ESTIMATION OF DISTRIBUTION Algorithms IN Machine Learning

Pedro Larrañaga

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

EvoStar 2022, Madrid, April 22, 2022
Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
Huge spaces to be optimized

- **Structures. Combinatorial optimization**
  - Number of possible feature subsets, \( f(n) \), for a supervised classification problem with \( n \) predictor variables (Saeys et al. 2007): \( f(n) = 2^n \)
  - Number of possible partitional clustering assignments, \( S(N, K) \), of \( N \) objects into \( K \) groups (Sharp 1968):
    \[
    S(N, K) = \frac{1}{K!} \sum_{i=0}^{K} (-1)^{K-i} \binom{K}{i} i^N
    \]
  - Number of Bayesian networks structures, \( f(n) \), is super-exponential in the number of nodes, \( n \) (Robinson 1977):
    \[
    f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n - i), \text{ for } n > 2,
    \]
    which is initialized with \( f(0) = f(1) = 1 \)

- **Parameters. Continuous optimization**
  - Maximum likelihood estimation is not always achieved by means of a closed form
Machine Learning

Methods

### Preprocessing
- Dimensionality reduction (PCA, MDS, t-SNE, ..)
- Visualization
- Discretization

### Supervised classification
- **Non probabilistic classifiers**
  - $k$-nearest neighbors
  - Classification trees
  - Rule induction
  - Artificial neural networks
  - Support vector machines
- **Probabilistic classifiers**
  - Discriminant analysis
  - Logistic regression
  - Bayesian classifier
- **Meta classifiers**
  - Random Forest. Hybrid classifiers

### Clustering
- Hierarchical clustering
- Partitional clustering
- Probabilistic clustering

### Reinforcement learning

### Probabilistic graphical models
- Bayesian networks
- Markov networks

---

Bielza and Larrañaga 2021

© Pedro Larrañaga

EDAs in ML

EvoStar 2022
Optimization in Machine Learning

References

Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
\[ \max O(x) = \sum_{i=1}^{6} x_i \]

with \( x_i = 0, 1 \)
EDAs. A Toy Example

\[ \max O(\mathbf{x}) = \sum_{i=1}^{6} x_i \]

with \( x_i = 0, 1 \)

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( O(\mathbf{x}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
**EDAs. A Toy Example**

\[
\max O(x) = \sum_{i=1}^{6} x_i
\]

with \( x_i = 0, 1 \)

<table>
<thead>
<tr>
<th></th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( O(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
EDAs. A Toy Example

Learning the probability distribution from the selected individuals

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
EDAs. A Toy Example

Learning the probability distribution from the selected individuals

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ p(x) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6) \]
Learning the probability distribution from the selected individuals

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ p(x) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6) \]

\[ p(X_1 = 1) = \frac{7}{10} \]
EDAs. A Toy Example

Learning the probability distribution from the selected individuals

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$p(x) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6)$

$p(X_1 = 1) = \frac{7}{10}$  
$p(X_2 = 1) = \frac{7}{10}$  
$p(X_3 = 1) = \frac{6}{10}$  

$p(X_4 = 1) = \frac{6}{10}$  
$p(X_5 = 1) = \frac{8}{10}$  
$p(X_6 = 1) = \frac{7}{10}$
EDAs. A Toy Example

Learning the probability distribution from the selected individuals

\[ p(x) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6) \]

\[ p(X_1 = 1) = \frac{7}{10} \quad p(X_2 = 1) = \frac{7}{10} \quad p(X_3 = 1) = \frac{6}{10} \]

\[ p(X_4 = 1) = \frac{6}{10} \quad p(X_5 = 1) = \frac{8}{10} \quad p(X_6 = 1) = \frac{7}{10} \]
EDAs. A Toy Example

Obtaining the new population by sampling from the probability distribution

\[ p(X_1 = 1) = \frac{7}{10}; p(X_2 = 1) = \frac{7}{10}; p(X_3 = 1) = \frac{6}{10} \]

\[ p(X_4 = 1) = \frac{6}{10}; p(X_5 = 1) = \frac{8}{10}; p(X_6 = 1) = \frac{7}{10} \]

\[ p(x) = p(x_1, \ldots, x_6) = p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)p(x_6) \]

