MSE 446: Intro To Machine Learning

University of Waterloo

Professor Lukasz Golab

December 3rd, 2024

Utku Çiçek (20910452)
Lisa Gavronsky (20882271)
Mathew Maradin (20877924)

Problem Introduction:

This project aims to derive insights from movie poster image data to classify films by genre. To classify movies, three approaches will be taken including a KNN algorithm using average RGB values, a logistic regression utilizing frequency of the most relevant colours and a neural network that will learn patterns based on colour pixels in an attempt to classify genres of movie posters. After gathering insights from each algorithm on the correlations between the colour distribution of movie posters and their genres, the algorithms' accuracy will be compared to draw conclusions on their efficiencies.

Colour is a powerful communication tool that can be used to convey mood, tone and even influence physiological reactions [1]. Colours have emotional associations that influence how people when they come across a poster. For example, warm colours portray excitement, intensity or danger whereas cool colours suggest calmness, mystery and sadness. We believe that the colour distribution of movie posters plays a distinctive role in conveying emotion and, therefore, is highly correlated with the genre. Based on visual inspection, movie posters that display genres like mystery or horror tend to have black and red as these colours display violence, danger, fear and death. In contrast, genres like romance tend to have warm colours like pink and soft red that display love, joy and warmth. We are trying to address the hypothesis: *"Can the colour distribution of a movie poster be a reliable predictor of its genre?"*.

Description of Data:

"Movie Posters" is a dataset on Kaggle and it consists of a collection of images of movie posters and metadata to explain the different categories each movie poster belongs to. The images provided come in various sizes and the metadata comes in a .csv format. In order to understand the data, Pandas library was utilized to convert the .CSV file into a dataframe, making it easier to perform operations on and increasing readability. Each row in the data represents a movie and includes a unique image identifier (ID), the movie's title, a list of genres associated with the movie and binary columns for each genre, where a value of 1 indicates the movie belongs to that genre, and 0 indicates it does not. There are a total of 25 distinct classes, however, this number drops to 22 when 3 classes with no corresponding images are removed. As the images come in various sizes, the resizing operation was done to fix the size of each poster to

(350, 350, 3), each poster was then converted into a NumPy array for efficient matrix calculations. When analyzing the data, it's visible that there is an imbalance of data with ~22% and ~19% of the data belonging to drama and comedy classes while the rest of the classes represent less than 8% of the data each. In order to eliminate a possible bias towards the classes, genres with less than 400 data points are removed from the dataset, leaving us with a total of 5 genres. Based on Figure 2, it can be seen that comedy, drama, and romance are a popular blend of genres for a film, possibly due to the appeal of combining humour, and romance with serious storytelling, therefore, we can assume that it will be hard to tell apart popular pairs as they tend to have the same movie poster. To make our data simpler and eliminate the co-occurence positive correlation between certain labels, at the recommendation of Professor Golab, we removed the following classes: comedy, crime and romance. Our dataset is now simplified to a binary classification problem with drama and action being the only 2 genres that films can belong to. Action was chosen as the second genre as it has the least co-occurrence with the drama class. Following this, the images belonging to both classes were removed to avoid the redundancy of having the same data point for both classes. Since the colour will be the main feature that will be used in all the variety of machine learning models, 20 random samples from the 2 different classes were selected and graphed using the mean RGB values against each other. This was done to assess whether the data looks visually separable and get an idea of how well our algorithms can possibly perform as well as selecting the feature that is most distinct. Based on Figure 3, mean red and mean blue values will be used as the input features for the machine learning algorithms. Mean green value was then removed from the data, as the two genres look to be the most distinguishable when graphed against each other for the mean red and blue values, compared to the other colour pairings. Despite removing the mean green value, the mean red and blue features will be referred to as the mean RGB values later in the report for ease of understanding.

