**PROJECT DESCRIPTION**

The data set is subset the computer vision benchmark data set  “German Street Sign Recognition Benchmark by Stallkamp et al”. The original data set consists of 39,209 RGB-coloured train and 12,630 RGB-coloured test images of different sizes displaying 43 different types of German traffic signs. These images are not centred and are taken during different times of the day.  
  
A sample of this data set which consists of 10 classes and 9690 images. The images have been converted to grey-scale with pixel values ranging from 0 to 255 and were rescaled to a common size of 48\*48 pixels. Hence, each row (= feature vector) in the data set has 2305 features and represents a single image in row-vector format (2304 features) plus its associated class label.  
  
Class labels are  labelled from 0 to 9 indicating the following traffic sign categories:

|  |  |
| --- | --- |
| **Class label** | **Traffic Sign** |
| 0 | **speed limit 20** |
| 1 | **speed limit 30** |
| 2 | **speed limit 50** |
| 3 | **speed limit 60** |
| 4 | **speed limit 70** |
| 5 | **left turn** |
| 6 | **right turn** |
| 7 | **beware pedestrian crossing** |
| 8 | **beware children** |
| 9 | **beware cycle route ahead** |

There are 5 different parts covered here, where different operations were perfomed over the given data set

Tasks covered in our project:

• Part1: Data Analysis and Exploration   
• Part2: Clustering   
• Part3: Decision Trees   
• Part4: Neural Networks and CNN   
• Part5: Research question

**Part 1. Data Analysis and Bayes Nets.**

This part of the assignment assumes that you are using a training set (keep your testing set aside for now). So, all algorithms should only be evaluated on the training data set.

• Visualization and initial data exploration help to gain insights on the data attributes and guides in choosing suitable features and building appropriate ML models. Examine your data through visualization and analysis and show how this helped you learn more about your data and has guided you for further analysis. Discuss how you fixed problems like missing values, errors or outliers -if applicable.

Did you need to apply any preprocessing or normalization procedures? If so, why?

• Run Naïve Bayes Classifier on your chosen data set, and record the major metrics: accuracy, TP rate, FP rate, precision, recall, F measure, the ROC area etc. (as explained in the lectures). Make conclusions. Use cross-validation on your training set to report your finding. Alternatively, you can generate a stratified train-test split version

**Part 2. Clustering**

• For the same data set, use k-means clustering to find clusters in your data. Evaluate the accuracy of this clustering, visualize the clusters, make conclusions.  
• For top marks, try different clustering algorithms for hard and soft clustering, such as EM, GMM, hierarchical clustering or any other algorithms of your choice. Compare their performance on your dataset, make conclusions.  
• Try also to vary the number of clusters manually and then research some of the existing algorithms to compute the optimal number of clusters. How does it affect the accuracy of clustering? Make conclusions.  
• (Optional) Look up methods to determine the optimal number of clusters. For example, look up: Elbow method, the silhouette method, cluster validity and similarity measures.  
• Using your experiments as a source, explain all pros and cons of using different clustering  
algorithms on the given data set. Compare the results of Bayesian classification on the same data set.

**Part 3. Supervised Learning: Generalisation & Overfitting; Decision trees.**

• Now you will start working with the provided test data sets.  
• Use Decision trees (the J48 algorithm) on the training set, measure the accuracy. Then measure the accuracy on the training set using 10-fold cross-validation. Record all your findings and explain them. Use the major metrics: accuracy, TP rate, FP rate, precision, recall, F measure, the ROC area if needed.  
• Repeat the experiment, this time using training and testing data sets instead of the cross validation. That is, build the J48 classifier using the training data set, and test the classifier using the test data set. Note the accuracy. Answer the question: Does the decision tree generalize well to new data? How do you tell?  
• Experiment with various decision tree parameters that control the size of the tree. For example: depth of the tree, confidence threshold for pruning, splitting criteria and the minimal number of instances permissible per leaf. Make conclusions about their influence on the classifier's performance.  
• Make new training and testing sets, by moving 30% of the instances from the original training set into the testing set. Note the accuracies on the training and the testing sets  
• Make new training and testing sets, by moving 60% of the instances from the original training set into the testing set. Note the accuracies on the training and the testing sets

**Part 4. Neural Networks and Convolutional Neural Networks.**

In this part, you will use the original training and testing data sets  
• Run a Linear classifier on the training data set, with 10-fold cross validation and without, mark the accuracies. Note also its accuracy on the test set. How well does the linear classifier generalize to new data? What hypothesis can you make about this data set being linearly separable or not?  
• Run the Multilayer Perceptron, experiment with various Neural Network parameters: modify the activation functions, experiment with the number and size of its layers, vary the learning rate, epochs and momentum, and validation threshold. Analyze relative performance of the resulting Neural Networks and changing parameters, using the training and the test data. What techniques can be used  
for performing hyperparameter tuning in a systematic way?

**Part 5. Research Question – Resnet, Fully CNN, Support Vector Machines**

Think about your own research question and/or research problem that may be  
raised in relation to the given data set, and your portfolio tasks. Formulate this question/problem clearly, explain why it is of research value. The problem may be of engineering nature (e.g. how to improve automation or speed of the algorithms), or it may be of exploratory nature (e.g. something about finding interesting properties in data), -- the choice is yours.