Outline

• Uncertainty
• Probability
• Syntax and Semantics
• Inference
• Independence and Bayes' Rule
Syntax

• Basic element: random variable
• Similar to propositional logic: possible worlds defined by assignment of values to random variables.
• Boolean random variables
  
  e.g., *Cavity* (do I have a cavity?) <true, false>

• Discrete random variables
• e.g., *Weather* is one of <sunny,rainy,cloudy,snow>
• Domain values must be exhaustive and mutually exclusive
• Elementary proposition constructed by assignment of a value to a random variable: e.g., *Weather* = sunny, *Cavity* = false
• (abbreviated as ¬*cavity*)
Syntax

• **Atomic event**: A complete specification of the state of the world about which the agent is uncertain

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  \[
  \begin{align*}
  Cavity &= \text{false} \land Toothache = \text{false} \\
  Cavity &= \text{false} \land Toothache = \text{true} \\
  Cavity &= \text{true} \land Toothache = \text{false} \\
  Cavity &= \text{true} \land Toothache = \text{true}
  \end{align*}
  \]

  • Atomic events are mutually exclusive and exhaustive
Axioms of probability

- For any propositions $A, B$
  
  - $0 \leq P(A) \leq 1$
  - $P(\text{true}) = 1$ and $P(\text{false}) = 0$
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Prior probability

• Prior or unconditional probabilities of propositions
e.g., \( P(Cavity = \text{true}) = 0.1 \) and \( P(Weather = \text{sunny}) = 0.72 \) correspond to belief prior to arrival of any (new) evidence

• Probability distribution gives values for all possible assignments:
\[
P(Weather) = <0.72,0.1,0.08,0.1> \text{ (normalized, i.e., sums to 1)}
\]

• Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables
\[
P(Weather,Cavity) = \begin{bmatrix}
0.144 & 0.02 & 0.016 & 0.02 \\
0.576 & 0.08 & 0.064 & 0.08
\end{bmatrix}
\]
### Joint Distribution


<table>
<thead>
<tr>
<th>Weather =</th>
<th>sunny</th>
<th>rainy</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td>0.144</td>
<td>0.02</td>
<td>0.016</td>
<td>0.02</td>
</tr>
<tr>
<td>Cavity = false</td>
<td>0.576</td>
<td>0.08</td>
<td>0.064</td>
<td>0.08</td>
</tr>
</tbody>
</table>

- Every question about the domain can be answered from joint probability distribution
Conditional probability

- **Conditional or posterior probabilities**
  e.g., $P(cavity \mid toothache) = 0.8$
  i.e., given that toothache is all I know
- (Notation for conditional distributions:
  $P(Cavity \mid Toothache) = 2$-element vector of 2-element vectors)
- If we know more, e.g., cavity is also given, then we have
  $P(cavity \mid toothache, cavity) = 1$
- New evidence may be irrelevant, allowing simplification, e.g.,
  $P(cavity \mid toothache, sunny) = P(cavity \mid toothache) = 0.8$
Conditional probability

- Definition of conditional probability:
  \[ P(a \mid b) = \frac{P(a \land b)}{P(b)} \]

- **Product rule** gives an alternative formulation:
  \[ P(a \land b) = P(a \mid b)P(b) = P(b \mid a)P(a) \]

- A general version holds for whole distributions, e.g.,
  \[ P(Weather, Cavity) = P(Weather \mid Cavity)P(Cavity) \]

- (View as a set of 4 × 2 equations, not matrix multiplication)

- This is analogous to logical reasoning, where logical agent cannot simultaneously believe A, B and \( \sim(A \text{ and } B) \)

- Where do probabilities come from? frequentist, objectivist, subjectivist (Bayesian)
Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.072</td>
<td>.008</td>
</tr>
<tr>
<td>cavity</td>
<td>.108</td>
<td>.016</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.072</td>
<td>.064</td>
</tr>
</tbody>
</table>

• For any proposition $\varphi$, sum the atomic events where it is true: $P(\varphi) = \sum_{\omega: \omega \models \varphi} P(\omega)$
Inference by enumeration

• Start with the joint probability distribution:

\[
P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
\]

For any proposition \( \varphi \), sum the atomic events where it is true: 
\[
P(\varphi) = \sum_{\omega: \omega \models \varphi} P(\omega)
\]

• \( P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2 \)
Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>0.016</td>
<td>0.064</td>
</tr>
<tr>
<td>cavity</td>
<td>0.072</td>
<td>0.008</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

• For any proposition \( \phi \), sum the atomic events where it is true:
  \[
P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)
  \]

• \( P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2 \)
• \( P(\text{toothache} \lor \text{cavity}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28 \)