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability of ( X_i = 1 )</th>
<th>Condition</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.23</td>
<td>( p(X_1 = 1) = \frac{7}{10} ) &gt; 0.23</td>
<td>( \rightarrow 1 )</td>
<td></td>
</tr>
<tr>
<td>0.65</td>
<td>( p(X_2 = 1) = \frac{7}{10} ) &gt; 0.65</td>
<td>( \rightarrow 1 )</td>
<td></td>
</tr>
<tr>
<td>0.89</td>
<td>( p(X_3 = 1) = \frac{6}{10} ) &lt; 0.89</td>
<td>( \rightarrow 0 )</td>
<td></td>
</tr>
<tr>
<td>0.12</td>
<td>( p(X_4 = 1) = \frac{6}{10} ) &gt; 0.12</td>
<td>( \rightarrow 1 )</td>
<td></td>
</tr>
<tr>
<td>0.48</td>
<td>( p(X_5 = 1) = \frac{8}{10} ) &gt; 0.48</td>
<td>( \rightarrow 1 )</td>
<td></td>
</tr>
<tr>
<td>0.54</td>
<td>( p(X_6 = 1) = \frac{7}{10} ) &gt; 0.54</td>
<td>( \rightarrow 1 )</td>
<td></td>
</tr>
</tbody>
</table>
EDAs. A Toy Example

Obtaining the new population by sampling from the probability distribution

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
<th>$X_6$</th>
<th>$O(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
# Probabilistic Models in EDAs

## Univariate EDAs. Mühlenbein and Paaß (1996) (UMDA)
- **Probabilistic model**: \( p_l(x) = \prod_{i=1}^{n} p_l(x_i) \)
- **Structural learning**: not necessary

## Bivariate EDAs. De Bonet et al. (1997) (MIMIC)
- **Probabilistic model**:
  \[
p_l^\pi(x) = p_l(x_{i_1} | x_{i_2})p_l(x_{i_2} | x_{i_3}) \cdots p_l(x_{i_{n-1}} | x_{i_n})p_l(x_{i_n})
  \]
- **Structural learning**: best permutation

## Multivariate EDAs. Etxeberria and Larrañaga (1999) (EBNA); Pelikan et al. (1999) (BOA); Harik et al. (1999) (EcGA); Mühlenbein and Mahnig (1999) (LFDA)
- **Probabilistic model**: \( p_l(x) = \prod_{i=1}^{n} p_l(x_i | pa_i) \)
- **Structural learning**: directed acyclic graph
Estimation of Distribution Algorithms

Probabilistic Models in EDAs

Univariate EDAs. Mühlenbein and Paaß (1996) (UMDA)
- **Probabilistic model:** \( p_l(x) = \prod_{i=1}^{n} p_l(x_i) \)
- **Structural learning:** not necessary

Bivariate EDAs. De Bonet et al. (1997) (MIMIC)
- **Probabilistic model:**
  \[
  p_{\pi}^l(x) = p_l(x_{i_1} | x_{i_2})p_l(x_{i_2} | x_{i_3}) \cdots p_l(x_{i_{n-1}} | x_{i_n})p_l(x_{i_n})
  \]
- **Structural learning:** best permutation

Multivariate EDAs. Etxeberria and Larrañaga (1999) (EBNA); Pelikan et al. (1999) (BOA); Harik et al. (1999) (EcGA); Mühlenbein and Mahnig (1999) (LFDA)
- **Probabilistic model:** \( p_l(x) = \prod_{i=1}^{n} p_l(x_i | pa_i) \)
- **Structural learning:** directed acyclic graph

EDAs in continuous domains. Assuming Gaussianity: Larrañaga et al. (2000)
- **Univariate:** \( \text{UMDA}_c^G \)
- **Bivariate:** \( \text{MIMIC}_c^G \)
- **Multivariate:** \( \text{EMNA}_{global}^G, \text{EMNA}_{ee}^G, \text{EGNA}^G \)
Estimation of Distribution Algorithms

Graphical Representation of EDAs

Population of candidate solutions

Selected candidates

Bayesian network or Gaussian network

New candidate solutions

© Pedro Larrañaga
EDAs in ML
EvoStar 2022
Evolution of Bayesian Network Structures in a EBNA Search (Bengoetxea 2002)

generation 0

generation 8

generation 16

generation 24

generation 32

generation 37
A Node for the Objective Function (Miquélez et al. 2004)

