CS 532 Project Update2

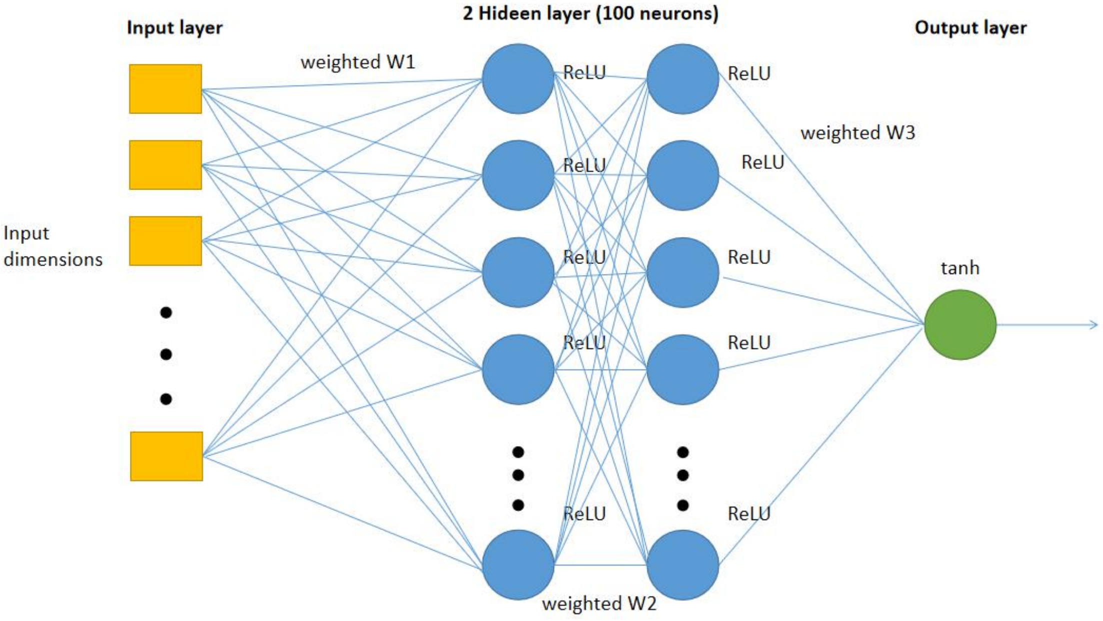
Topic: Machine learning approaches to predict forest fires

Name: Sicheng Fu

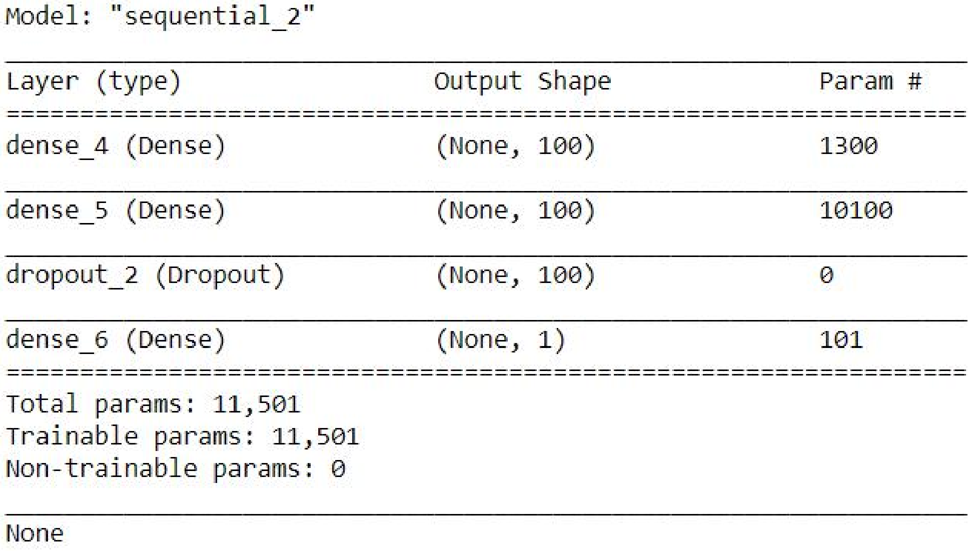
Email: sfu42@wisc.edu

This project update mainly focus on the neural network and SVM method.

