CS5014 - Machine Learning

P2 - Classification of object colour using optical spectroscopy

Report

140014952

April 20, 2018

1 Overview

The objectives of this practical were to come up with classification model for binary and multi-class classification problems. This submission investigates both binary and multi-class tasks. The solution python scripts can be found in */binaryML/* and */multiclassML/* directories. Corresponding predicted class files are also in these directories.

As later sections suggest, the amount of features it takes to determine the class for each task is very low. This is hypothesized based on input analysis and later machine learning observations support these hypotheses.

2 Binary Classification Task

2.1 Cleaning Data and Feature Extraction

As the very first step input data were split into training and testing data with 70-to-30 ratio. The analysis were first done on the training set. Sklearn's *train_test_split* was used to do so. Also, a seed was used to ensure the same samples are used every time python notebook is fully executed.

From observations and further analysis in python, data cleaning was not necessary. Data contains negative and positive values that according to the practical specification make sense.

However, a functionality to remove all records with null values in was implemented to ensure that the samples are fully prepared.

2.2 Data Analysis and Visualization

To visualize training data X it was plotted with provided wavelength data that contains information about each feature wave length. Additionally, Y training set was used to indicate visually how colours are distributed and can be distinguished. This gave insights as to which features are likely to be good indicators for predicting the colour.

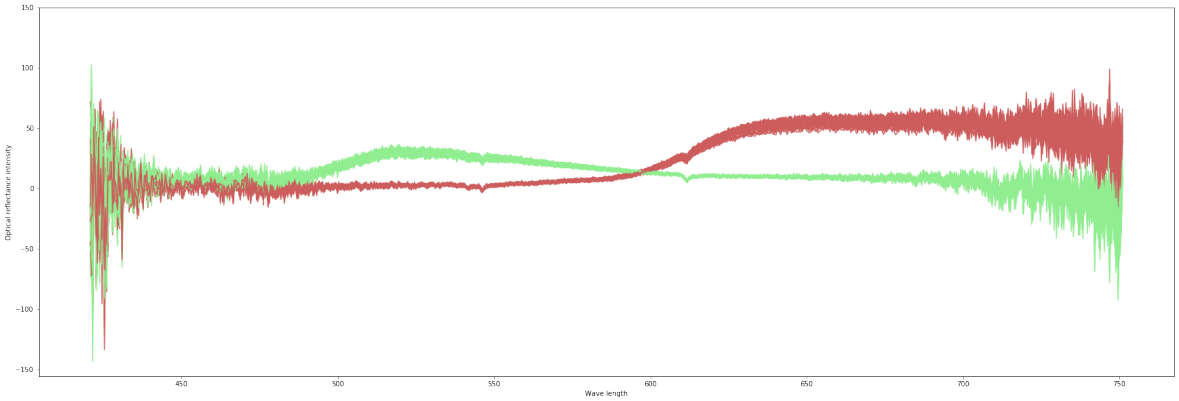

Figure 1: Input feature visualization for binary classification task. Red and green predicting features are indicated by colour.

Fig.1 shows that there is a clear distinction between red and green colours in terms of features. From the same figure it can be said that feature values between 420 and 450 wavelengths are more or less shared between both, red and green colours. Around 600 wavelength both colour features overlap. Similarly, towards the final features similar observations can be made. The graph insights suggest that overlapping features are not great for determining the class because a feature value can be shared by both colours therefore reducing possibility of determining the right colour. However, a single feature that is around 530 or 650 should be good enough to determine the class. From fig.1 it is clear that intensity at those wavelengths are distinct to each colour.

The hypothesis therefore is that any feature that distinguishes the two colours at particular wavelength will be good enough to determine the class. According to the fig.1 there are many of these features: at wavelengths 500-to-580 and 610-to-720. Therefore, it one feature should be enough to determine the class.

