MACHINE LEARNING ENGINEER NANODEGREE

CAPSTONE PROPOSAL

Vibhav Chaturvedy

September 14th, 2017

PROPOSAL

Domain Background

In order to deploy a project with a consistent context, taking into account elements like data quality, relevance, ease for comparing and community delibration, i decided to work over a kaggle competition:

https://www.kaggle.com/c/leaf-classification

There are estimated to be nearly half a million species of plants in the world. Classification of species has been historically problamatic and often results in duplicate identifications, Automating plant recognition might have many applications, including:

- Species population tracking and preservation
- Plant-based medicinal research
- Crop and food supply management

In the recent few years the emergence of new organised data ,high computaional power and successful researchs in machine learning has given us the ability to use a computer to predict the species of a plant .

Problem Statement

The objective of this project is to use binary leaf images and extracted features, including shape, margin and texture, to accurately identify 99 species of plants. Leaves due to their volume, prevelance, unique characteristics, are an effective means of differentiating plant species. They also provide a fun introduction to applying techniques that involve image-based features.

The project consist of deploying a deep learning model trying to find accuracy for the given dataset to classify the given image of a leaf ,using for that objective a set of variables of different types (like boolean and categorical), divided into train and test datasets.

Datasets and inputs

The dataset consists approximately 1,584 images of leaf specimens (16 samples each of 99 species) which have been converted to binary black leaves against white backgrounds. Three sets of features are also provided per image: a shape contiguous descriptor, an interior texture histogram, and a fine-scale margin histogram. For each feature, a 64-attribute vector is given per leaf sample.

Note that of the original 100 species, i have eliminated one on account of incomplete associated data in the original dataset.

File Description:

- **train.csv :** the training set
- **test.csv :** the testing set
- **sample_submission.csv :** a sample submission file in correct format
- **images :** the image files (each image is named with its corresponding id)

Data Fields:

- **id** :an anonymous id unique to an image
- **margin_1, margin_2, margin_3, ..., margin_64 :** each of the 64 attribute vectors for the margin feature
- **shape_1, shape_2, shape_3, ..., shape_64** :each of the 64 attribute vectors for the shape feature
- **texture_1, texture_2, texture_3, ..., texture_64** : each of the 64 attribute vectors for the texture feature.

Solution Statement:

As it is expected to build a model or classifier that uses the provided pre-extracted features and next creating a set of our own features. The solution will be focused on developing a neural network. For that purpose , the process will be focused over two main stages: the first one, a descriptive analysis focused on define the best strategy over the feature engineering process (like labeling, data preparation and variable descomposition) as features are already well defined in given dataset in form of 64 attribute vectors,we don't need to do any feature extraction. Labels are textual so i will encode them catgorically and second, implementing a feed forward neural network and train it to give the expected ouput.

Benchmark Model:

The main models that will be used as bechmark are those published by the community over the discussion forum of the competition. The main references that will be used in a first stage will be:

https://www.kaggle.com/vibhav8/neural-network-through-keras

https://www.kaggle.com/vibhav8/feature-extraction-from-images

In the first kernel, they have implemented a three layered neural network by using keras . They have plotted an error versus no of iterations graph to depict the decrement of error with increase in number of iterations.

In the second kernal, they have used many machine learning models like RandomForestClassifier and calculated their accuracy and log loss score to compare each other.

In relation to the performance of the model (log loss) the benchmark will be those provided for the kaggle platform.

Evalutation metrics:

As it is required in the kaggle competition, the evaluation Metric will be the multi-class logarithmic loss i.e. logloss over the dataset. For each image, it is required to submit a set of predicted probabilities(one for every species).The formula is then,

$$logloss=-1N\sum i=1N\sum j=1Myij\log(pij),$$

where N : no of images in test set
M : no of species labels
yij : is 1 if observation i is in class j otherwise 0
pij :is predicted probability that observation i belongs to class j.

Project Design:

The solution will be constructed using the following main steps:

- Deploy a descriptive analysis over the information provided, focusing to analysis about distribution of variables, correlation, volumetry, and relevance.

- Preprocessing the data using into the model by using techniques like LabelEncoder, StandardScaler

- Evaluate the variables over different machine and deep learning(like feed forward neural network) models and compare their performance using log loss method. Finally using a feed forward neural network which includes developing a layered model for Neural Networks and using softmax layer to predict a uniform probalistic distribution of outcomes. Input dimensions will be kept equal to number of given features and Error will be measured as categorical crossentropy or multiclass logloss.

- Finally ,fitting the model on the whole training data and converting the test predictions in a dataframe as depicted by sample submission.

Acknowledgement:

[1] https://www.kaggle.com/c/leaf-classification#description

[2] https://deeplearning.org

[3] For test and training dataset, https://www.kaggle.com/c/leaf-classification

[4] Texture based leaf identification research paper
http://cmp.felk.cvut.cz/~sulcmila/papers/Sulc-TR-2014-10.pdf