{P(B)} = P(A|B)$

Bayes Theorem Example 1

- Assuming that a school has 60% boys and 40% girls.
- The number of girls wearing pants equals to the number of girls wearing skirts.
- All boy are wearing pants.
- What is the probability of that when you saw a person wearing pants and that person is a girl in the school?
- Let *A* is the event of girl, *B* is the event of pant wearing ➔ The answer is *P*(*A*|*B*)
  - $P(A) = 0.4 \rightarrow P(\bar{A}) = 1 - P(A) = 0.6$, which is the probability of boy
  - $P(B|A) = 0.5$, which is the probability of a girl wearing pants
  - $P(B|\bar{A}) = 1.0$, which is the probability of a boy wearing pants
  - $P(B) = P(B, A) + P(B, \bar{A})$
    $= P(B|A)P(A) + P(B|\bar{A})P(\bar{A}) = 0.5 \times 0.4 + 1.0 \times 0.6 = 0.8$
  - $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8} = 0.25$

Bayes Theorem Example 2

- A doctor informs a patient that he has both bad news and good news.
- The bad news is that the patient has tested positive for a serious disease and that **the test is 99% accurate**
  - the probability is 0.99 ➔ testing positive when a patient has the disease.
  - the probability is 0.01 ➔ testing positive when a patient does not have the disease.
  - the probability is also 0.99 ➔ testing negative when a patient does not have the disease.
- The good news is that the disease is extremely rare, striking **only 1 in 10,000 people**.
- What is the actual probability that the patient has the disease?
- Why is the rarity of the disease good news given that the patient has tested positive for it?

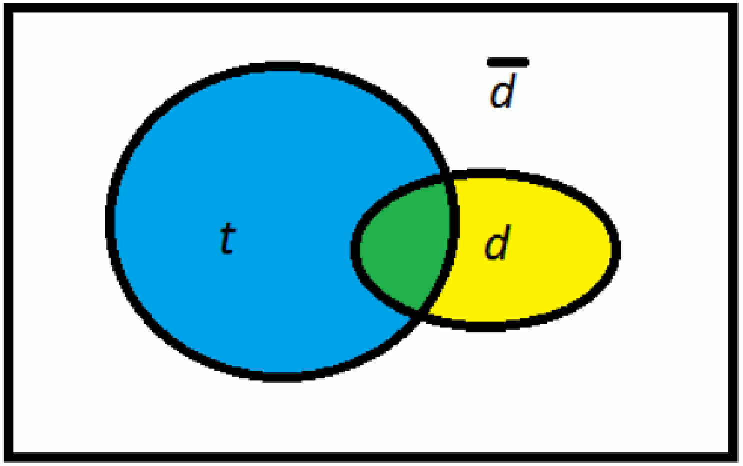
Bayes Theorem Example 2

- d: a patient has the disease
- t: the test is positive
- $P(d|t) = \frac{P(t|d)P(d)}{P(t)}$
- $P(t) = P(t|d)P(d) + P(t|\bar{d})P(\bar{d})$