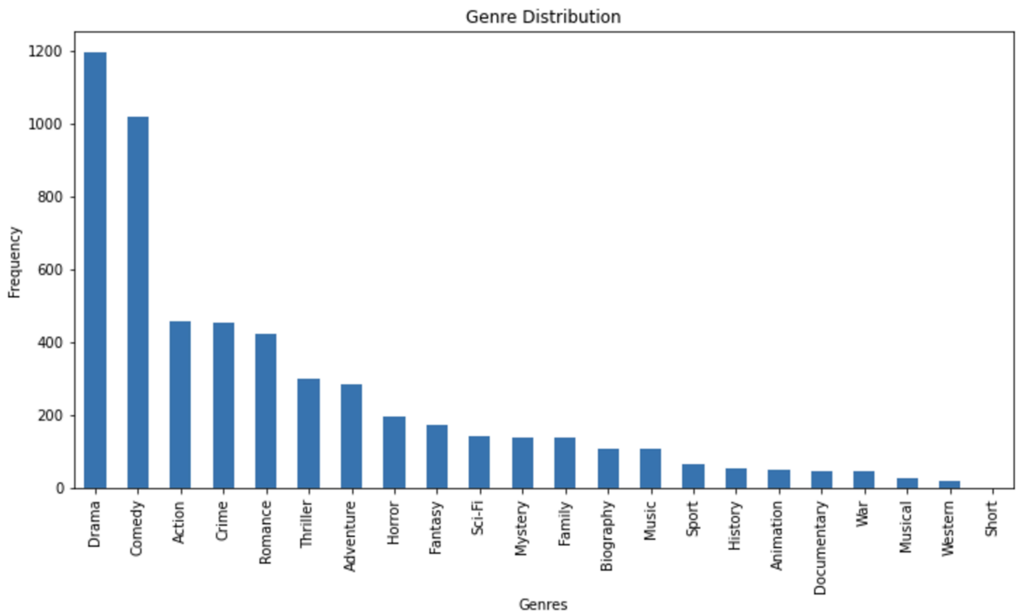

Figure 1: Frequency plot of movie genres

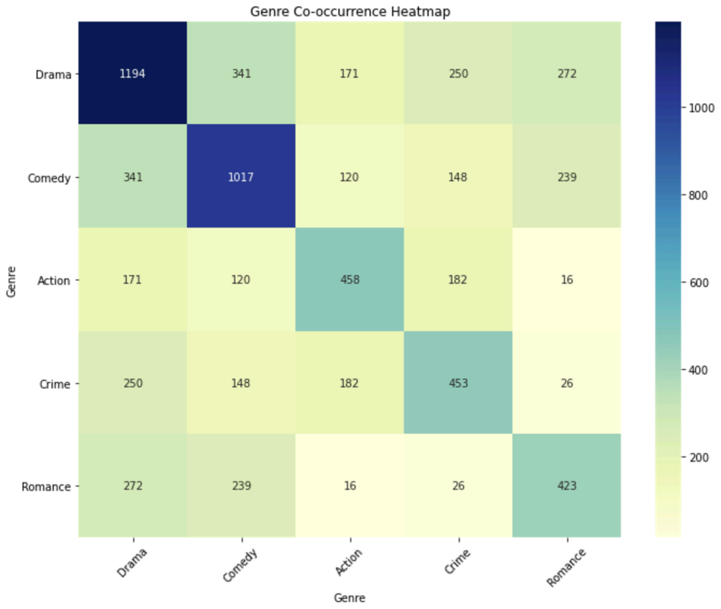

Figure 2: Confusion matrix of co-occurring genres

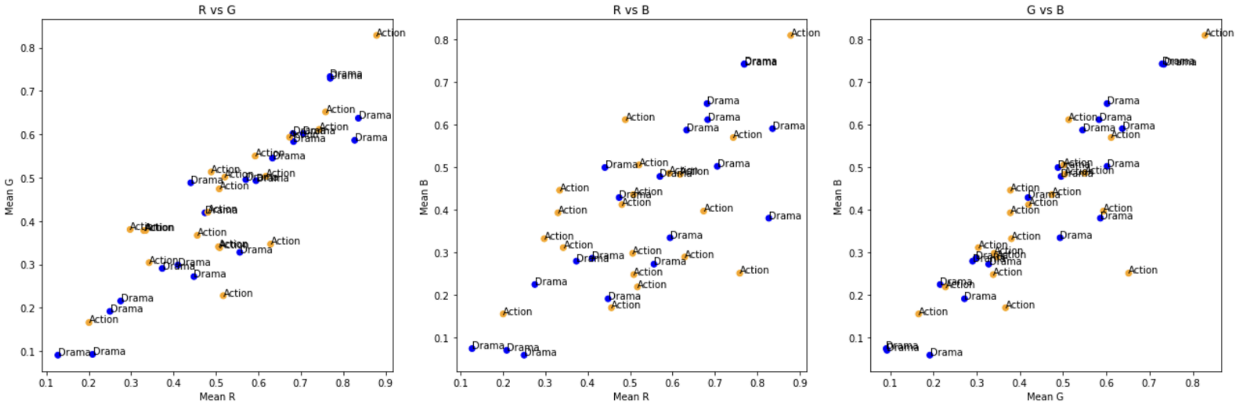

Figure 3: Mean RGB values of movie posters and associated genres

Methods:

To gain insight on the relationship between movie poster colour distributions and their genres, the following machine learning algorithms were tested: K-Nearest Neighbours (KNN), logistic regression, and a neural network. Each of the models were tested and compared using 5-fold cross-validation, with accuracy scores being reported for each one. This testing framework was selected to ensure reliable evaluation of unseen data while maximizing the amount of data used for training, and as a result, reducing the risk of overfitting.

K-Nearest Neighbours (KNN)

KNN is a simple supervised machine learning algorithm used for classification and regression tasks. The algorithm makes decisions based on the principle that similar data points exist close to each other. KNN is easy to understand and simple to implement, therefore, it can be a perfect choice to get quick results. It's also ideal for non-linear data since there are no assumptions made about the underlying data, and it can handle multi-class classification. However, every algorithm comes with its own issues, KNN algorithms tend to be sensitive to noise and computationally expensive. It also tends to perform poorly when it comes to complex data with high dimensions, such as images. For example, an image with a size of (350, 350, 3) has a total of 367,500 pixels and each pixel can be considered a feature leading to high dimension feature space. In order to address the curse of dimensionality for the KNN algorithm, the mean of RGB values was utilized in an attempt to describe the image in a low dimension and use the algorithm to classify a variety of images into their respective classes. However, images are much more complex than just pixel values and they represent complex information when the pixels are arranged correctly. It's hard to capture features like context, and hierarchical features when using the KNN algorithm.

Logistic Regression