2.3 Preparing Inputs and Choosing Features

To choose an appropriate feature an experiment was conducted. Since a single feature could possibly determine a class a training set was

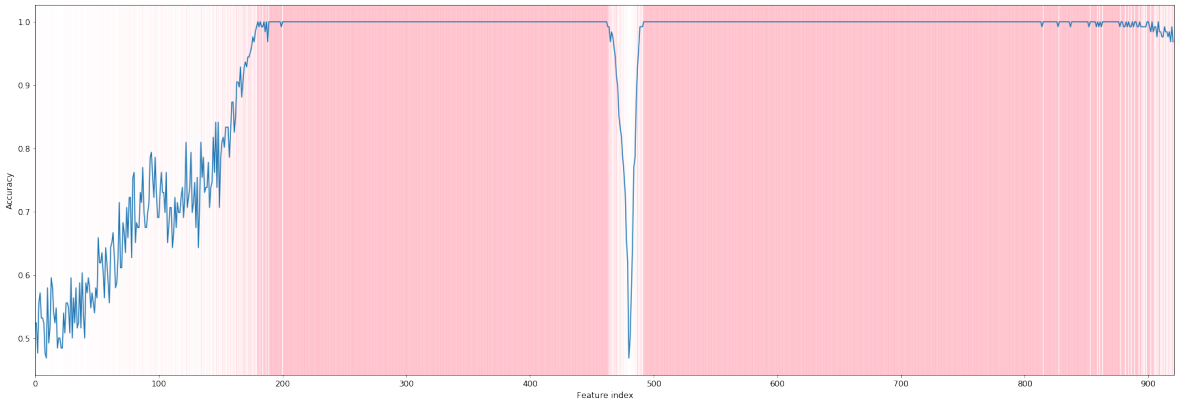

Figure 2: One feature accuracy scores for binary classification.

2.4 Selecting and Training Classification Models

2.5 Evaluating and Comparing Model Performance

2.6 Result Discussion

3 Multi-Class Task

3.1 Cleaning Data and Feature Extraction

Similarly to binary task, no cleaning or feature extraction was required.

3.2 Data Analysis and Visualization

Similarly, fig.3 shows different colour reflectance intensities for five different colours. Similar observation can be seen here too. However, due to larger number of colour classes some seem to overlap slightly.

From fig.**??** it can be seen that red and pink colour intensities over different wavelengths are very similar. Green and blue are also similar. Yellow on the other hand is very distinct. Just from looking at the graph it can be seen that there are not many features that

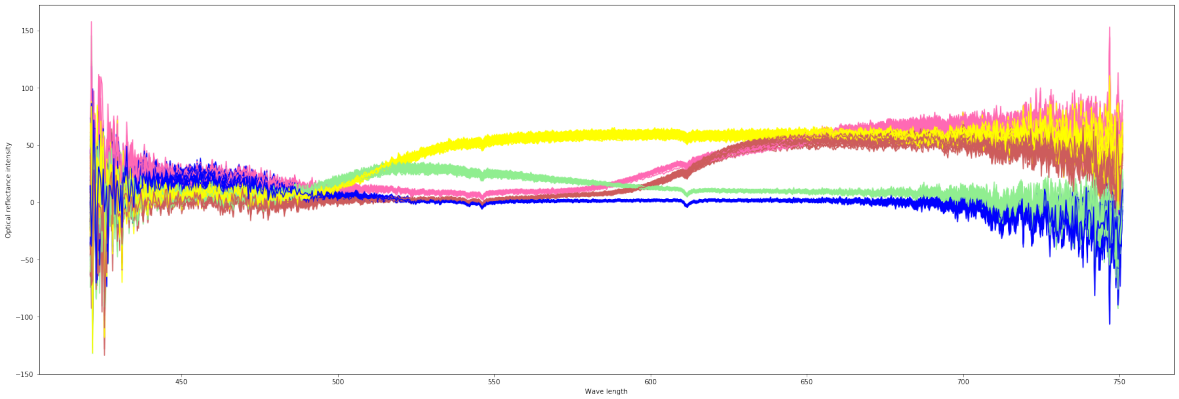

Figure 3: Input feature visualization for multi-class task. Five colour predicting features are indicated by corresponding colour.