$$= (0.99 \times 0.001) + (0.01 \times 0.9999)$$

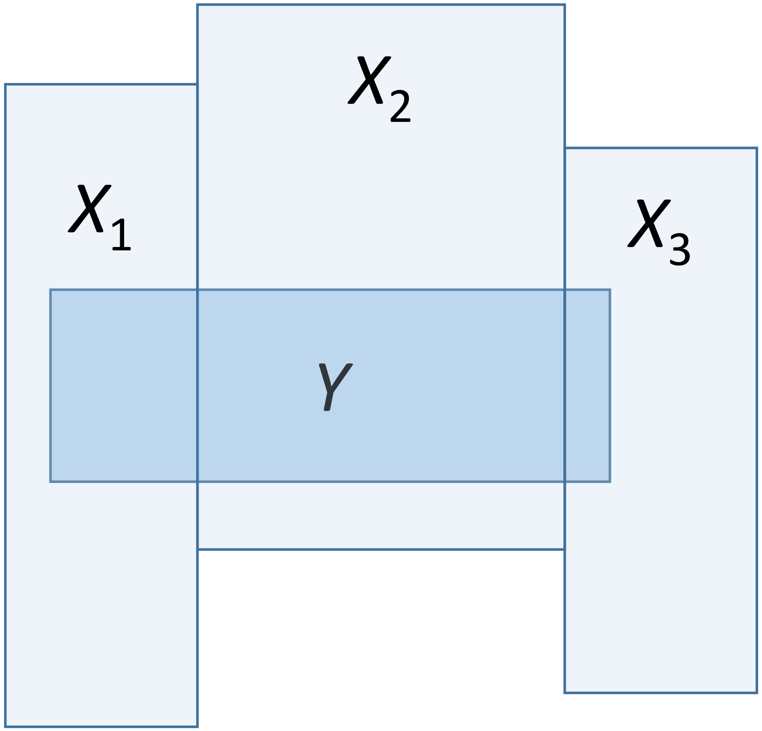
$$= 0.0101$$

- $P(d|t) = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$



Theorem of Total Probability

- $P(Y) = \sum_{i=1}^{n} P(Y|X_i)P(X_i)$
- where $\{X_i: i = 1,2,3 \ldots\}$ is a set of pairwise disjoint events whose union is the entire sample space



Theorem of Total Probability

- Example
  - Suppose that two factories supply light bulbs to the market. Factory X's bulbs work for over 5000 hours in 99% of cases, whereas factory Y's bulbs work for over 5000 hours in 95% of cases.
  - It is known that factory X supplies 60% of the total bulbs available.
  - What is the chance that a purchased bulb will work for longer than 5000 hours?

    - Let *A* is the event of that a purchased bulb will work for longer than 5000 hours

$$P(X) = 0.6, P(Y) = 0.4$$

$$P(A|X) = 0.99, P(A|Y) = 0.95$$

$$P(A) = P(A|X)P(X) + P(A|Y)P(Y)$$
$$= 0.99 \times 0.6 + 0.95 \times 0.4 = 0.974$$

Generalized Bayes' Theorem

- Given *m* random variables, $\{X_1, X_2, \ldots, X_m\}$

$$P(Y \mid X_1, X_2, \ldots, X_m) = \frac{P(X_1, X_2, \ldots, X_m \mid Y) P(Y)}{P(X_1, X_2, \ldots, X_m)}$$

Chain Rule

- Given *m* random variables, $\{X_1, X_2, \dots, X_m\}$

  - $P(X_1, X_2, \dots, X_m)$

    $= P(X_1)\ P(X_2|X_1)\ \dots\ P(X_m|X_{m-1}, \dots, X_2, X_1)$

    $$= P(X_1) \prod_{i=2}^{m} P(X_i|X_{i-1}, \dots, X_2, X_1)$$

  - Proof:
    - $P(X_1, X_2) = P(X_1|X_2)P(X_2)$
    - $P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2, X_3)$
      $= P(X_1|X_2, X_3)P(X_2|X_3)P(X_3)$
    - …

Probability-based Learning Model

- Given a query **q** with *m* features
  - $\mathbf{q} = \{X_1, X_2, \dots, X_m\}$
- And there are *n* target levels
  - $\mathbf{T} = \{Y_1, Y_2, \dots, Y_n\}$
- We want to predict which target level **q** should belong to.
  - $M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} P(Y \mid X_1, X_2, \dots, X_m)$

Probability-based Learning Model

- Example:

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- Whether MENININGFITIS is true if
  q = {HEADACHE = true, FEVER = false, VOMITIMG = true}

Probability-based Learning Model

- According the generalized Bayes' theorem

  - $P(Y \mid X_1, X_2, \dots, X_m) = \frac{P(X_1, X_2, \dots, X_m \mid Y) P(Y)}{P(X_1, X_2, \dots, X_m)}$

- Let Y_1 be MENININGFITIS = true
  - $P(Y_1) = \frac{3}{10} = 0.3$

- Then Y_2 is MENININGFITIS = false
  - $P(Y_2) = 1.0 - P(Y_1) = 0.7$

- And the probability of **q** in the training data set
  - $P(\mathbf{q}) = P(X_1, X_2, \dots, X_m) = \frac{6}{10} = 0.6$

