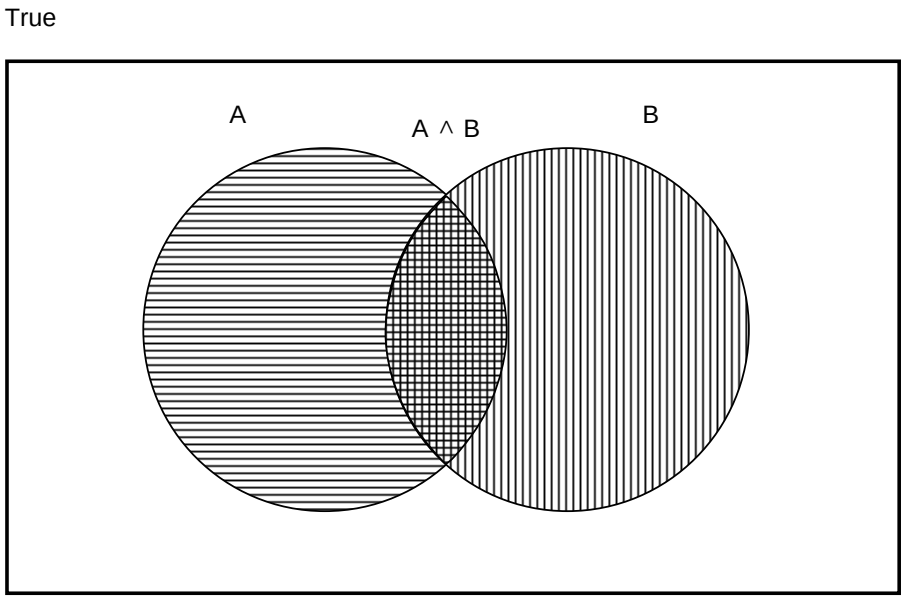
Naive Bayes Classifiers

COMP4211

Axioms for Probability

- All probabilities are between 0 and 1: $0 \leq P(A) \leq 1$
- Necessarily true propositions have probability 1: $P(True) = 1$
- Necessarily false propositions have probability 0: $P(False) = 0$
- The probability of a disjunction:
  $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



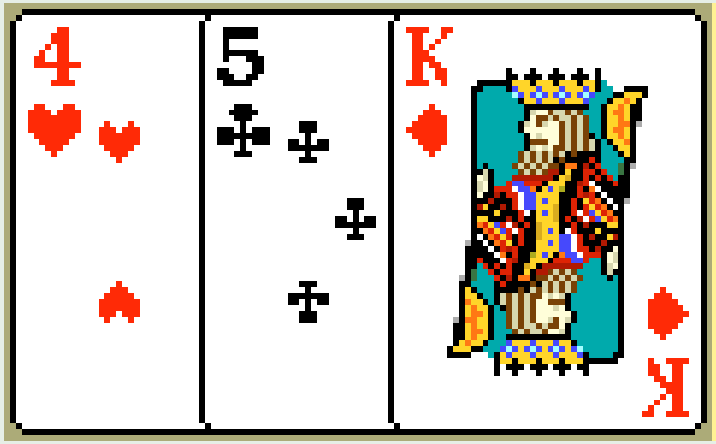
Mutually Exclusive Events

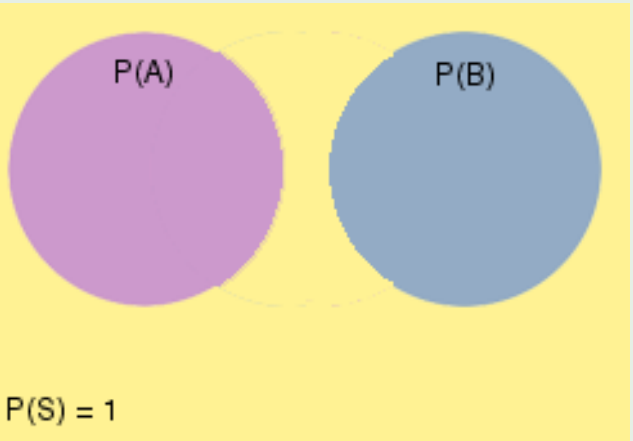
Two events are mutually exclusive if they cannot occur at the same time