• Process of summing out – marginalization – sum out all possible values of the other variables

\[
P(Y) = \sum_{z \in Z} P(Y, z) \quad P(\text{Cavity}) = \sum_{z \in \{\text{Catch,Toothache}\}} P(\text{Cavity}, z)
\]
Inference by enumeration

• Start with the joint probability distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>.108</td>
<td>.072</td>
</tr>
<tr>
<td>¬ catch</td>
<td>.012</td>
<td>.008</td>
</tr>
<tr>
<td>cavity</td>
<td>.016</td>
<td>.144</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>.064</td>
<td>.576</td>
</tr>
</tbody>
</table>


• Can also compute conditional probabilities:

\[
P(\neg \text{cavity} \mid \text{toothache}) = \frac{P(\neg \text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{(0.016 + 0.064)}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]
Normalization

- Denominator can be viewed as a normalization constant $\alpha$

$$P(Cavity \mid toothache) = \alpha \cdot P(Cavity, toothache)$$
$$= \alpha \cdot [P(Cavity, toothache, catch) + P(Cavity, toothache, \neg catch)]$$
$$= \alpha \cdot [<0.108, 0.016> + <0.012, 0.064>]$$
$$= \alpha \cdot <0.12, 0.08> = <0.6, 0.4>$$

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables
Inference by enumeration, contd.

- Typically, we are interested in the posterior joint distribution of the query variables $X$ given specific values $e$ for the evidence variables $E$
- Let the hidden variables be $Y = X - E$
- Then the required summation of joint entries is done by summing out the hidden variables:

\[
P(X \mid E = e) = \alpha P(X,E = e) = \alpha \Sigma_h P(X,E = e, Y = y)
\]

- The terms in the summation are joint entries because $X$, $E$ and $Y$ together exhaust the set of random variables

Obvious problems:
1. Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries?
Independence

• $A$ and $B$ are independent iff
  \[ P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A) P(B) \]

\[ P(\text{Toothache, Catch, Cavity, Weather}) = P(\text{Toothache, Catch, Cavity}) P(\text{Weather}) \]

• 32 entries reduced to 12; for $n$ independent biased coins, $O(2^n) \rightarrow O(n)$

• Absolute independence powerful but rare

• Dentistry is a large field with hundreds of variables, none of which are independent.
Bayes' Rule

- Product rule \( P(a \land b) = P(a \mid b)P(b) = P(b \mid a)P(a) \)

\( \Rightarrow \) Bayes' rule:

\[
P(a \mid b) = \frac{P(b \mid a)P(a)}{P(b)}
\]

- or in distribution form

\[
P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}
\]

- Useful for assessing diagnostic probability from causal probability:

\[
P(\text{Cause} \mid \text{Effect}) = P(\text{Effect} \mid \text{Cause}) \cdot P(\text{Cause}) / P(\text{Effect})
\]

Note: posterior probability of meningitis still very small!
Bayes' Rule

- Baye's rule
  \[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \]

- More general version conditionalized on some evidence
  \[ P(Y \mid X, e) = \frac{P(X \mid Y, e)P(Y \mid e)}{P(X \mid e)} \]

- E.g., let \( M \) be meningitis, \( S \) be stiff neck:
  \[ P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.8 \times 0.001}{0.1} = 0.008 \]

- Normalization same for \( m \) and \( \sim m \)
  \[ P(Y \mid X) = \alpha P(X \mid Y)P(Y) \]
Bayes' Rule and combining evidence

\[ P(Cavity \mid toothache \land catch) = \alpha P(toothache \land catch \mid Cavity) P(Cavity) \]

We can assume independence in the presence of Cavity

\[ = \alpha P(toothache \mid Cavity) P(catch \mid Cavity) P(Cavity) \]

Given Cavity, toothache and catch are independent

In general **Conditional Independence**

\[ P(X,Y \mid Z) = P(X \mid Z)P(Y \mid Z) \]
Conditional independence

- $P(\text{Toothache, Cavity, Catch})$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  \[(1) P(\text{catch} | \text{toothache, cavity}) = P(\text{catch} | \text{cavity})\]
- The same independence holds if I haven't got a cavity:
  \[(2) P(\text{catch} | \text{toothache, } \neg \text{cavity}) = P(\text{catch} | \neg \text{cavity})\]
- $\text{Catch}$ is conditionally independent of $\text{Toothache}$ given $\text{Cavity}$:
  \[P(\text{Catch} | \text{Toothache, Cavity}) = P(\text{Catch} | \text{Cavity})\]
Conditional independence contd.