Estimation of Bayesian classifiers optimization algorithms

- Initial population of candidate solutions with their fitness values
- Selected candidates with a discretization of the fitness
- Bayesian classifier learning
- Sampling and fitness evaluation of new candidates
EDAs for multiobjective optimization

A Node for Each Objective Function (Karshenas et al. 2014)
EDAs

Books

Estimation of Distribution Algorithms

Larrañaga and Lozano 2002

Hierarchical Bayesian Optimization Algorithm

Pelikan 2005

Towards an New Evolutionary Computation

Lozano et al. 2006

Special issues

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

2002

Evolutionary Computation

2005

IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION

2009
References I

References II

References III

References

Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
Bayesian Networks

DAG + CPTs

- **Conditional independence**: \( W \) and \( T \) are conditionally independent given \( Z \) \( \iff \) \( p(W|T, Z) = p(W|Z) \)
- **Directed acyclic graph** (DAG)
- **Conditional probability tables** (CPTs)

\[
p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | \text{Pa}(X_i))
\]

\[
p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N, S)p(P|S)
\]
Bayesian Networks

DAG + CPTs

- Conditional independence: W and T are conditionally independent given Z ⇔ \( p(W|T, Z) = p(W|Z) \)
- Directed acyclic graph (DAG)
- Conditional probability tables (CPTs)
- \( p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | Pa(X_i)) \)

Inference

- Exact: variable elimination, message passing
- Approximate: sequential simulation and MCMC

\[ p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N, S)p(P|S) \]

Bielza and Larrañaga 2021

© Pedro Larrañaga

EDAs in ML

EvoStar 2022
Bayesian Networks

Conditional Independence

\[ p(X_i) \]

\[ p(X_i | \text{Stroke}=yes) \]

\[ p(X_i | \text{Stroke}=yes, \text{Neural Atropy}=yes) \]

\[ p(X_i | \text{Stroke}=yes, \text{Neural Atropy}=yes, \text{Age}=young) \]
Learning Bayesian Networks from Data

Two tasks

- **Parameters** $p(X_i = x_i | Pa(X_i) = pa_i^j)$: MLE or Bayesian
- **Structure**: conditional independence tests or by optimizing a score

Scores

- **Penalized** likelihood: avoid structural overfitting
- **Bayesian**: $\arg \max_G p(G|D)$, with marginal likelihood prior

$p(G|D) \propto \widehat{p(D|G)} \ p(G)$, with

$p(D|G) = \int p(D|G, \theta) \ f(\theta|G) \ d\theta$

Scores

- **Penalized likelihood**: avoid structural overfitting
- **Bayesian**: $\arg \max_G p(G|D)$, with marginal likelihood prior

$p(G|D) \propto \widehat{p(D|G)} \ p(G)$, with

$p(D|G) = \int p(D|G, \theta) \ f(\theta|G) \ d\theta$

Scores

- **Penalized likelihood**: avoid structural overfitting
- **Bayesian**: $\arg \max_G p(G|D)$, with marginal likelihood prior

$p(G|D) \propto \widehat{p(D|G)} \ p(G)$, with

$p(D|G) = \int p(D|G, \theta) \ f(\theta|G) \ d\theta$
Bayesian Networks

References

Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
**The problem.** The way rows and columns are ordered in a table is a very sensitive issue that affects its readability. The optimal ordering of tables is equivalent to solving two travelling salesman problems, one for the $R$ rows and the other for the $C$ columns $\Rightarrow$ cardinality of the search space: $R! \cdot C!$

**Individual representation.**

(a) Discrete: $x = (x_1, \ldots, x_R, x_{R+1}, \ldots, x_{R+C})$, where $x_i = k$ means that the order of the original $i$th row is $k$, and $x_{R+j} = l$ means that the order for the $j$th column is $l$

(b) Continuous: the real vectors of values were transformed into permutations as the respective order in the continuous individual

**EDAs.** Univariate, bivariate and multivariate in both discrete and continuous domains
**The problem.** The way rows and columns are ordered in a table is a very sensitive issue that affects its readability. The optimal ordering of tables is equivalent to solving two travelling salesman problems, one for the $R$ rows and the other for the $C$ columns $\Rightarrow$ cardinality of the search space: $R! \cdot C!$

