Overview

- Fuzzy Sets
- Information Retrieval
- Machine Learning
- Statistics & Estimation Techniques
- Similarity Measures
- Decision Trees

Fuzzy Sets and Logic

Fuzzy Set: Set where the set membership function is a real valued function with output in the range [0,1].
- \( f(x) \): Probability \( x \) is in \( F \).
- \( 1-f(x) \): Probability \( x \) is not in \( F \).

Example
- \( T = \{ x \mid x \text{ is a person and } x \text{ is tall} \} \)
- Let \( f(x) \) be the probability that \( x \) is tall
- Here \( f \) is the membership function

DM: Prediction and classification are often fuzzy.

Fuzzy Sets

Crisp Sets

Fuzzy Sets

Classification/Prediction is Fuzzy

Information Retrieval

Information Retrieval (IR): retrieving desired information from textual data.
- Library Science
- Digital Libraries
- Web Search Engines
- Traditionally has been keyword based
- Sample query:
  - Find all documents about “data mining”.

DM: Similarity measures; Mine text or Web data.
Information Retrieval (cont’d)

**Similarity**: measure of how close a query is to a document.
- Documents which are “close enough” are retrieved.
- Metrics:
  - Precision = \( \frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|} \)
  - Recall = \( \frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|} \)

Machine Learning

- **Machine Learning (ML)**: area of AI that examines how to devise algorithms that can learn.
- Techniques from ML are often used in classification and prediction.
- **Supervised Learning**: learns by example.
- **Unsupervised Learning**: learns without knowledge of correct answers.
- Machine learning often deals with small or static datasets.

**DM**: Uses many machine learning techniques.

Statistics

- Usually creates simple descriptive models.
- **Statistical inference**: generalizing a model created from a sample of the data to the entire dataset.
- **Exploratory Data Analysis**: - Data can actually drive the creation of the model.
  - Opposite of traditional statistical view.
- Data mining targeted to business users.

**DM**: Many data mining methods are based on statistical techniques.

Point Estimation

**Point Estimate**: estimate a population parameter.
- May be made by calculating the parameter for a sample.
- May be used to predict values for missing data.
- Ex:
  - \( R \) contains 100 employees
  - 99 have salary information
  - Mean salary of these is $50,000
  - Use $50,000 as value of remaining employee’s salary.

Is this a good idea?

Estimation Error

**Bias**: Difference between expected value and actual value.
\[
\text{Bias} = E(\hat{\Theta}) - \Theta
\]

**Mean Squared Error (MSE)**: expected value of the squared difference between the estimate and the actual value:
\[
\text{MSE}(\hat{\Theta}) = E((\hat{\Theta} - \Theta)^2)
\]

- Why square?
- Root Mean Square Error (RMSE).
Jackknife Estimate

- **Jackknife Estimate**: estimate of parameter is obtained by omitting one value from the set of observed values.
- Ex: estimate of mean for $X = \{x_1, \ldots, x_n\}$

$$\hat{\theta}_k = \frac{\sum_{j=1}^{n-1} x_j + \sum_{j=i+1}^n x_j}{n-1}$$

$$\hat{\theta}_k = \frac{\sum_{j=1}^{n} \hat{\theta}(j)}{n}$$

Maximum Likelihood Estimate (MLE)

- Obtain parameter estimates that maximize the probability that the sample data occurs for the specific model.
- Joint probability for observing the sample data by multiplying the individual probabilities. Likelihood function:

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

- Maximize $L$.

MLE Example

- Coin toss five times: (H,H,H,H,T)
- Assuming a perfect coin with H and T equally likely, the likelihood of this sequence is:

$$L(p \mid 1,1,1,1,0) = \prod_{i=1}^{5} P = 0.5 = 0.03.$$  

- However if the probability of a H is 0.8 then:

$$L(p \mid 1,1,1,1,0) = 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.2 = 0.08.$$  

MLE Example (cont’d)

General likelihood formula:

$$L(p \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}.$$  

$$l(p) = \log L(p) = \sum_{i=1}^{n} x_i \log(p) + (5-\sum_{i=1}^{n} x_i) \log(1-p)$$  

$$\frac{\partial l(p)}{\partial p} = \sum_{i=1}^{n} \frac{x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p},$$  

$$p = \frac{\sum_{i=1}^{n} x_i}{n}.$$  

Estimate for $p$ is then $4/5 = 0.8$.

Expectation-Maximization (EM)

Solves estimation with incomplete data.

**Algorithm**

- Obtain initial estimates for parameters.
- Iteratively use estimates for missing data and continue refinement (maximization) of the estimate until convergence.