- Write out full joint distribution, given the conditional independence assumption

\[
P(\text{Toothache, Catch, Cavity})
= P(\text{Toothache Catch } | \text{ Cavity}) \cdot P(\text{Cavity})
\]

\[
= P(\text{Toothache } | \text{ Cavity}) \cdot P(\text{Catch } | \text{ Cavity}) \cdot P(\text{Cavity})
\]

\[= P(\text{Toothache } | \text{ Cavity}) \cdot P(\text{Catch } | \text{ Cavity}) \cdot P(\text{Cavity})
\]

I.e., \(2 + 2 + 1 = 5\) independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in \(n\) to linear in \(n\).
Naïve Bayes

- **Naïve Bayes** model (the effects are independent given the cause)

\[ P(Cause, \text{Effect}_1, \ldots, \text{Effect}_n) = P(Cause) \prod_i P(\text{Effect}_i | Cause) \]

- This simplifying assumption, often works well

- Example SPAM classification
Example: Spam Filter

- Input: email
- Output: spam/ham

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
Example: Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
Example: Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails
- Features: The attributes used to make the ham / spam decision

---

Dear Sir.
First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.
99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
**Example: Spam Filter**

- **Input:** email
- **Output:** spam/ham
- **Setup:**
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails

- **Features:** The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - …

---

**Incorrect Examples:**

- Dear Sir.
  - First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. …

- TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.
  - 99 MILLION EMAIL ADDRESSES FOR ONLY $99

- Ok, I know this is blatantly OT but I’m beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
Probabilistic Classification

- **MAP classification rule**
  - **MAP**: Maximum A Posterior
  - Assign $x$ to $c^*$ if
    \[ P(C = c^* \mid X = x) > P(C = c \mid X = x) \quad c \neq c^*, \ c = c_1, \ldots, c_L \]

- **Generative classification with the MAP rule**
  - Apply Bayesian rule to convert them into posterior probabilities
    \[ P(C = c_i \mid X = x) = \frac{P(X = x \mid C = c_i) P(C = c_i)}{P(X = x)} \]
    \[ \propto P(X = x \mid C = c_i) P(C = c_i) \]
    for $i = 1, 2, \ldots, L$
  - Then apply the MAP rule
Naïve Bayes

- Bayes classification
  \[ P(C \mid X) \propto P(X \mid C)P(C) = P(X_1, \ldots, X_n \mid C)P(C) \]
  Difficulty: learning the joint probability \( P(X_1, \ldots, X_n \mid C) \)

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent
  \[
  P(X_1, X_2, \ldots, X_n \mid C) = P(X_1 \mid X_2, \ldots, X_n, C)P(X_2, \ldots, X_n \mid C)
  = P(X_1 \mid C)P(X_2, \ldots, X_n \mid C)
  = P(X_1 \mid C)P(X_2 \mid C) \cdots P(X_n \mid C)
  \]
  - MAP classification rule: for \( x = (x_1, x_2, \ldots, x_n) \)
  \[
  [P(x_1 \mid c^*) \cdots P(x_n \mid c^*)]P(c^*) > [P(x_1 \mid c) \cdots P(x_n \mid c)]P(c), \quad c \neq c^*, c = c_1, \ldots, c_L
  \]
Naïve Bayes

- **Naïve Bayes Algorithm** (for discrete input attributes)
  - **Learning Phase**: Given a training set $S$,
    For each target value of $c_i$ ($c_i = c_1, \cdots, c_L$)
    \[ \hat{P}(C = c_i) \leftarrow \text{estimate } P(C = c_i) \text{ with examples in } S; \]
    For every attribute value $x_{jk}$ of each attribute $X_j$ ($j = 1, \cdots, n; k = 1, \cdots, N_j$)
    \[ \hat{P}(X_j = x_{jk} \mid C = c_i) \leftarrow \text{estimate } P(X_j = x_{jk} \mid C = c_i) \text{ with examples in } S; \]
    Output: conditional probability tables; for $X_j, N_j \times L$ elements
  - **Test Phase**: Given an unknown instance $X' = (a'_1, \cdots, a'_n)$,
    Look up tables to assign the label $c^*$ to $X'$ if
    \[ [\hat{P}(a'_1 \mid c^*) \cdots \hat{P}(a'_n \mid c^*)] \hat{P}(c^*) > [\hat{P}(a'_1 \mid c) \cdots \hat{P}(a'_n \mid c)] \hat{P}(c), \quad c \neq c^*, \ c = c_1, \cdots, c_L \]
Example

Example: Play Tennis

PlayTennis: training examples

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>PlayTennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>
Example

- **Learning Phase**

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>2/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Overcast</td>
<td>4/9</td>
<td>0/5</td>
</tr>
<tr>
<td>Rain</td>
<td>3/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temperature</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot</td>
<td>2/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Mild</td>
<td>4/9</td>
<td>2/5</td>
</tr>
<tr>
<td>Cool</td>
<td>3/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Humidity</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>3/9</td>
<td>4/5</td>
</tr>
<tr>
<td>Normal</td>
<td>6/9</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wind</th>
<th>Play=Yes</th>
<th>Play=No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>3/9</td>
<td>3/5</td>
</tr>
<tr>
<td>Weak</td>
<td>6/9</td>
<td>2/5</td>
</tr>
</tbody>
</table>

\[
P(\text{Play=Yes}) = \frac{9}{14} \quad P(\text{Play=No}) = \frac{5}{14}\]
Example

