Probabilistic Relationship between Attributes and Class

Ten middle-aged, divorced, male borrowers have defaulted on their loans, but would the 11th one default as well?

Bayes Theorem

\[ P(A | B) = \frac{P(B | A)P(A)}{P(B)} \]

Prior and posterior probabilities
- \( P(A) \) and \( P(A|B) \)
- \( P(B) \) and \( P(B|A) \)

Bayesian Classification

\[ P(C_i | X) = \frac{P(X | C_i)P(C_i)}{P(X)} \]

\( X \) is a given record with attribute values \( (x_1, x_2, \ldots, x_n) \), and \( C_i \) is a class
- \( P(C_i | X) \) is the probability of \( X \) belonging to class \( C_i \) given \( X \)'s attribute values
- We predict that \( X \) belong to \( C \) if \( P(C_i | X) > P(C_j | X) \) for \( j \neq i \)

Calculate \( P(C_i | X) \)

- \( P(X) \) does not need to be calculated
  - Why??
- \( P(C_i)?? \)
- \( P(X|C_i)?? \)

Sample Data

<table>
<thead>
<tr>
<th>TID</th>
<th>Home Owner</th>
<th>Marital Status</th>
<th>Annual Income</th>
<th>Defaulted Borrower</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
<tr>
<td>11</td>
<td>No</td>
<td>Single</td>
<td>120K</td>
<td>??</td>
</tr>
</tbody>
</table>
Naive Bayesian Classification

\[ X = (x_1, x_2, \ldots, x_n) \]

Assume the attribute values are conditionally independent of one another (the \textit{naive assumption})

\[
P(X | C) = \prod_{i=1}^{n} P(x_i | C)
= P(x_1 | C) \times P(x_2 | C) \times \cdots \times P(x_n | C)
\]

Attribute \( A_k \) is Categorical

\[ P(x_1 | C_i) \] is the fraction of number of records in \( C_i \) with value \( x_1 \) for attribute \( A_k \)

Attribute \( A_k \) is Continuous-valued

\[ \text{Assume } A_k \text{ follows a Gaussian distribution with a mean } \mu \text{ and standard deviation } \sigma \]

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \\
\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2}
\]

\( \sigma \): sample standard deviation

Naive Bayesian Classification Example ...

\[ P(\text{Default}=N | \text{HO}=N, MS=S, AI=120K) \]
\[ P(\text{Default}=Y | \text{HO}=N, MS=S, AI=120K) \]

\[ P(\text{Default}=N | \text{HO}=N, MS=S, AI=120K) \]
\[ P(\text{Default}=Y | \text{HO}=N, MS=S, AI=120K) \]

\[ \text{Annual Income, Default}=\text{No} \]
- \( \mu=110, \sigma=54.54 \)
- \( P(\text{AI}=120K | \text{No})=0.0072 \)

\[ \text{Annual Income, Default}=\text{Yes} \]
- \( \mu=90, \sigma=5 \)
- \( P(\text{AI}=120K | \text{Yes})=1.2 \times 10^{-9} \)
Avoid Zero $P(x_k | C_i)$

- A zero $P(x_k | C_i)$ would make the whole $P(X | C_i)$ zero
- To avoid this problem, add 1 to each count – assuming the training set is sufficiently large, the effect of adding one is negligible
- Example: $P(\text{Default}=\text{Y}|\text{HO}=\text{N}, \text{MS}=\text{M}, \text{AI}=120\text{K})$?

About Naive Bayesian Classification

- The most accurate classification if the conditional independence assumption holds
- In practice, some attributes may be correlated
  - E.g. education level and annual income

Bayesian Belief Network (BBN)

- A directed acyclic graph (dag) encoding the dependencies among a set of variables
- A conditional probability table (CPT) for each node given its immediate parent nodes

A BBN Example

![BBN Diagram](image1)

Construct a BBN

- Create the structure of the network
  - From domain knowledge
  - From training data
- Calculate the CPT for each node $X$
  - $P(X)$ if $X$ does not have any parent
  - $P(X|Y)$ if $X$ has one parent $Y$
  - $P(X|Y_1, Y_2, ..., Y_k)$ if $X$ has multiple parents $\{Y_1, Y_2, ..., Y_k\}$

BBN Construction Example

- Use the Vertebrate dataset to construct the following BBN
BBN Terminology

- If there is a directed arc from \( X \) to \( Y \)
  - \( X \) is a parent of \( Y \)
  - \( Y \) is a child of \( X \)
- If there is a directed path from \( X \) to \( Y \)
  - \( X \) is an ancestor of \( Y \)
  - \( Y \) is a descendent of \( X \)

Conditional Independence in BBN

- A node in a Bayesian network is conditionally independent of its non-descendants if its parents are known

\[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(x_i | Parents(A_i))
\]

Bayesian Classification Examples

- Output node – Heart Disease
- Testing data
  - ()
  - (BP=high)
  - (BP=high,D=Healthy,E=Yes)

Bayesian Classification Examples – 1

\[
P(HD=Yes) = \sum_{i=1}^{m} \sum_{j=1}^{n} P(HD=Yes | E=a_i, D=b_j) P(E=a_i, D=b_j)
\]

Bayesian Classification Examples – 2

\[
P(HD = Yes | BP = High)
= \frac{P(BP = High | HD = Yes) P(HD = Yes)}{P(BP = High)}
= \frac{P(BP = High | HD = Yes) P(HD = Yes)}{\sum_{i=1}^{m} P(BP = High | HD = a_i) P(HD = a_i)}
= 0.80
\]

Bayesian Classification Examples – 3

\[
P(HD = Yes | BP = High, D = Healthy, E = Yes)
= \frac{P(BP = High | HD = Yes, D = Healthy, E = Yes) P(HD = Yes | D = Healthy, E = Yes)}{P(BP = High | HD = Yes, D = Healthy, E = Yes)}
= \frac{P(BP = High | HD = Yes) P(HD = Yes | D = Healthy, E = Yes)}{\sum_{i=1}^{m} P(BP = High | HD = a_i) P(HD = a_i | D = Healthy, E = Yes)}
= 0.59
\]
**About BBN**
- Does not assume attribute independence
- Provides a way to encode domain knowledge
  - Robust to model overfitting
- Any node can be used as an output node

**Bayes Error Rate**
- If the relationship between attributes and class is probabilistic, it is impossible to be 100% correct.
- Bayes Error Rate – minimum achievable error rate for a given classifier

**Bayes Error Rate Example ...**
- Identify alligators and crocodiles based on their lengths $X$
- $P(X|\text{Crocodile})$ is Gaussian with $\mu=15$ and $\sigma=2$
- $P(X|\text{Alligator})$ is Gaussian with $\mu=12$ and $\sigma=2$

**...Bayes Error Rate Example...**
- Figure 5.11. Comparing the likelihood functions of a crocodile and an alligator.

**... Bayes Error Rate Example**
- \[ Error = \int P(X|\text{Crocodile})dX + \int P(X|\text{Alligator})dX \]
- \[ P(X = \hat{x}|\text{Crocodile}) = P(X = \hat{x}|\text{Alligator}) \]
- \[ \hat{x} = 13.5 \]

**Readings**
- Textbook 8.3 and 9.1.1