**Individual representation.**
- (a) Discrete: $x = (x_1, \ldots, x_R, x_{R+1}, \ldots, x_{R+C})$, where $x_i = k$ means that the order of the original $i$th row is $k$, and $x_{R+j} = l$ means that the order for the $j$th column is $l$
- (b) Continuous: the real vectors of values were transformed into permutations as the respective order in the continuous individual

**EDAs.** Univariate, bivariate and multivariate in both discrete and continuous domains

**References**

Wrapper Multidimensional Discretization (Flores et al. 2007)

EDAs in Preprocessing

Multidimensional discretization

- **The problem.** Optimal (maximizing the estimated accuracy) multidimensional discretization
- **Individual representation** depends on the discretization process (number of bins and cutpoints for each predictor variable)
  - An individual in the EDA represents a discretization policy that transforms the original dataset into a discretized one
- **EDAs.** UMDA$_c^G$

Original data set

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-17</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

Individuals of the EDA (multidimensional discretization)

<table>
<thead>
<tr>
<th>$d_1$</th>
<th>$t_1^1$</th>
<th>$t_1^2$</th>
<th>$t_1^3$</th>
<th>$t_1^4$</th>
<th>$t_1^5$</th>
<th>$t_1^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-12</td>
<td>-6</td>
<td>12</td>
<td>87.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-10</td>
<td>-3</td>
<td>9</td>
<td>71.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discretized data sets

<table>
<thead>
<tr>
<th>$X_1^{d_1}$</th>
<th>$X_2^{d_1}$</th>
<th>$X_3^{d_1}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X_1^{d_2}$</th>
<th>$X_2^{d_2}$</th>
<th>$X_3^{d_2}$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Accuracy 87.13% Accuracy 71.14%
**EDAs in Preprocessing**

**Wrapper Multidimensional Discretization**

(Flores et al. 2007)

<table>
<thead>
<tr>
<th>Original data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>$X_3$</td>
</tr>
<tr>
<td>$C$</td>
</tr>
<tr>
<td>0 -17</td>
</tr>
<tr>
<td>17 3</td>
</tr>
<tr>
<td>4 3</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>5 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Individuals of the EDA (multidimensional discretization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
</tr>
<tr>
<td>$d_2$</td>
</tr>
<tr>
<td>$t_1^1$</td>
</tr>
<tr>
<td>$t_1^2$</td>
</tr>
<tr>
<td>$t_1^3$</td>
</tr>
<tr>
<td>5 -12</td>
</tr>
<tr>
<td>4 -10</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>87.13</td>
</tr>
<tr>
<td>71.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discretized data sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1^{d_1}$</td>
</tr>
<tr>
<td>$X_2^{d_1}$</td>
</tr>
<tr>
<td>$X_3^{d_1}$</td>
</tr>
<tr>
<td>$C$</td>
</tr>
<tr>
<td>1 1 2 1</td>
</tr>
<tr>
<td>2 3 2 1</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>2 3 1 1</td>
</tr>
</tbody>
</table>

| $X_1^{d_2}$ |
| $X_2^{d_2}$ |
| $X_3^{d_2}$ |
| $C$ |
| 1 1 2 1 |
| 2 3 2 1 |
| ...    |
| 2 3 2 1 |

**Multidimensional discretization**

- **The problem.** Optimal (maximizing the estimated accuracy) multidimensional discretization
- **Individual representation** depends on the discretization process (number of bins and cutpoints for each predictor variable)
  - An individual in the EDA represents a discretization policy that transforms the original dataset into a discretized one
- **EDAs.** UMDA$_G^c$

**References**

Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
The problem. Optimal (maximizing the accuracy) subset of features for a given supervised classification paradigm

Individual representation. \( \mathbf{x} = (x_1, \ldots, x_n) \) where \( x_i = 1 \) if variable \( X_i \) is selected, and 0 otherwise

EDAs. EBNA for ID3 and naive Bayes
EDAs in Supervised Classification

Feature Subset Selection

References


$d(x, x^i) = \sum_{j=1}^{n} w_j \delta(x_j, x^i_j)$
\[ d(x, x^i) = \sum_{j=1}^{n} w_j \delta(x_j, x_j^i) \]