- **Test Phase**
  - Given a new instance,
    \(x' = (\text{Outlook}=\text{Sunny}, \text{Temperature}=\text{Cool}, \text{Humidity}=\text{High}, \text{Wind}=\text{Strong})\)
  - Look up tables
    
    \[
    \begin{align*}
    P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{Yes}) &= 2/9 & P(\text{Outlook}=\text{Sunny} | \text{Play}=\text{No}) &= 3/5 \\
    P(\text{Temperature}=\text{Cool} | \text{Play}=\text{Yes}) &= 3/9 & P(\text{Temperature}=\text{Cool} | \text{Play}=\text{No}) &= 1/5 \\
    P(\text{Humidity}=\text{High} | \text{Play}=\text{Yes}) &= 3/9 & P(\text{Humidity}=\text{High} | \text{Play}=\text{No}) &= 4/5 \\
    P(\text{Wind}=\text{Strong} | \text{Play}=\text{Yes}) &= 3/9 & P(\text{Wind}=\text{Strong} | \text{Play}=\text{No}) &= 3/5 \\
    P(\text{Play}=\text{Yes}) &= 9/14 & P(\text{Play}=\text{No}) &= 5/14
    \end{align*}
    \]
  - MAP rule
    
    \[
    \begin{align*}
    P(\text{Yes} | x') &= [P(\text{Sunny} | \text{Yes})P(\text{Cool} | \text{Yes})P(\text{High} | \text{Yes})P(\text{Strong} | \text{Yes})]P(\text{Play}=\text{Yes}) = 0.0053 \\
    P(\text{No} | x') &= [P(\text{Sunny} | \text{No})P(\text{Cool} | \text{No})P(\text{High} | \text{No})P(\text{Strong} | \text{No})]P(\text{Play}=\text{No}) = 0.0206
    \end{align*}
    \]

Given the fact \(P(\text{Yes} | x') < P(\text{No} | x')\), we label \(x'\) to be “No”. 


Relevant Issues

- Violation of Independence Assumption
  - For many real world tasks, \( P(X_1, \cdots, X_n \mid C) \neq P(X_1 \mid C) \cdots P(X_n \mid C) \)
  - Nevertheless, naïve Bayes works surprisingly well anyway!

- Zero conditional probability Problem
  - If no example contains the attribute value \( X_j = a_{jk} \), \( \hat{P}(X_j = a_{jk} \mid C = c_i) = 0 \)
  - In this circumstance, \( \hat{P}(x_1 \mid c_i) \cdots \hat{P}(a_{jk} \mid c_i) \cdots \hat{P}(x_n \mid c_i) = 0 \) during test
  - For a remedy, conditional probabilities estimated with Laplacian smoothing
  - \[
  \hat{P}(X_j = a_{jk} \mid C = c_i) = \frac{n_c + mp}{n + m}
  \]
    - \( n_c \): number of training examples for which \( X_j = a_{jk} \) and \( C = c_i \)
    - \( n \): number of training examples for which \( C = c_i \)
    - \( p \): prior estimate (usually, \( p = 1/ t \) for \( t \) possible values of \( X_j \))
    - \( m \): weight to prior (number of "virtual" examples, \( m \geq 1 \))
Relevant Issues

• Continuous-valued Input Attributes
  – Numberless values for an attribute
  – Conditional probability modeled with the normal distribution
    \[ \hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi \sigma_j}} \exp \left( -\frac{(X_j - \mu_j)^2}{2\sigma_j^2} \right) \]
    \( \mu_j \): mean (average) of attribute values \( X_j \) of examples for which \( C = c_i \)
    \( \sigma_j \): standard deviation of attribute values \( X_j \) of examples for which \( C = c_i \)
  – Learning Phase: for \( X = (X_1, \ldots, X_n) \), \( C = c_1, \ldots, c_L \)
    Output: \( n \times L \) normal distributions and \( P(C = c_i) \) \( i = 1, \ldots, L \)
  – Test Phase: for \( X' = (X'_1, \ldots, X'_n) \)
    • Calculate conditional probabilities with all the normal distributions
    • Apply the MAP rule to make a decision
Conclusions

• Naïve Bayes based on the independence assumption
  – Training is very easy and fast; just requiring considering each attribute in each class separately
  – Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions

• A popular generative model
  – Performance competitive to most of state-of-the-art classifiers even in presence of violating independence assumption
  – Many successful applications, e.g., spam mail filtering
  – A good candidate of a base learner in ensemble learning
  – Apart from classification, naïve Bayes can do more...