Neural Network
We apply multilayer perceptron, which is a classical artificial neural network (ANN),to train our model. After normalize the data to range 0-1, we use keras package to build the mlp structure. The model involves four layers, one input layer with input dimensions equals to number of features, two hidden layers with 100 neurons, one output layer with one unit. Each hidden layers utilize ReLU as activation function, and output layer utilize tanh as activation function. The structure of model is shown below:



The model summary is shown below:



```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 100)               1300
_________________________________________________________________
dense_5 (Dense)              (None, 100)               10100
_________________________________________________________________
dropout_2 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 101
=================================================================
Total params: 11,501
Trainable params: 11,501
Non-trainable params: 0
_________________________________________________________________
None
```

To avoid overfitting problem, we can add a dropout layer in the hidden layer, randomly remove 20% of data to avoid neurons assign too much weights in one feature.

To compile the model, we choose binary cross entropy as loss function and Adam (which is an algorithm combines Momentum and RMSprop, gives better performance than SGD) as optimizer, and evaluate each epoch by accuracy. 20% of data assign as validation data, 100 times training epochs and 128 batch size is setting. After training and testing our data in this model, we can get results below:

	Accuracy	Square error
Training	60.22	1.59
Testing	54.19	1.83

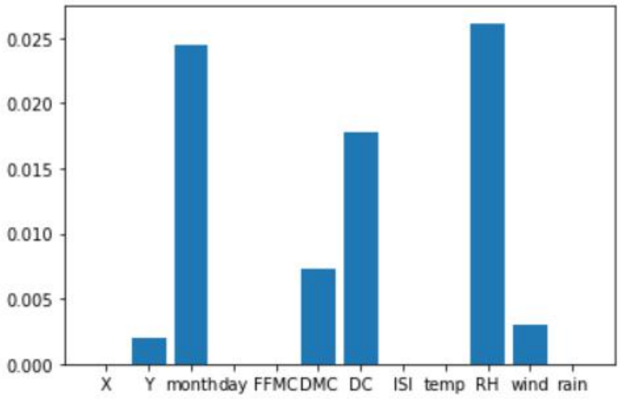
SVM

Support Vector Machine is a perfect algorithm for binary classification problems. Since the input data might be non-linearly separable, both linear support vector classification models and support vector classification models using non-linear kernels will be used. The function sklearn.svm.LinearSVC is used for linear support vector classification and the function sklearn.svm.SVC(kernel='rbf') with Radial basis function kernel is used for kernel based SVM. After training and testing our data in SVM model, we can get results below:

	Accuracy	Square error
SVC(kernel=Radial Basis Function) training	68.51	1.26
SVC(kernel=Radial Basis Function) testing	52.26	1.91
LinearSVC training	58.01	1.68
LinearSVC testing	50.98	1.96

Feature Selection

Since the prediction results did give a good performance, features selection should be considered to improve the accuracy. Only the most principal features regarding the labels should be used, which means the most weighted features should be selected. To score the importance of the features and show the correlation between the input features and the output label, the following bar chart is created using the sklearn.feature_selection.SelectKBest function.



It is easy to observe that month, DMC, DC, RH and wind have higher scores, so these features should be selected. Same method and parameters have been implemented to test the feature selected results.

	Accuracy	Square error
MLP training	61.05	1.56
MLP testing	52.26	1.91
SVC(kernel=Radial Basis Function) training	58.01	1.67
SVC(kernel=Radial Basis Function) testing	48.39	2.06

Next step:

As we can see, the prediction results did not improve after feature selection. If we look back to our data, this forest fire dataset may not easy to classification, close values for feature of two entries may cause different classification result. Since the reason of fire forest burned area is complicated, to improve the prediction accuracy, more detailed analyze and processing for original data need to be implemented before final submission. Or more classifications of burned area could be considered in next step.

Link of project page:

https://github.com/SichengIce/CS532_project