Machine Learning Engineer Nanodegree

Capstone Proposal

Domain background:

The dataset is Fashion MNIST database with gary scale images pixel data.In the data itself each sample is classified into any one of the 10 categories.This is also comes under object detection.Because here we are classifying image as Tshirt,Trouser,Dress,Coat likewise 6 other categories.

Which is similar to MNIST database i.e handwritten digits database.Which is used in Google house number detection from images.Similarly it is also used in vehicle number detection from number plate images.

I took this project because I want to work on a dataset with multiclass classification and image data. In which I want to implement different Machine Learning models and explore them using this dataset.

This Model can be used in online clothing sites.When a person search about any of the category name it can detect from the images whether it is a Tshirt,Trouser or Coat and display the result accordingly.From the video also we can detect specific class object.

The project dataset is taken from kaggle research dataset.
https://www.kaggle.com/zalando-research/fashionmnist

Problem Statement:

In this project, By training with the given dataset I want to classify images into 10 classes. Further, we can predict the new data belongs to which class. Here classification is nothing but detects whether the given image is a

0-Tshirt/top

1 Trouser

2 Pullover

3 Dress

4 Coat

5 Sandal

6 Shirt

7 Sneaker

8 Bag

9 Ankle boot.

Datasets and Inputs:

I am using a dataset with 70000 images. Data is divided into training and testing. The training set contains 60000 samples. Testing set contains 10000 samples. The dataset contains two CSV files one is fashion-mnist_train.csv and second one is fashion-mnist_test.csv.Each CSV file consists of 785 columns. Out of 785 columns, the first column is the label which is the target for our model. Remaining 784 values in each sample of 784 pixel values. Each row is a sample image with size a 28X28. Each value is the darkness of the pixel 1 to 255.In both training and testing data number of samples for each class are equal.That is in Training dataset each class contains 6000 samples.In the testing set each class contains 1000 samples.

Solution Statement:

Classify the image given to the model accurately to the respective model.It is multilabel classification problem with many features.My idea is to implement skmultilearn models and Convolutional Neural networks on the training data and test the accuracy of the model with test data.We need to take another measure training time and prediction time for the specific model.By measuring all above metric Choose the model which is perfect for the dataset

Benchmark Models:

The base model for this project is BinaryRelevance with Gaussian Naive Bayes from sci-kit multilearn models.Based on it's accuracy score we can improve it using other models or other multilearn models.Here is the benchmark model link for the dataset.

http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/

Evaluation Metrics:

In the dataset classes are equally balanced For this project, I am considering accuracy score as an evaluation metric. Along with accuracy score, I will also compare training and prediction time for every model. I am considering time to know my model better but I am not fully depend on this time.Because accuracy is more important for my model. The model with Highest accuracy is our best model.

Project Design:

Understanding the data

Features: 784 pixels of 28X28 grayscale image. Values are the intensity of pixel from 0 to 255

Output: 10 labels(0 to 9) this is the classification of the grey scale image.

Data Preprocessing

First divided the data as features and target.Means store the first column data of the data file into output variable. Remove this column data and store it as features.I need to check and remove Nan values from dataset.To work on multilabel data I need to make different labels for each class and place one for the label if the specific sample is for the class otherwise place zero.Which can be done easily with to_catogorical function from keras.

The data is already divided into training and testing data. Now we need to split the training data into training and validation data.using train_test_split function from skleran model_selection.After we have divided the data into training and testing data.

Normalization of data is not needed in the case because each and every value is range from 0 to 255.

Best Model Selection

Fit the different multileran models from sci-kit learn with different classifiers with Training data.Calculate each and every time training time. Predict the output of the model with testing data.Calculate accuracy score for each model. Decide which one is a better model. Tune hyperparameters for that model and maximize accuracy. If it is crossed our bench models performance it is the best model.

Otherwise, we can model the best Convolutional network for the given dataset. If I want to use the data with CNN I need to reshape my data. So that I can give it as input for the model. Once the model is decided then we can run with the suitable error functions and maximize our model by running with a suitable number of epochs. Calculate accuracy scores for each case and make sure to get maximum accuracy.

http://scikit.ml/#classifiers

Testing with Outside data

Once the best model is selected with maximum accuracy we can give a new image and check its with our prediction model. If it is an RGB image with the different size we need to change it to gray image with size 28X28 for CNN model.Whereas for the multilearn model we need to change these pixels into an array.

Output

Now, Our model is working perfectly. We need to print the output with the name of class, not as number range from 0 to 9.

Summary

Summarise the results by comparing the best model with BenchMark Models.