**Feature weighting**

- **The problem**: Search for the optimal (in terms of accuracy) feature weighting
- **Individual representation**: discrete (three possible values), or continuous
- **EDAs**: EBNA and EGNA

**References**

Classification Trees (Cagnini et al. 2017)

Optimal classification tree

- **The problem.** Optimal (maximizing the estimated accuracy) classification tree for continuous predictors
- **Individual representation.** Binary tree with a given maximal depth. Root node and internal nodes selected from probability distributions
- **EDAs.** UMDA
Classification Trees (Cagnini et al. 2017)

Optimal classification tree

- **The problem.** Optimal (maximizing the estimated accuracy) classification tree for continuous predictors
- **Individual representation.** Binary tree with a given maximal depth. Root node and internal nodes selected from probability distributions
- **EDAs.** UMDA

References

Rule Induction (Sierra et al. 2002)

\[
\{ \mathcal{R} : \begin{align*}
\text{IF} & \quad (X_{25} = 2 \; \text{AND} \; X_{56} = 3) \\
\text{OR} & \quad (X_{2} = 1 \; \text{AND} \; X_{5} \neq 1) \\
\text{THEN} \quad & \quad C = I
\end{align*} \} 
\]

Pittsburgh-like approach

- **The problem.** Optimal (maximizing the estimated accuracy) rule. Classifier system
- **Individual representation.** The antecedent of the rule consists on disjunction of simple antecedents, where a single rule dimension is given by \( n \), the number of predictor variables, allowing for each variable to take values that are equal to, different from, and any possible value
- **EDAs.** UMDA, EBNA
Rule Induction (Sierra et al. 2002)

\[
\{ \begin{array}{l}
R : \text{IF} \quad (X_{25} = 2 \text{ AND } X_{56} = 3) \quad \text{OR} \\
\quad \quad \quad (X_2 = 1 \text{ AND } X_5 \neq 1) \quad \text{THEN} \quad C = I
\end{array} \}
\]

Pittsburgh-like approach

- **The problem.** Optimal (maximizing the estimated accuracy) rule. Classifier system
- **Individual representation.** The antecedent of the rule consists on disjunction of simple antecedents, where a single rule dimension is given by \( n \), the number of predictor variables, allowing for each variable to take values that are equal to, different from, and any possible value.
- **EDAs.** UMDA, EBNA

References

Optimal weights

- **The problem.** Minimization of $E(w, w') = \frac{1}{N} \sum_{k=1}^{N} (c^k - \hat{c}^k)^2$. Alternative to the backpropagation algorithm, a gradient descent method.

- **Individual representation.** $k$-dimensional vectors of real numbers, where $k$ denotes the number of weights in the artificial neural network (a feed-forward multilayer perceptron).

- **EDAs.** PBIL: $p_{l+1}(x) = (1 - \alpha)p_l(x) + \alpha \frac{1}{N} \sum_{k=1}^{N} x_{k,M}^l$.
EDAs in Supervised Classification

Artificial Neural Networks

References


EDAs in Supervised Classification

Logistic Regression (Robles et al. 2008)

- The logistic regression model: \( p(C = 1|\mathbf{x}, \mathbf{\beta}) = \theta_x = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n}} \)

- The likelihood function is: \( L(\mathbf{\beta}|\mathbf{x}^1, \ldots, \mathbf{x}^N) = p(c^1, \ldots, c^N|\mathbf{x}, \theta_x) = \prod_{i=1}^N \theta_i^{c_i} (1 - \theta_i)^{1-c_i} \)

- The likelihood equations are:

\[
\begin{align*}
\frac{\partial \ln L(\mathbf{\beta})}{\partial \beta_0} &= \sum_{i=1}^N c_i - \sum_{i=1}^N \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} = 0 \\
\frac{\partial \ln L(\mathbf{\beta})}{\partial \beta_1} &= \sum_{i=1}^N c_i x_i - \sum_{i=1}^N x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} = 0 \\
&\vdots \\
\frac{\partial \ln L(\mathbf{\beta})}{\partial \beta_n} &= \sum_{i=1}^N c_i x_i - \sum_{i=1}^N x_i \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} = 0
\end{align*}
\]