3.3 Preparing Inputs and Choosing Features

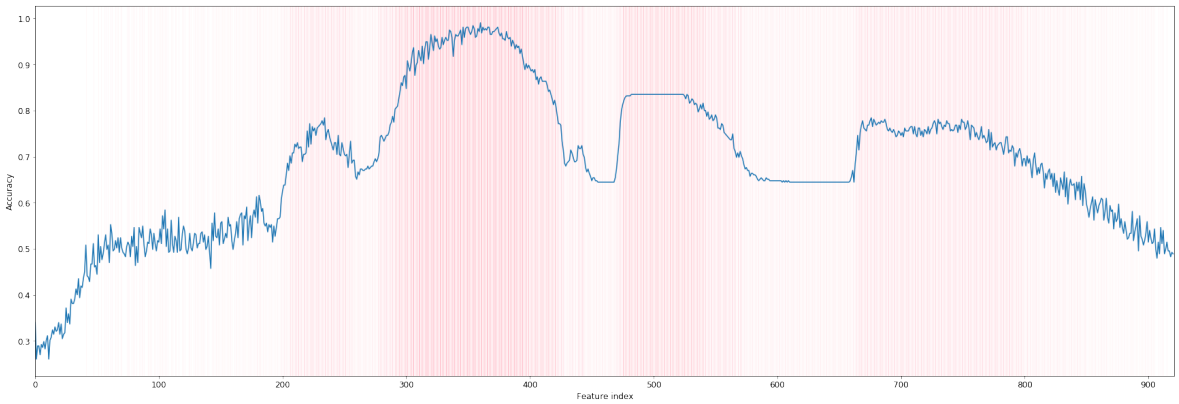

Figure 4: One feature accuracy scores for multi-class classification.

3.4 Selecting and Training Classification Models

3.5 Evaluating and Comparing Model Performance

3.6 Result Discussion

4 Conclusion

No	Feature Indexes	Accuracy
1	421	0.810
2	421, 429	0.851
3	421, 429, 250	0.997
4	421, 429, 250, 251	0.997
5	421, 429, 250, 251, 86	1.0
6	421, 429, 250, 251, 86, 586	1.0
7	421, 429, 250, 251, 86, 586, 88	1.0
8	421, 429, 250, 251, 86, 586, 88, 66	1.0
9	421, 429, 250, 251, 86, 584, 88, 66, 586	1.0
10	421, 429, 250, 251, 86, 584, 88, 66, 586, 914	1.0

Table 1: My caption

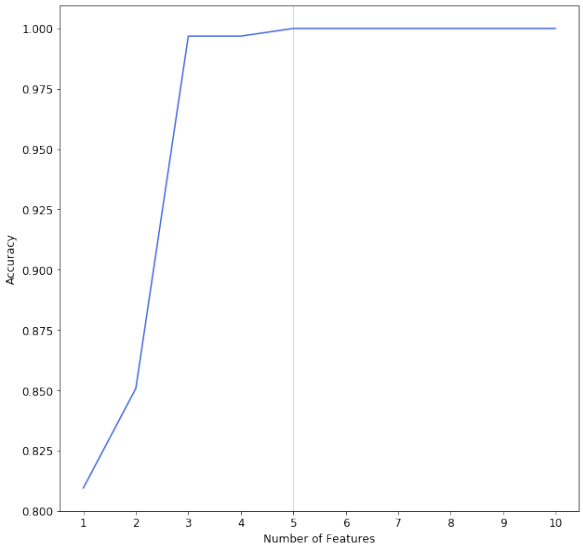

Figure 5: Accuracy depending on number of features extracted by REF.