- So, $P(\mathbf{q}|Y) = P(X_1, X_2, \dots, X_m \mid Y) =?$

Probability-based Learning Model

- $P(X_1, X_2, \dots, X_m \mid Y) = \frac{P(Y, X_1, X_2, \dots, X_m)}{P(Y)}$

  - Or we can apply the chain rule

$$= \frac{P(Y)P(X_1|Y)\,P(X_2|X_1, Y)\ \dots\ P(X_m|X_{m-1}, \dots, X_2, X_1, Y)}{P(Y)}$$

$$= P(X_1|Y)\,P(X_2|X_1, Y)\ \dots\ P(X_m|X_{m-1}, \dots, X_2, X_1, Y)$$

Probability-based Learning Model

- **q** = {HEADACHE = true, FEVER = false, VOMITIMG = true}
- $P(\mathbf{q}|Y_1) = P(H, \bar{F}, V\ |Y_1)$

$$= P(H|Y_1) \times P(\bar{F}|H, Y_1) \times P(V\ |H, \bar{F}, Y_1)$$

$$= \frac{2}{3}$$

- $P(\mathbf{q}|Y_2) = P(H, \bar{F}, V\ |Y_1)$

$$= P(H|Y_2) \times P(\bar{F}|H, Y_2) \times P(V\ |H, \bar{F}, Y_2)$$

$$= \frac{4}{7}$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

Probability-based Learning Model

- Then,
  - $P(Y_1|\mathbf{q}) = \frac{P(\mathbf{q}|Y_1)P(Y_1)}{P(\mathbf{q})} = \frac{\frac{2}{3}\times\frac{3}{10}}{\frac{6}{10}} = \frac{1}{3} = 0.3333$
  - $P(Y_2|\mathbf{q}) = \frac{P(\mathbf{q}|Y_2)P(Y_2)}{P(\mathbf{q})} = \frac{\frac{4}{7}\times\frac{7}{10}}{\frac{6}{10}} = \frac{2}{3} = 0.6667$
- Therefore,
  - MENININGFITIS = false if
    q = {HEADACHE = true, FEVER = false, VOMITIMG = true}

Probability-based Learning Model

- What if
  q = {HEADACHE = true, FEVER = true, VOMITIMG = true}
  - No such training data!
  - Data insufficient ➔Model overfitting

Independence

- Two events, *X* and *Y*, are independent if knowledge of *Y* has no effect on the probability of *X*.

$$P(X|Y) = P(X)$$

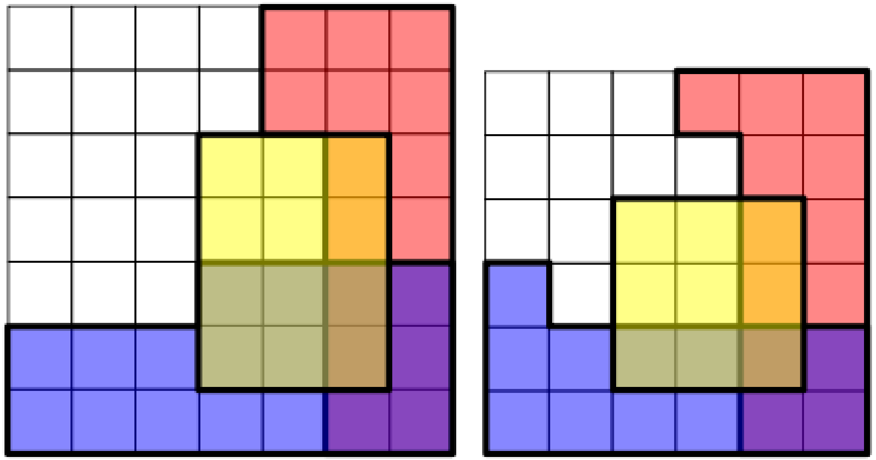
Then,

$$P(X,Y) = P(X|Y)P(Y) = P(X)P(Y)$$

Conditional Independence

- Two events *R* and *B* are conditionally independent given a third event *Y*,

$$P(R, B \mid Y) = P(R \mid Y) P(B \mid Y)$$



$$P(R, B \mid Y) = \frac{6}{12} \times \frac{4}{12} = \frac{2}{12}$$

$$P(R, B \mid Y) = \frac{3}{9} \times \frac{3}{9} = \frac{1}{9}$$

Conditional Independence

- Example 1.
  - Height (H) and vocabulary (V) are not independent
    - A taller kid could know more vocabulary than a shorter kid because the age of taller kid is larger then the shorter kid.
  - H and V are conditionally independent given a certain Age (A).
    - $P(H \mid A)$ and $P(V \mid A)$ are conditionally independent.
    - $P(H, V \mid A) = P(H \mid A)\, P(V \mid A)$
  - H and V are NOT conditionally independent given a gender (G).

