Problem Set 3

Dataset	Type	Length	IMDb Rating	Liked?
Train	Comedy	Short	7.2	Yes
Train	Drama	Medium	9.3	Yes
Train	Comedy	Medium	5.1	No
Train	Drama	Long	6.9	No
Train	Drama	Medium	8.3	Yes
Train	Drama	Short	4.5	No
Train	Comedy	Short	8.0	Yes
Train	Drama	Medium	7.5	Yes
Train		Medium	6.7	Yes
Train	Comedy		3.6	Yes
Train	Drama	Long		Yes
Test	Drama	Medium	8.1	Yes
Test	Comedy	Short		Yes
Test		Long	7.5	Yes

Table 1: Training and test data.

1. **Data Clean-Up (8 points)**: pre-process the data in **Table 1** for this exercise.

   (a) (One-Hot Encoding) Report in a new table the resulting dataset after encoding the categorical features with a one-hot encoding.

   (b) (Imputing Missing Values) Report in a new table the resulting dataset after imputing all missing values using the feature mean for examples belonging to the same class.

   (c) (Feature Scaling) Report in a single table the IMDb values for all data as follows:

   - Column 1: resulting values after min-max scaling
   - Column 2: resulting values after standardization

2. **Dimensionality Reduction (5 points)**

   (a) Name three uses for dimensionality reduction techniques.

   (b) In your own words, describe when to use Principle Component Analysis (PCA) versus Locally Linear Embedding (LLE) to reduce the feature dimensionality.

3. **Classification Evaluation (4 points)**

   (a) What is a "precision-recall curve" (PR curve) and how can it be used when designing a machine learning system?

   (b) What is the relationship between a PR curve and a confusion matrix?

(c) What is a "ROC curve" and how can it be used when designing a machine learning system?

(d) What is the relationship between a ROC curve and a confusion matrix?

4. **Challenge Analysis (8 points)**: find two machine learning competitions (challenges) on the platform Kaggle. For each competition, write a response to the following items:

   (a) Describe the motivation for the competition.

   (b) Describe the machine learning task. You must include a discussion of what is the raw input (e.g., images, text, etc) and what are the target labels a machine learning must predict.

   (c) Describe how the dataset is partitioned for training and evaluation (e.g., train/val/test split?) and how many examples are included for each partition.

   (d) Describe the evaluation metrics used to assess the performance of algorithms submitted for the competition.