Expectation Maximization Algorithm

Input:

- $\theta = \{\theta_1, \ldots, \theta_j\}$  //Parameters to be Estimated
- $x_{obs} = \{x_1, \ldots, x_n\}$  //Input Database Values Observed
- $x_{mis} = \{x_{n+1}, \ldots, x_o\}$  //Input Database Values Missing

Output:

- $\hat{\theta}$  //Estimates for $\theta$

**EM Algorithm**:

1. $i := 0$;
2. Obtain initial parameter MLE estimate, $\hat{\theta}$;
3. repeat
   - Estimate missing data, $\hat{x}_{mis}$;
   - $i := i + 1$;
   - Obtain next parameter estimate, $\hat{\theta}$ to maximize data; until estimate converges;
**Expectation Maximization Example**

\[ \begin{align*}
\mu_1 &= \frac{\sum_{i=1}^n x_i}{n} + \frac{\sum_{j=1}^m x_j}{m} = 3.33 + \frac{2 + 3}{6} = 4.33 \\
\mu_2 &= \frac{\sum_{i=1}^n x_i}{n} + \frac{\sum_{j=1}^m x_j}{m} = 3.33 + \frac{4.33 + 4.33}{6} = 4.77 \\
\mu_3 &= \frac{\sum_{i=1}^n x_i}{n} + \frac{\sum_{j=1}^m x_j}{m} = 3.33 + \frac{4.77 + 4.77}{6} = 4.92 \\
\mu_4 &= \frac{\sum_{i=1}^n x_i}{n} + \frac{\sum_{j=1}^m x_j}{m} = 3.33 + \frac{4.92 + 4.92}{6} = 4.97 
\end{align*} \]

---

**Models Based on Summarization**

- **Visualization:** Frequency distribution, mean, variance, median, mode, etc.
- **Box Plot:**

**Bayes Theorem**

- **Posterior Probability:** \( P(h|x) \)
- **Prior Probability:** \( P(h) \)
- **Bayes Theorem:**

\[
P(x) = \sum_{j=1}^{m} P(x | h_j) P(h_j).
\]

- Assign probabilities of hypotheses given a data value.

**Bayes Theorem Example**

- Credit authorizations (hypotheses):
  - \( h_1 = \) authorize purchase,
  - \( h_2 = \) authorize after further identification,
  - \( h_3 = \) do not authorize,
  - \( h_4 = \) do not authorize but contact police
- Assign twelve data values for all combinations of credit and income:

<table>
<thead>
<tr>
<th>ID</th>
<th>Income</th>
<th>Credit</th>
<th>Class</th>
<th>ID</th>
<th>Income</th>
<th>Credit</th>
<th>Class</th>
<th>ID</th>
<th>Income</th>
<th>Credit</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>Excellent</td>
<td>x_1</td>
<td>2</td>
<td>3</td>
<td>Good</td>
<td>x_2</td>
<td>3</td>
<td>2</td>
<td>Excellent</td>
<td>x_3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Good</td>
<td>x_4</td>
<td>5</td>
<td>4</td>
<td>Good</td>
<td>x_5</td>
<td>6</td>
<td>2</td>
<td>Excellent</td>
<td>x_6</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>Bad</td>
<td>x_7</td>
<td>8</td>
<td>2</td>
<td>Bad</td>
<td>x_8</td>
<td>9</td>
<td>3</td>
<td>Bad</td>
<td>x_9</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>Bad</td>
<td>x_10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- From training data: \( P(h_1) = 60\% \); \( P(h_2) = 20\% \); \( P(h_3) = 10\% \); \( P(h_4) = 10\% \).
Bayes Example (cont’d)

- Calculate $P(x_i|h_j)$ and $P(x_i)$.
- Ex: $P(x_1|h_1)=2/6$; $P(x_2|h_1)=1/6$; $P(x_3|h_1)=2/6$; $P(x_4|h_1)=1/6$; and $P(x_i|h_j)=0$ for all other $x_i$.
- Predict the class for $x_4$:
  - Calculate $P(h_j|x_4)$ for all $h_j$.
  - Place $x_4$ in class with largest value.
  - Ex:
    - $P(h_1|x_4) = (P(x_1|h_1)P(h_1))/P(x_4) = (1/6)0.6/0.1 = 1.$
    - $x_4$ in class $h_1$.

Hypothesis Testing

- Find model to explain behavior by creating and then testing a hypothesis about the data.
- Exact opposite of usual DM approach.
- $H_0$ - Null hypothesis; Hypothesis to be tested.
- $H_A$ - Alternative hypothesis.