Conditional Independence

- Example 2.
  - Lung cancer (*L*) and Smoking (*S*) are not independent
    - There are many people do smoking and have lung.
  - *L* and *S* are NOT conditionally independent given the condition of Regular Exercise (*E*).
    - $P(L = true \mid E = true)$ may still high if $P(S = true \mid E = true)$ is high.

  - *L* and *E* are not independent
    - Many people without lung cancer have a regular exercise.
  - *L* and *E* are conditionally independent given *S*.
    - $P(L \mid S)$ and $P(E \mid S)$ are conditionally independent.
    - $P(L, E \mid S) = P(L \mid S)\, P(E \mid S)$

Conditional Independence

- Given *m* random variables, $\{X_1, X_2, \dots, X_m\}$ and an event *Y*, if $X_1, X_2, \dots, \text{and } X_m$ are conditional independent under *Y*, then

$$P(X_1, X_2, \dots, X_m \mid Y)$$

$$= P(X_1 \mid Y) \times P(X_2 \mid Y) \times \cdots \times P(X_m \mid Y)$$

$$= \prod_{i=1}^{m} P(X_i \mid Y)$$

Conditional Independence

- Then,

$$P(Y \mid X_1, X_2, \dots, X_m)$$

$$= \frac{P(Y) \prod_{i=1}^{m} P(X_i \mid Y)}{P(X_1, X_2, \dots, X_m)}$$

Naive Bayes' Classifier

- Apply conditional independence to the learning model

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}}\, P(Y \mid X_1, X_2, \dots, X_m)$$

$$= \underset{Y \in \mathbf{T}}{\operatorname{argmax}} \frac{P(Y) \prod_{i=1}^{m} P(X_i \mid Y)}{P(X_1, X_2, \dots, X_m)}$$

Naive Bayes' Classifier

- However, the divider of $M(\mathbf{q})$, $P(X_1, X_2, \ldots, X_m)$, can be ignored in the maximum comparison.
- Therefore,

$$M(\mathbf{q}) = \underset{Y \in \mathbf{T}}{\operatorname{argmax}} P(Y) \prod_{i=1}^{m} P(X_i \mid Y)$$

Naive Bayes' Classifier

- What if
  q = {HEADACHE = true, FEVER = true, VOMITIMG = true}

- $P(\mathbf{q}|Y_1) = P(H, F, V\,|Y_1)$

$$= P(H|Y_1) \times P(F|Y_1) \times P(V\,|Y_1)$$

$$= \frac{2}{3} \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{27} = 0.1481$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- $P(\mathbf{q}|Y_2) = P(H, F, V\,|Y_2)$

$$= P(H|Y_1) \times P(F|Y_1) \times P(V\,|Y_1)$$

$$= \frac{5}{7} \times \frac{3}{7} \times \frac{4}{7} = \frac{60}{343} = 0.1749$$

21

Naive Bayes' Classifier

- Then,
  - $P(Y_1|\mathbf{q})P(Y_1) = \frac{4}{27} \times \frac{3}{10} = 0.0444$

  - $P(Y_2|\mathbf{q})P(Y_2) = \frac{60}{343} \times \frac{7}{10} = 0.1224$

- Therefore,
  - MENININGFITIS = false if
    q = {HEADACHE = true, FEVER = true, VOMITIMG = true}

Naive Bayes' Classifier

- An example of a loan application fraud detection

ID	Credit History	Guarantor/ CoApplicant	Accomodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Naive Bayes' Classifier

- Query *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = none*
  - *ACCOMODATION (ACC) = rent*

Naive Bayes' Classifier

- For *FR* = true
  - $P(fr) = \frac{6}{20} = 0.3$
  - $P(CH = paid \mid fr) = \frac{1}{6}$
  - $P(GC = none \mid fr) = \frac{5}{6}$
  - $P(ACC = rent \mid fr) = \frac{2}{6}$
  - $\frac{6}{20} \times \frac{1}{6} \times \frac{5}{6} \times \frac{2}{6} = 0.0139$
- For *FR* = false
  - $P(\overline{fr}) = \frac{14}{20} = 0.7$
  - $P(CH = paid \mid \overline{fr}) = \frac{4}{14}$
  - $P(GC = none \mid \overline{fr}) = \frac{12}{14}$
  - $P(ACC = rent \mid \overline{fr}) = \frac{2}{14}$
  - $\frac{14}{20} \times \frac{4}{14} \times \frac{12}{14} \times \frac{2}{14} = \mathbf{0.0245}$

