Classical Machine Learning for NLP

Summer

Introduction

Classical Machine Learning (ML) for NLP focuses on traditional algorithms used before the advent of deep learning. These models are simpler, faster, and require smaller datasets, making them suitable for many NLP tasks.

Key Algorithms in Classical ML for NLP

- **Naïve Bayes**:
  - Probabilistic classifier based on Bayes' theorem.
  - Assumes conditional independence between features given the class label.
  - Fast and requires low storage.
- **Support Vector Machines (SVMs)**: Finds the hyperplane that maximizes class separation.
- **Decision Trees**: Tree-based approach for classification or regression.

Key Features

- Simpler models compared to deep learning.
- Require feature engineering but are more interpretable.
- Effective for smaller datasets and cheaper to compute.

Probability Basics

- **Conditional Probability**: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.
- **Bayes' Rule**: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.
- **Independent Events**: $P(A \cap B) = P(A)P(B)$.

Text Representation

- **Bag of Words (BoW)**:
  - Represents a document as an unordered collection of word counts.
  - Does not encode word order or semantics.
- **TF-IDF (Term Frequency-Inverse Document Frequency)**:
  - Highlights important words by weighting terms inversely proportional to their frequency in the corpus.
  - Formula: $\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$.
- **N-Grams**: Captures local word dependencies by using contiguous sequences of n words or characters.

Evaluation Metrics

- **Accuracy**: Proportion of correct predictions.
- **Precision**: $\frac{TP}{TP+FP}$ (relevant retrieved items).
- **Recall**: $\frac{TP}{TP+FN}$ (relevant items correctly predicted).
- **F1-Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Trade-off between true positive rate and false positive rate.

Example: Naïve Bayes Classifier for Spam Detection

- Assumes conditional independence between words given the class (spam/ham).
- Uses Laplace smoothing to handle unseen words.
- Compares approaches using BoW and TF-IDF.

Conclusion

Classical ML methods like Naïve Bayes, SVMs, and Decision Trees remain relevant in NLP for their simplicity, interpretability, and efficiency. Proper text representation and evaluation are crucial for achieving robust performance.

References

- Chadha, Aman. *Sentiment Analysis Using Naive Bayes.* Notes from Coursera NLP Specialization.
- Karen Spärck Jones. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval.*

- Hugging Face Enron Dataset: `https://huggingface.co/datasets/SetFit/enron_spam`