Chi Squared Statistic

- $O$ - observed value
- $E$ - Expected value based on hypothesis.

\[ \chi^2 = \sum \frac{(O - E)^2}{E} \]

Ex:
- $O = (50, 93, 67, 78, 87)$
- $E = 75$
- $\chi^2 = 15.55$ and therefore significant

Regression

- Predict future values based on past values
- Linear Regression assumes that a linear relationship exists.
  \[ y = c_0 + c_1 x_1 + \ldots + c_n x_n \]
- Find values to best fit the data

Linear Regression

Correlation

- Examine the degree to which the values for two variables behave similarly.
- Correlation coefficient $r$:
  - $1$ = perfect correlation
  - $-1$ = perfect but opposite correlation
  - $0$ = no correlation

\[ r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \]
Similarity Measures

- Determine similarity between two objects.
- Characteristics of a good similarity measure:
  - \( \forall t_i \in D, \text{sim}(t_i, t_i) = 1 \)
  - \( \forall t_i, t_j \in D, \text{sim}(t_i, t_j) = 0 \) if \( t_i \) and \( t_j \) are not alike at all.
  - \( \forall t_i, t_j \in D, \text{sim}(t_i, t_j) < \text{sim}(t_i, t_k) \) if \( t_i \) is more like \( t_j \) than it is like \( t_k \).
- Alternatively, distance measures indicate how unlike or dissimilar objects are.

Commonly Used Similarity Measures

- Dice: \( \text{sim}(t_i, t_j) = \frac{2 \sum_{k=1}^{K} t_{ik} t_{jk}}{\sum_{k=1}^{K} t_{ik}^2 + \sum_{k=1}^{K} t_{jk}^2} \)
- Jaccard: \( \text{sim}(t_i, t_j) = \frac{\sum_{k=1}^{K} t_{ik} t_{jk}}{\sum_{k=1}^{K} t_{ik}^2 + \sum_{k=1}^{K} t_{jk}^2 - \sum_{k=1}^{K} t_{ik} t_{jk}} \)
- Cosine: \( \text{sim}(t_i, t_j) = \frac{\sum_{k=1}^{K} t_{ik} t_{jk}}{\sqrt{\sum_{k=1}^{K} t_{ik}^2 \sum_{k=1}^{K} t_{jk}^2}} \)
- Overlap: \( \text{sim}(t_i, t_j) = \frac{\sum_{k=1}^{K} t_{ik} t_{jk}}{\min(\sum_{k=1}^{K} t_{ik}, \sum_{k=1}^{K} t_{jk})} \)

Distance Measures

Measure dissimilarity between objects

- Euclidean: \( \text{dis}(t_i, t_j) = \sqrt{\sum_{k=1}^{K} (t_{ik} - t_{jk})^2} \)
- Manhattan: \( \text{dis}(t_i, t_j) = \sum_{k=1}^{K} |(t_{ik} - t_{jk})| \)

Twenty Questions Game

Decision Trees

Decision Tree (DT):
- Tree where the root and each internal node is labeled with a question.
- The arcs represent each possible answer to the associated question.
- Each leaf node represents a prediction of a solution to the problem.

Popular technique for classification: Leaf nodes indicate classes to which the corresponding tuples belong.

Decision Tree Example
### Decision Trees

- A **Decision Tree Model** is a computational model consisting of three parts:
  - Decision Tree
  - Algorithm to create the tree
  - Algorithm that applies the tree to data
- Creation of the tree is the most difficult part.
- Processing is basically performing a search similar to that in a binary search tree (although DT may not be binary).

### Decision Tree Algorithm

**Input:**
- $T$ // Decision Tree
- $D$ // Input Database

**Output:**
- $M$ // Model Prediction

DTPred Algorithm:
- \( \text{// Illustrates Prediction Technique using DT} \)
  - for each \( t \in D \) do
    - $n \leftarrow \text{root node of } T$
    - while \( n \) not leaf node do
      - Obtain answer to question on \( n \) applied to \( t \)
      - Identify arc from \( t \) which contains correct answer
      - $n \leftarrow \text{node at end of this arc}$
    - Make prediction for \( t \) based on labeling of \( n \)

### Decision Trees: Advantages & Disadvantages

**Advantages:**
- Easy to understand.
- Easy to generate rules from.

**Disadvantages:**
- May suffer from overfitting.
- Classify by rectangular partitioning.
- Do not easily handle nonnumeric data.
- Can be quite large—pruning is necessary.