BS Data Science

Subject: Text Mining

Day 3: Date: 19/11/2024

Topic : Feature Extraction in Data Analysis

Goal

To introduce students to the concept of feature extraction and its significance in preparing data for machine learning and analytical models.

Objectives:

1. Explain the purpose and importance of feature extraction in data preprocessing.
2. Demonstrate common techniques for extracting features from text, images, and audio data.
3. Provide practical examples to illustrate how feature extraction improves model performance and insights.

Slide 2: Introduction to Feature Extraction

- **Definition:**
  Feature extraction is the process of transforming raw data into a set of features that can be effectively used by machine learning models.
- **Importance:**
  - Simplifies raw data.
  - Enhances model performance.
  - Reduces computational complexity.

Slide 3: Goals of Feature Extraction

1. Reduce data dimensionality while retaining meaningful information.
2. Identify key patterns and characteristics in the data.
3. Prepare data for machine learning or statistical analysis.

Slide 4: Types of Data for Feature Extraction

1. **Numerical Data**: Sales figures, temperatures, etc.
2. **Textual Data**: Emails, tweets, documents.
3. **Image Data**: Pictures, videos.
4. **Audio Data**: Speech, music.

Slide 5: Feature Extraction in Text Data

Common Techniques:

1. **Bag of Words (BoW):**
   - Converts text into a frequency matrix.
   - Example:
     - Input: "Text mining is fun."
     - Output: {"text": 1, "mining": 1, "is": 1, "fun": 1}.
2. **TF-IDF (Term Frequency-Inverse Document Frequency):**
   - Measures the importance of words in a document relative to a collection of documents.
   - Example: High TF-IDF for unique terms like "neural networks" in an AI article.
3. **Word Embeddings (e.g., Word2Vec, GloVe):**
   - Converts words into dense vector representations.
   - Example: The word "king" might have a vector close to "queen".

Slide 6: Feature Extraction in Numerical Data

1. **Descriptive Statistics:**
   - Extract features like mean, median, variance, etc.
   - Example: Stock prices → Mean daily return, volatility.
2. **Normalization and Scaling:**
   - Normalize features between 0 and 1.
   - Example: Standardizing heights and weights for comparison.
3. **Polynomial Features:**
   - Create higher-degree terms from features.
   - Example: From feature x, create x^2, x^3.

Slide 7: Feature Extraction in Images

Techniques:

1. **Edge Detection:**
   - Highlights object boundaries.
   - Example: Canny edge detector in image processing.
2. **Histograms of Oriented Gradients (HOG):**
   - Captures object shapes and orientations.
   - Example: Detecting vehicles in a traffic image.
3. **Deep Features with CNNs (Convolutional Neural Networks):**
   - Automatically learns hierarchical features like edges, textures, and shapes.
   - Example: Recognizing faces in photos.

Slide 8: Feature Extraction in Audio

1. **Spectral Features:**
   - Extract pitch, frequency, and energy.
   - Example: Recognizing speech patterns.
2. **Mel Frequency Cepstral Coefficients (MFCC):**
   - Used for speech recognition.
   - Example: Transcribing spoken commands to text.
3. **Zero-Crossing Rate:**
   - Measures how often the signal crosses zero amplitude.
   - Example: Differentiating voiced and unvoiced speech.

Slide 9: Practical Example: Feature Extraction in Text Classification

Dataset: Movie Reviews (Positive/Negative).
Steps:

1. **Tokenization:** Split sentences into words.
2. **Feature Engineering:**
    - Extract BoW and TF-IDF features.
    - Generate embeddings with Word2Vec.
3. **Output Features:** Use features as input for a machine learning model (e.g., logistic regression).

Slide 10: Practical Example: Feature Extraction in Images

Task: Object Detection in Traffic Images.
Steps:

1. Apply HOG to detect vehicle shapes.
2. Use CNN to extract deep features for advanced classification.
3. Train a machine learning model to classify objects (cars, bikes, pedestrians).

Slide 11: Challenges in Feature Extraction

1. Identifying the right features for the task.
2. Dealing with noisy or irrelevant data.
3. Balancing complexity and interpretability.

Slide 12: Tools for Feature Extraction

- **Text Data:** NLTK, SpaCy, Scikit-learn, Gensim.
- **Image Data:** OpenCV, TensorFlow, PyTorch.
- **Audio Data:** Librosa, PyDub.

Slide 13: Conclusion

- Feature extraction transforms raw data into usable forms.
- Effective features improve model performance significantly.
- Proper tools and techniques simplify the process.

Lab 1:

Formulate bag of word model. Apply Bag of Word Model to the following Example.

Example of Bag of Words

Corpus of Documents:

1. **Document 1:** "Text mining is fun."
2. **Document 2:** "Text mining helps extract insights."

.