Anime Score Classification using Support Vector Machine

1. Introduction

The goal of this assignment is to classify anime shows into different score categories based on metadata such as number of episodes, duration, genres, producers, studios, licensors, source material, and rating. Machine learning provides a systematic way to extract patterns from such data and predict the popularity or quality of shows. Support Vector Machine (SVM) was chosen as the classification algorithm due to its effectiveness in handling high-dimensional feature spaces.

2. Dataset Description

The dataset (anime.csv) contains metadata about anime shows, including numerical attributes (episodes, duration), categorical attributes (source, rating), and multi-label categorical attributes (genres, producers, studios, licensors). The target variable is the score class, derived from the numerical score as follows: - Very Good: Score ≥ 8.0 - Good: 7.0 ≤ Score < 8.0 - Average: 6.0 ≤ Score < 7.0 - Low: Score < 6.0 The dataset had missing values, which were handled by imputing medians or assigning 'Unknown' for categorical features.

3. Methodology

The methodology involved the following steps: 1. Data cleaning: Removed invalid entries and handled missing values. 2. Feature engineering: - Duration converted to minutes. - Multi-label categorical features encoded into binary columns using top-K frequent items. - Source and Rating features one-hot encoded. 3. Feature scaling applied using StandardScaler. 4. Model training: A Support Vector Machine (SVC with default parameters) was trained on the processed data. Sample preprocessing code:

Code Snippet

```
df['Score'] = pd.to_numeric(df['Score'], errors='coerce')
df = df.dropna(subset=['Score']).reset_index(drop=True)

def bucket_score(s):
    if s >= 8.0: return "Very Good"
    elif s >= 7.0: return "Good"
    elif s >= 6.0: return "Average"
    else: return "Low"

df['Score_Class'] = df['Score'].apply(bucket_score)
```

4. Results & Analysis

The SVM classifier achieved an accuracy of approximately 63.5%. The performance was better for the 'Average' and 'Low' classes compared to 'Very Good', which had fewer samples and lower recall (~0.40). The confusion matrix showed that most misclassifications occurred between 'Good' and 'Average'. Key Observations: - Multi-label features such as genres and studios were influential. - Class imbalance affected performance. - The model can benefit from hyperparameter tuning and advanced techniques.

5. Discussion & Conclusion

This project demonstrated the application of SVM to classify anime shows based on metadata. Although the accuracy (~63.5%) indicates moderate success, there is significant room for improvement. Future work can include: - Hyperparameter tuning (kernel, C, gamma). - Ensemble methods (Random Forest, XGBoost, Gradient Boosting). - Dimensionality reduction (PCA or feature selection). - Embedding-based representation for multi-label categorical features. Overall, the experiment highlighted both the potential and limitations of traditional machine learning approaches on complex, high-dimensional categorical datasets.