Classification Algorithms for Mobile Phone Dataset - Proposal

Team: Rakesh Kumar Busa, Jeremy Driscoll, Daniel Ackom

Introduction

Effective product classification is essential for organizing and navigating e-commerce catalogs. Traditional rule-based and manual tagging systems are increasingly inadequate due to their labor-intensive nature and susceptibility to error. Recent advancements in machine learning provide scalable alternatives for automating this process. This study investigates the application of five machine learning algorithms—Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Naive Bayes, and Support Vector Machine (SVM)—to categorize products based on textual descriptions. The objective is to evaluate each model in terms of classification accuracy, interpretability, and computational efficiency, thereby identifying the most effective approach for this task.

Research Question

Which machine learning algorithm—LDA, QDA, KNN, Naive Bayes, or SVM—offers the most accurate and interpretable solution for product categorization based on textual input?

Dataset

The dataset, sourced from the UC Irvine Machine Learning Repository, includes 35,311 product listings across 10 categories from 306 merchants. Each entry consists primarily of a product title along with associated metadata (e.g., Product ID, Merchant ID, and Category Label). The dataset presents typical challenges in text classification: high variability in language, class imbalance, noisy and sparse data—all of which make it suitable for evaluating algorithmic robustness in real-world conditions.

Methodology

The study consists of four stages:

1. **Preprocessing**: Product titles are cleaned and vectorized using TF-IDF or word embeddings. Categorical features are encoded accordingly.
2. **Model Selection**: The selected models represent a spectrum of statistical and algorithmic approaches—parametric (LDA, QDA), non-parametric (KNN), probabilistic (Naive Bayes), and margin-based (SVM).
3. **Evaluation**: Models are assessed via cross-validation using accuracy, precision, recall, F1-score, and confusion matrices.
4. **Hyperparameter Tuning**: Grid search or random search methods are employed to optimize model performance.

Justification of Models

These algorithms were chosen to capture diverse modeling philosophies. LDA and QDA are useful for understanding class distributions; KNN is adaptable to high-dimensional sparse data; Naive Bayes offers efficiency and is well-suited for text; SVM is robust in high-dimensional spaces, making it a strong candidate for textual input. Together, these methods provide a comprehensive framework for evaluating automated product classification.

Conclusion

This study aims to benchmark the performance of five machine learning models for product categorization in e-commerce. By analyzing their performance on a realistic and challenging dataset, the research will inform the best practices in text-based classification tasks and contribute to the ongoing development of intelligent product recommendation and organization systems