Naive Bayes' Classifier

- How about that *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
-

Naive Bayes' Classifier

- For *FR* = true
  - $P(fr) = \frac{6}{20} = 0.3$
  - $P(CH = paid \mid fr) = \frac{1}{6}$
  - $P(GC = guarator \mid fr) = \frac{5}{6}$
  - $P(ACC = free \mid fr) = \mathbf{\frac{0}{6}}$
- For *FR* = false
  - $P(\overline{fr}) = \frac{14}{20} = 0.7$
  - $P(CH = paid \mid \overline{fr}) = \frac{4}{14}$
  - $P(GC = guarator \mid \overline{fr}) = \mathbf{\frac{0}{14}}$
  - $P(ACC = free \mid \overline{fr}) = \frac{1}{14}$

Naive Bayes' Classifier

- Smoothing
  - To take some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.
- There are several different ways to smooth probabilities.
  - Average smoothing
  - Gaussian smoothing
  - **Laplacian smoothing** is commonly used to smooth categorical data.
  - Given a constant *k* and a random variable *X* with *m* events,

$$P(X = x \mid y) = \frac{N(X=x \mid y)+k}{N(y)+km},$$

  - where $N(X=x|y)$ is number of sample under event $X=x|y$ and $N(y)$ is number of sample under event *y*.

Naive Bayes' Classifier

- Let *k* = 3
- For *ACC* = free and *FR* = true
  - The number of types of *ACC* (*m*) is 3 (own, rent, and free)
  - $P(ACC = free|\,fr)$

$$= \frac{N(ACC = free|\,fr) + 3}{N(fr) + 3 \times 3} = \frac{0 + 3}{6 + 9}$$
$$= 0.2$$

- For *GC* = guarantor and *FR* = false
  - The number of types of *GC* (*m*) is 3 (none, guarantor, and coapplicant)
  - $P(GC = guarator|\overline{fr})$

$$= \frac{N(GC = guarator|\overline{fr}) + 3}{N\left(\overline{fr}\right) + 3 \times 3} = \frac{0 + 3}{14 + 9}$$
$$= 0.1304$$

Naive Bayes' Classifier

- Therefor, after applying Laplace smoothing

$$
\begin{array}{rcl@{\qquad}rcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = none|fr) & = & 0.2222 & P(CH = none|\neg fr) & = & 0.1154 \\
P(CH = paid|fr) & = & 0.2222 & P(CH = paid|\neg fr) & = & 0.2692 \\
P(CH = current|fr) & = & 0.3333 & P(CH = current|\neg fr) & = & 0.2692 \\
P(CH = arrears|fr) & = & 0.2222 & P(CH = arrears|\neg fr) & = & 0.3462 \\
P(GC = none|fr) & = & 0.5333 & P(GC = none|\neg fr) & = & 0.6522 \\
P(GC = guarantor|fr) & = & 0.2667 & P(GC = guarantor|\neg fr) & = & 0.1304 \\
P(GC = coapplicant|fr) & = & 0.2 & P(GC = coapplicant|\neg fr) & = & 0.2174 \\
P(ACC = own|fr) & = & 0.4667 & P(ACC = own|\neg fr) & = & 0.6087 \\
P(ACC = rent|fr) & = & 0.3333 & P(ACC = rent|\neg fr) & = & 0.2174 \\
P(ACC = Free|fr) & = & 0.2 & P(ACC = Free|\neg fr) & = & 0.1739
\end{array}
$$

Naive Bayes' Classifier

- How about that *FRAUDULENT (FR) = ?* if
  - *CREDIT HISTORY (CH) = paid*
  - *GUARANTOR/COAPPLICANT (GC) = guarantor*
  - *ACCOMODATION (ACC) = free*
- For *FR* = true
  - $P(fr) \times P(CH = paid \mid fr) \times P(GC = guarator \mid fr) \times P(ACC = free \mid fr)$
  - $= 0.3 \times 0.2222 \times 0.2667 \times 0.2 = \mathbf{0.016}$
- For *FR* = false
  - $P(fr) \times P(CH = paid \mid \overline{fr}) \times P(GC = guarator \mid \overline{fr}) \times P(ACC = free \mid \overline{fr})$
  - $= 0.7 \times 0.2692 \times 0.1304 \times 0.1739 = 0.0042$