Logistic regression is another supervised learning algorithm that is used to classify the genre of movie posters. Similar to the KNN algorithm, the complexity of input data was reduced by using the mean RGB values from each movie poster as the feature, instead of a (350, 350, 3) high-dimensional feature space. Given the mean RGB value that describes the movie poster, the logistic regression algorithm returns a value of 0 or 1, indicating what genre the movie's average RGB values are associated with, being either drama or action. The logistic regression model is beneficial for its simplicity and interpretability, as it does not require class labels to follow any specific distribution. However, it models the target variable using a linear relationship with probabilities, which can be overly simplistic for certain machine learning applications, such as predicting movie genres based on the pixel distributions of movie posters.

Neural Network

A neural network consists of layers of interconnected nodes that process and transform input data to learn patterns and make predictions. The neural network chosen consisted of a total of 7 layers, beginning with 512 neurons and halving with each layer to 256, 128, 64, 32, 16, and 2 output neurons to account for binary classification. Neural networks are regarded as non-linear models and this non-linearity stems from the activation functions applied to each neuron. For every hidden and input layer, ReLu activation was used to introduce non-lineary and address the vanishing gradient problem. Between each layer, a dropout of 0.3 was applied to manage overfitting. This dropout was initially set to 0.5 but was changed as it was too high of a dropout given the size of the neural network and reduced learning too much. Dropout function randomly sets a number of neurons to zero, and encourages the model not to rely on the output of a single neuron activation. This forces the model to generalize better by challenging the model to learn more robust and distributed representations of the data. The mean red and mean blue data were used to train the neural network. Due to the simplification to binary classification, the sigmoid activation function was chosen over softmax, as the sigmoid function outputs a probability between 0 and 1, scaling the favored class closer to 1 and the other closer to 0.

