Information Visualization: Fundamental Concepts

NOVA IMS Course Notes

Introduction to Information Visualization

Information visualization is the graphical representation of data to enhance understanding and communication. It serves two main purposes:

- **Presentation**: Communicating insights to others
- **Analysis**: Exploring data to discover patterns and relationships

Before Starting Analysis

1. Curse of Dimensionality

- As the number of features (dimensions) increases, data becomes sparse
- Fixed-size training sets cover decreasing fractions of input space
- Makes generalization harder and requires more data
- High-dimensional spaces exhibit "weird" effects and difficult visualization

2. Separability and Bayes Error

- **Separable**: Classes can be perfectly distinguished (zero error possible)
- **Not separable**: Always some error exists
- **Bayes Error**: Lowest possible error rate for any classifier
- Important for understanding fundamental limits of classification

3. Types of Measurements

- **Nominal**: Categories without order (colors, labels)
- **Ordinal**: Ordered categories (satisfaction levels)
- **Interval**: Equal intervals, arbitrary zero (temperature in °C)
- **Ratio**: True zero point, meaningful ratios (height, weight)

4. Exploratory Data Analysis (EDA)

- Initial investigation of data to discover patterns
- Detect outliers, test hypotheses, check assumptions
- Uses visual methods to understand data structure

Information Visualization Guidelines

Tufte's Principles of Graphical Excellence

- Show the greatest number of ideas in shortest time
- Use least ink in smallest space
- Tell the truth about the data
- Maximize data-ink ratio
- Minimize chartjunk (decorative elements)

Lie Factor

- Measures distortion in graphical representation:

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

- Ideal range: 0.95 ¡ Lie Factor ¡ 1.05
- Avoids exaggeration or minimization of effects

Graphics for Presentation

Effective Presentation Graphics

- **Bar Charts**: Compare categorical data
- **Line Charts**: Show trends over time
- **Pie Charts**: Show parts of a whole (use sparingly)
- **Stacked Bars**: Show composition and comparison

Best Practices

- Maximize contrast between data and background
- Use meaningful ordering (alphabetical, by value, chronological)
- Choose appropriate scales and ranges
- Provide clear labels and titles
- Use consistent color schemes

Graphics for Analysis

Exploratory Visualization Types

- **Scatter Plots**: Relationships between two continuous variables
- **Histograms**: Distribution of single variable
- **Box Plots**: Distribution summary with outliers
- **Correlation Matrices**: Relationships between multiple variables

Advanced Visualization Techniques

- **Parallel Coordinates**: Multivariate data analysis
- **Small Multiples**: Multiple similar graphs for comparison
- **Heat Maps**: Matrix data with color coding
- **Tree Maps**: Hierarchical data as nested rectangles
- **Radar Charts**: Multivariate data on radial axes
- **Geo-visualization**: Spatial data on maps
- **Linked Views**: Multiple coordinated visualizations

Analysis-Specific Considerations

- Focus on data exploration rather than polished appearance
- Use interactive features for deeper investigation
- Consider data density and overplotting issues
- Support drill-down capabilities for detailed analysis

Key Takeaways

- Choose visualization type based on data type and analysis goal
- Follow design principles to ensure accurate representation
- Balance aesthetics with functionality
- Use appropriate tools for presentation vs. analysis
- Always consider the audience and communication objective