- The (iterative) Newton-Raphson method: \( \mathbf{\beta}^{\text{new}} = \mathbf{\beta}^{\text{old}} - \left( \frac{\partial^2 \ln L(\mathbf{\beta})}{\partial \mathbf{\beta} \partial \mathbf{\beta}^T} \right)^{-1} \frac{\partial \ln L(\mathbf{\beta})}{\partial \mathbf{\beta}} \)

Pareto front for calibration and discrimination

- Each individual in the EDA is represented as a vector of real numbers with cardinality \( n + 1 \)
- Two UMDA\(_G\)s were developed, one for calibration (log-likelihood) and the other for discrimination (area under the ROC curve)
- The best individuals obtained with each of these UMDA\(_G\)s were evaluated in the other objective, thus obtaining an approximation to the Pareto front for the bi-objective problem
Logistic Regression

References


EDAs in Supervised Classification

Bayesian Classifiers (Robles et al. 2004)

\[ p(c|x_1,\ldots,x_n) \propto p(c,x_1,\ldots,x_n) \]

**Semi-naive Bayes**

- \( p(c|x_1,\ldots,x_4) \propto p(c)p(x_1,x_3|c)p(x_4|c) \)
- \( p(c|x_1,x_2,x_3,x_4) \propto p(c)p(x_1,x_2,x_3,x_4|c) \)

- **UMDA based approach.** Individuals will have \( n \) variables each one with an integer value in \( \{0, 1, 2, \ldots, n\} \) representing the (super)node each variable belongs to.
Bayesian Classifiers (Robles et al. 2004)

$p(c|x_1,\ldots,x_n) \propto p(c) p(x_1,\ldots,x_n|c)$

$= p(c) p(x_1|\text{pa}(c)) \prod_{i=1}^{n} p(x_i|\text{pa}(x_i))$

UMDA based approach. Individuals will have $n$ variables each one with an integer value in $\{0, 1, 2, \ldots, n\}$ representing the (super)node each variable belongs to

References

The base classifiers are classification trees. Each classifier $\phi_i$ is trained on data set $D_i$ of size $N$ sampled from $D$ which focuses more on the mistakes of the previous classifier $\phi_{i-1}$.

Voting weights given by AdaBoost as starting point for the UMDA$^G_c$.
EDAs in Supervised Classification

Boosting (Cagnini et al. 2018)

The base classifiers are classification trees. Each classifier $\phi_i$ is trained on data set $D_i$ of size $N$ sampled from $D$ which focuses more on the mistakes of the previous classifier $\phi_{i-1}$.

Voting weights given by AdaBoost as starting point for the UMDA$_C^G$.

References

Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
EDAs in Clustering

Hierarchical Clustering (Fan 2019)

Stochasticity in the merging operations in agglomerative hierarchical clustering

- **UMDA** promoting that the subsets with more instances have a greater probability of being joined as long as the value of the distance with the centroid linkage does not exceed a certain threshold.

- The constructed **dendrogram** does not necessarily have to be complete.
EDAs in Clustering

Hierarchical Clustering (Fan 2019)

Stochasticity in the merging operations in agglomerative hierarchical clustering

- UMDA promoting that the subsets with more instances have a greater probability of being joined as long as the value of the distance with the centroid linkage does not exceed a certain threshold
- The constructed dendrogram does not necessarily have to be complete

References

K-means. Forgy (1965)

K-means (hill-climbing strategy) with computations of the new centroids once all objects are assigned to their respective clusters.

EDAs with an object membership representation

- Each individual is a string of length $N$ (number of objects to clusters) where each position can take one value in $\{1, \ldots, K\}$, with $K$ denoting the number of clusters.
- The $i$th position of the string represents the cluster number to which object $x^i$ belongs.
- EDAs: MIMIC, EBNA$_{BIC}$
EDAs in Clustering

Partitional Clustering

References

The value of variable $C$ is unknown, it is estimated with the EM algorithm.