Example

A single card is chosen at random from a standard deck of 52 playing cards

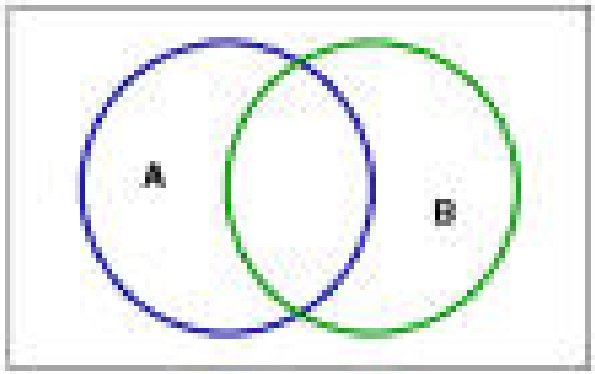
- E_1: the card chosen is a five, E_2: the card chosen is a king
- mutually exclusive?





Conditional Probability

- Let A and B be two events such that $P(A) > 0$
- $P(B|A)$: probability of B given that A has occurred



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

- probability that both A and B occur is equal to the probability that A occurs times the probability that B occurs given that A has occurred

Conditional Probability...

For any three events A_1, A_2, A_3:

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

If events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_n)P(B|A_n)$$

Bayes Rule

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

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{P(D|h)P(h)}{\sum_h P(D|h)P(h)}$$

- $P(h)$: prior probability of hypothesis h
  - initial probability that h holds, before observing the training data
- $P(h|D)$: posterior probability of h after observing the data D
- $P(D|h)$: likelihood of observing the data D given hypothesis h
- $P(D)$: probability that training data D will be observed

Example: Medical Diagnosis

Given:

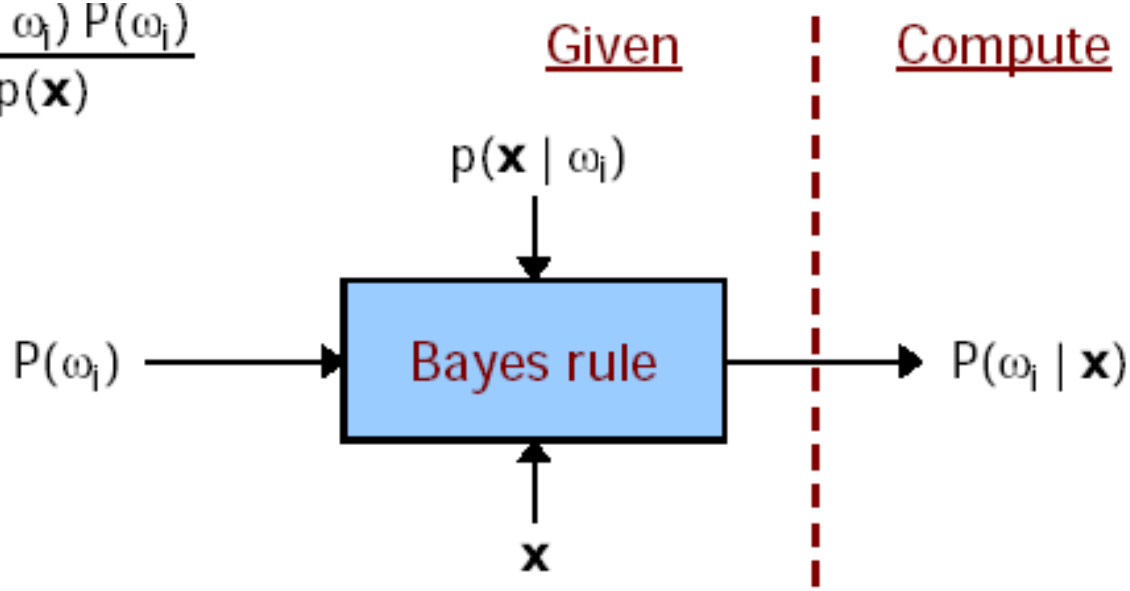
- $P(Cough|LungCancer) = 0.8$
- $P(LungCancer) = 0.005$
- $P(Cough) = 0.05$

Find: $P(LungCancer|Cough)$

$$
\begin{aligned}
& P(LungCancer|Cough) \\
= \quad & \frac{P(Cough|LungCancer)P(LungCancer)}{P(Cough)} \\
= \quad & \frac{0.8 \times 0.005}{0.05} = 0.08
\end{aligned}
$$

Bayes Rule...