Results:

K-Nearest Neighbours (KNN)

Through data exploration, we realized that images belonging to the same class tend to be relatively close as outlined in Figure 3, therefore, in order to test our hypothesis we believe that KNNs would be an optimal choice to get quick results. When the model was fitted and used to make predictions based on the test data, it achieved an average accuracy of 75% using a 5-fold cross-validation with the number of neighbours n equal to 5. The accuracy seems to be promising despite the KNN's sensitivity to outliers and the significant loss of information when representing an entire image using only mean RGB values. Overall, KNN performed very effectively for an image dataset and achieved an accuracy of 75%

Logistic Regression

The logistic regression model resulted in an average accuracy of 78.1% using 5-fold cross validation. Similar to the KNN algorithm, the logistic regression model is a simple model which can handle multi-class classification. However, logistic regression assumes a linear relationship between the input features and the logarithm of probability of the target class, which may not necessarily be the case for the relationship between the pixel values of movie posters and their corresponding genres. This limitation makes it challenging for logistic regression to capture the true relationship for the data in our given application.

Neural Network

The neural network performed marginally better than the other approaches with an accuracy of 78.12% utilizing k-fold validation for 5 folds. This improvement in accuracy is welcome as it showcases the validity of neural networks as a strong approach to making predictions. However, the neural network comes at the cost of being significantly more involved in setting up and understanding the theory that goes into the code and being much more computationally expensive which is reflected in the long it takes to run when compared to the alternatives tested.

Conclusions:

After analyzing a dataset of movie posters and their corresponding genres and applying various machine learning algorithms, including KNN, logistic regression, and a neural network, we gained valuable insights into both the performance of these algorithms and the relationship between the colour pixels in a movie poster and the corresponding movie genre. Even though the complexity of the input was reduced from a (350, 350, 3) high-dimensional feature space to two features, being the mean red and mean blue values from each movie poster, all of the algorithms performed fairly well in predicting the genre of the poster (action or drama). The final accuracies of each algorithm are as follows: 75% for KNN, 78.1% for logistic regression, and 78.12% for the neural network. These three algorithms achieved similar accuracy, which is likely due to the simplicity of the input. With the reduced feature set, simpler algorithms like KNN and logistic regression have an advantage, as they require less computational power while delivering comparable results. However, if the complexity of the input data was increased, such as using a greater set of pixels from each poster, the neural network algorithm would most likely be able to capture more complex patterns, and thus, achieve an accuracy than the simpler algorithms.

To further develop the idea of classifying movie posters into genres based on color pixels and continue building on this project, we could use a larger dataset with more pixel data by applying less dimensionality reduction and training similar algorithms. Additionally, we could extend the approach beyond just pixel values and incorporate higher-level visual features, such as shapes, by applying convolutional neural networks for classification. Another idea for expanding this project would be to collect a more diverse and balanced dataset of movie posters and their genres. If a dataset with an even distribution of data points across different genres were available, the project could be expanded to test these algorithms on a wider variety of genres, rather than just drama and action.

In conclusion, the machine learning algorithms we trained suggest a correlation between the color distribution of a movie poster and its corresponding genre. This implies that the pixel values of a poster align with the underlying meaning or theme of the movie. Identifying such trends could provide valuable marketing or business insights, allowing colours to be leveraged in designs to make posters for individual genres more effective and appealing. This analysis overall provides an interesting take-away, being that the genre of a movie can be inferred solely from the colour pixels of its posters, and rather than having to rely on textual descriptions, like movie summaries.

Bibliography:

[1] Cherry, Kendra, MSEd. "Color Psychology: Does It Affect How You Feel?" *Verywell Mind*, Dotdash Meredith, 20 Feb. 2024, www.verywellmind.com/color-psychology-2795824.

[2] Raman77786. “MovieClassifier”. Kaggle, 2024, “www.kaggle.com/datasets/raman77768/movie-classifier”.