UMDA to search for the optimal structure of dependency between the variables. Each individual in the UMDA represents an upper triangular connectivity matrix $H$ with $\frac{n^2-n}{2}$ elements $h_{ij}$, such that:

$$h_{ij} = \begin{cases} 
  1 & \text{if } X_j \in Pa_i \\
  0 & \text{otherwise}
\end{cases}$$
Probabilistic Clustering (Peña et al. 2004)

The value of variable $C$ is unknown, it is estimated with the EM algorithm
UMDA to search for the optimal structure of dependency between the variables. Each individual in the UMDA represents an upper triangular connectivity matrix $H$ with $\frac{n^2-n}{2}$ elements $h_{ij}$, such that:

$$h_{ij} = \begin{cases} 1 & \text{if } X_j \in Pa_i \\ 0 & \text{otherwise} \end{cases}$$

References


Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
EDAs in Reinforcement Learning

Reinforcement Learning (Handa and Nishimura 2008)

Handa and Nishimura 2008
Reinforcement Learning (Handa and Nishimura 2008)

References

EDAs in Bayesian Networks

Learning from Data (Blanco et al. 2003)

Learning structure and parameters

<table>
<thead>
<tr>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>X₄</th>
<th>X₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>p(X₁=0) = 0.60</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(X₂=0</td>
</tr>
<tr>
<td>p(X₂=0</td>
</tr>
<tr>
<td>p(X₃=0</td>
</tr>
<tr>
<td>p(X₃=0</td>
</tr>
</tbody>
</table>

Parameter learning

| p(X₂=0|X₁=0,X₃=0) = 0.60  |
| p(X₂=0|X₁=0,X₃=1) = 0.20  |
| p(X₄=0|X₁=0,X₃=0) = 0.60  |
| p(X₄=0|X₁=0,X₃=1) = 0.20  |

| p(X₅=0|X₁=0) = 0.60  |
| p(X₅=0|X₁=0) = 0.60  |

Space of DAGs

- **Univariate EDAs: UMDA**
- Individual **representation**: connectivity matrix (Bayesian network structure)
- If no total ordering between the variables is assumed: simple **repair operator** (randomly delete cycles)
## References


Outline

1. Introduction
2. Estimation of Distribution Algorithms
3. Bayesian Networks
4. EDAs in Preprocessing
5. EDAs in Supervised Classification
6. EDAs in Clustering
7. EDAs in Reinforcement Learning
8. EDAs in Bayesian Networks
9. Conclusions and Further Topics
Conclusions and Further Topics

Estimation of Distribution Algorithms in Machine Learning

Methods

<table>
<thead>
<tr>
<th>Preprocessing</th>
<th>Clustering</th>
<th>Supervised classification</th>
<th>Reinforcement learning</th>
<th>Probabilistic graphical models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensionality reduction (PCA, MDS, t-SNE, ..)</td>
<td>Hierarchical clustering</td>
<td><strong>Non probabilistic classifiers</strong></td>
<td>Bayesian networks</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Partitional clustering</td>
<td><em>k</em>-nearest neighbors</td>
<td></td>
<td>Markov networks</td>
</tr>
<tr>
<td>Discretization</td>
<td>Probabilistic clustering</td>
<td>Classification trees</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rule induction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Artificial neural networks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Support vector machines</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Probabilistic classifiers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discriminant analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logistic regression</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian classifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Metaclassifiers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stacking</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cascading</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bagging</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Boosting</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Random Forest</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hybrid classifiers</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Multidimensional classification</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions and Further Topics

Conclusions

- Estimation of distribution algorithms competitive with the state of the art heuristics in machine learning
- Bayesian networks as a framework providing interpretability for machine learning and optimization

Further topics

- EDAs
  - Development of EDAs that incorporates advances in methodology for learning Bayesian networks from data
  - Continuous optimization: Semiparametric Bayesian networks

- EDAs in machine learning
  - Preprocessing: Parallel coordinates
  - Feature subset selection: Multivariate filtering approach
  - $k$-nearest neighbors: Prototype selection, distance election
  - Multilabel classification: Chain classifiers
  - Clustering: Divisive hierarchical clustering, $K$-medians, $K$-modes, fuzzy $C$-means, SOM, biclustering, clustering multi-view

- Bayesian networks
  - Mallows distribution for the best order in (a) structure learning in the space of ordering, (b) triangulation of the moral graph
  - Evidence explanation: Most relevant explanation, MAP-independence explanation, counterfactual reasoning
E STIMATION OF D ISTRIBUTION A LGORITHMS
I N
M ACHINE L EARNING

Pedro Larrañaga

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

EvoStar 2022, Madrid, April 22, 2022