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)\, P(\omega_i)}{p(\mathbf{x})}$$



- relates the prior probability (before observing D) and the posterior probability (after observing D)

Example: PlayTennis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Question

New instance: $\langle Sunny, Cool, High, Strong \rangle$, Play tennis?

Solution

- Instance x: attributes $\langle a_1, a_2 \ldots a_n \rangle$, and target v
- Most probable value of v:

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n) \\ &= \arg\max_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \\ &= \arg\max_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j) \end{aligned}$$

Question

How to estimate $P(v_j)$ and $P(a_1, a_2 \ldots a_n | v_j)$?

- n: number of training examples for which $v = v_j$
- N: total number of training examples

$$\hat{P}(v_j) \leftarrow n/N, \quad P(a_1, a_2 \ldots a_n | v_j) \leftarrow ?$$

Solution...

Example

100 binary attributes, and a binary class variables. How many entries in the joint probability table?

- ANS: 2^{101}

Problems

- too many to handle
- too many to estimate from data
- overfitting

The naive Bayes model reduces the number of model parameters by making independence assumption

Independence

- Two random variables X and Y are independent if

$$P(X|Y) = P(X), \text{ or } P(Y|X) = P(Y)$$

- Knowledge about X contains no information about Y
- Equivalently, $P(X, Y) = P(X)P(Y)$
- If n Boolean variables $(X_1, \ldots, X_n)$ are independent

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i)$$

  - the full joint can be specified by just n numbers

Example

- X: result of tossing a fair coin for the first time; Y: result of second tossing of the same coin
- X: result of US election; Y: your grades in this course

Question: Are these independent?

X: midterm exam grade; Y: final exam grade

Conditional Independence

- Absolute independence is a very strong requirement, seldom met
- Two random variables X and Y are conditionally independent given Z if

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

- Given Z, knowledge about Y contains no information about X
  - Y might contain some information about X
  - However all the information about X contained in Y are also contained in Z
- Equivalently, $P(Y|X,Z) = P(Y|Z)$ (why?)

$$\begin{aligned} P(Y|X,Z) &= P(X|Y,Z)P(Y|Z)/P(X|Z) \\ &= P(X|Z)P(Y|Z)/P(X|Z) = P(Y|Z) \end{aligned}$$

- Equivalently, $P(X,Y|Z) = P(X|Z)P(Y|Z)$ (why?)

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

Example

- There is a bag of 100 coins. 10 coins were made by a malfunctioning machine and are biased toward head. Tossing such a coin results in head 80% of the time. The other coins are fair.
- Randomly draw a coin from the bag and toss it a few time
- X_i : result of the ith tossing; Y: whether the coin is produced by the malfunctioning machine

Question

Are X_i's independent of each other?

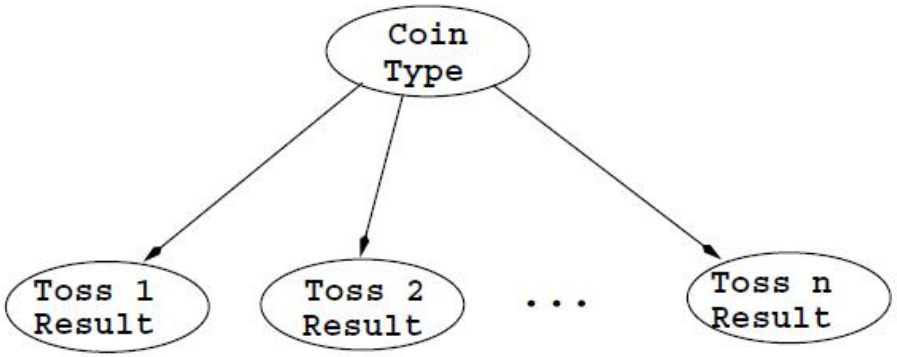
- If I get 9 heads in first 10 tosses, then the coin is probably a biased coin. Hence the next tossing will be more likely to result in a head than a tail.

Example...

Question

Are X_i's conditionally independent given Y?

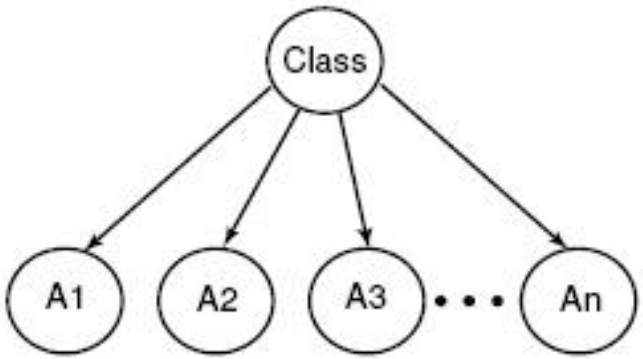
- If the coin is not biased, the probability of getting a head in one toss is $1/2$ regardless of the results of other tosses
- If the coin is biased, the probability of getting a head in one toss is 80% regardless of the results of other tosses
- If I already knew whether the coin is biased or not, learning the value of X_i does not give me additional information about X_j



Naive Bayes Classifier

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

- Naive Bayes assumption: Attributes are conditionally independent of each other given the class variable



$$P(a_1, a_2 \ldots a_n | v_j) = \prod_{i=1}^{n} P(a_i | v_j)$$

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i | v_j)$$

Learning the Naive Bayes Classifier

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i|v_j)$$

Learning amounts to estimating from data: $P(v_j)$'s, $P(a_1|v_j)$, ..., $P(a_n|v_j)$

Question

How to estimate $P(v_j)$?

- Straightforward

Question

How to estimate $P(a_i|v_j)$?

- n: number of training examples for which $v = v_j$
- n_c: number of examples for which $v = v_j$ and $a = a_i$

$$\hat{P}(a_i|v_j) \leftarrow n_c/n$$

Naive Bayes Algorithm

Naive_Bayes_Learn(*examples*)

begin
  for *each target value* v_j **do**
    $\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$;
    for *each attribute value* a_i *of each attribute* a **do**
      $\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$;
    end
  end
end

Classify_New_Instance(x)

begin

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

end

Example: PlayTennis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P(PlayTennis = y) = 9/14 \quad P(PlayTennis = n) = 5/14$$
$$P(Outlook = sunny|y) = 2/9 \quad P(Outlook = sunny|n) = 3/5$$
$$P(Outlook = overcast|y) = 4/9 \quad P(Outlook = overcast|n) = 0/5$$
$$P(Outlook = rain|y) = 3/9 \quad P(Outlook = rain|n) = 2/5$$
$$P(Temp = hot|y) = 2/9 \quad P(Temp = hot|PlayTennis = n) = 2/5$$
$$P(Temp = mild|y) = 4/9 \quad P(Temp = mild|n) = 2/5$$
$$P(Temp = cool|y) = 3/9 \quad P(Temp = cool|n) = 1/5$$
$$P(Humidity = high|y) = 3/9 \quad P(Humidity = normal|n) = 1/5$$
$$P(Humidity = normal|y) = 6/9 \quad P(Humidity = high|n) = 4/5$$
$$P(Wind = strong|y) = 3/9 \quad P(Wind = strong|n) = 3/5$$
$$P(Wind = weak|y) = 6/9 \quad P(Wind = weak|n) = 2/5$$

New instance : $\langle sunny, cool, high, strong \rangle$

$$P(y)P(sunny|y)P(cool|y)P(high|y)P(strong|y) = .005$$
$$P(n)P(sunny|n)P(cool|n)P(high|n)P(strong|n) = .021$$
$$\rightarrow v_{NB} = n$$

Zero Counts

Question

What if none of the training instances with target v_j have attribute a_i?

- $\hat{P}(a_i|v_j) = 0 \rightarrow \hat{P}(v_j)\prod_i \hat{P}(a_i|v_j) = 0$
- no chance to be classified as v_j, even if all other attribute values suggest v_j

Laplace correction: add a virtual count of 1 to each attribute value

- n: number of training examples for which $v = v_j$,
- n_c: number of examples for which $v = v_j$ and $a = a_i$
- v can take values of $v_1, v_2, \ldots, v_C$
- instead of

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c}{n}$$

- use

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + 1}{n + |a|}$$

Continuous Attributes

Discretize them

- typically, simply discretize into equal-length intervals (say, 10)

Conditional Independence Assumption

- often violated
- but it works surprisingly well anyway!
- don't need estimated posterior $\hat{P}(v_j|x)$ to be correct
- need only that

$$\arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = \arg\max_{v_j \in V} P(v_j) P(a_1 \ldots, a_n|v_j)$$

Advantages of Naive Bayes Classifier

- competitive performance
- fast
  - on training, requires only a single pass over the training set
  - on testing, also fast
- simple to update upon additions or deletions of training examples
  - easy to maintain