Learning System

Project Report: Classification and Regression

Ahamed Buhari

Chandraprakash Sahu

Introduction: Machine learning is categorized in two types Supervised and Unsupervised learning. Supervised Learning is an AI system where data which is labelled, which means that each data tagged with the correct label. This learning is also categorized in two types Classification and Regression.

Regression: It is called when the target variable is continuous. It consist of real values. For an instance, predicting the price of houses, predicting the power consumption .

Classification: When the target variable is discrete. It consists of categories, for example: predicting if an image contains a cat or a dog, predicting if the mail is spam or not spam.

Unsupervised Learning: It is presented with data which is unlabeled, un-categorized data and the system's algorithms act on the data without prior training. It is categorized into Clustering and Association.

In our project, we have given to study of Supervised machine learning, and in which three tasks for Regression and three for Classification.

Regression:

In Task 1, we have to see whether it is possible or not to estimate the cetane number for diesel fuel from the near infrared spectrum for the fuel using linear models of Regression.

In Task 2, It is given that This is a non-linear regression task, the output is the valve opening is connected to one heat exchanger. This opening is part of a feed-back control loop based on the temperature of the fluid that passes thru the heat exchanger. The goal with the control is to keep the temperature constant and the fluid in liquid format. So, we have to construct a model with all or part of the variables available to model the valve opening.

In Task 3, This is a non linear regression model to predict the power load for Puget Sound Power & Light Co. 24 hours in advance, at 8 in the morning, when the current day is a working day and tomorrow is a working day.

Classification Task:

In Task 4, we have to construct a model to tell if a particular set of measurements comes from a person who is normal, or suffers from being hypothyroid or hyperthyroid (i.e. 3 output categories) using classification models.

In Task 5, we have to check if a patient has a benign or malign breast cancer, based on image features from a Fine Needle Aspiration (FNA). It has been told that the diagnosis is done in the following ways. An FNA is taken from the breast mass. This material is then mounted on a microscope slide and stained to highlight the cellular nuclei. A portion of the slide in which the cells are well-differentiated is then scanned using a digital camera and a frame-grabber board.The user then isolates the individual nuclei using an image processing software. When all (or most) of the nuclei have been isolated, values for each of ten characteristics of each nuclei are computed, measuring size, shape and texture. The mean, standard error and extreme values of these features are computed, resulting in a total of 30 nuclear features for each sample.

In Task 6, we have to tell if a patient suffers from Transmural Ischemia (TI) or not, based on the signal from a 12 channel electrocardiogram (ECG). The 12 ECG channels are called V1, V2, V3, V4, V5, V6, aVL, I, -aVR, II, aVF, and III.There are 300 observations: 150 control subjects and 150 subjects that suffers from TI.

2. STATE-OF-THE-ART:

Regression:

For the purpose of accurate and reliable predictions of a cetane number for diesel fuel from the near infrared spectrum for the fuel using linear models of Regression. Based on the validation results of the developed regression models on the testing data set, the performance of RF regression in predicting the cetane number for diesel fuel data was found more accurate than KNN model. [1].

In order to calculate the gas compressibility factor (z-factor), truncated regularized KNN algorithm is used. The natural gas compressibility factor (z) is one of the critical parameters in the computations used for the upstream and downstream zones of petroleum/chemical industries. The KNN predicts the z-factor by building a nonlinear regression model in terms of the pressure and temperature. It is also observed that KNN is much more computationally efficient than the support vector regression (SVR) method, while both methods provide an accurate way for calculating the z-factor, but KNN is a time efficient method for large data sets. [2]

For predicting the power load for Puget Sound Power & Light Co , SVR method based on forecasting models is used. SVR forecasting model is compared with the other seven traditional forecasting models. The accuracy of the SVR model in forecasting power load is much better than traditional calculated methods; The SVR model proposed in this paper is suitable for real-time calculation, to expend the application and improve its efficiency. [3].

Classification:

In this corresponding paper, maximization estimation is done in order to get more accuracy. Better results are obtained by increasing the intensity class in the estimation. Usual shape features can't be used for this purpose because it considers the entire image for feature extraction and classification. Also by using Hough transform normal and abnormal classes are effectively classified. Use of more intensity features like mean, variance and entropy can improve the results. By having ANN model, we obtained the accuracy range of 97% which is higher when compared with other classifier like KNN, it has only 93% of accuracy. [4]

The paper focuses on breast cancer diagnosis by using ML algorithms. To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed four main algorithms: SVM, NB, k-NN and C4.5 on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy of SVM reaches and accuracy of 97.13% and outperforms, therefore, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate. [5]

This paper deals with ultrasonic classification of Electrocardiogram. It aims to identify if a patient suffers from Transmural Ischemia (TI) or not, based on the signal from a 12 channel electrocardiogram (ECG). Significant features have been used to diagnose in all models were hypoechoic, irregular margin, and microcalcification.In conclusion, KNN has proven its efficiency than DT in prediction of suffers from Transmural Ischemia (TI) and diagnosis and achieves the best performance in terms of precision and low error rate.[6]

3. Methodology :

We can define the machine learning workflow in below steps:

- Gathering data : Dataset is loaded and changed into the appropriate format i.e. by using transpose if needed.

- Data pre-processing: This is most important process i.e. cleaning the raw data i.e. whenever the data is gathered from different sources, it is collected in a raw format and this data is not feasible for the analysis, so certain steps are executed to convert the data into a small clean data and that can be used to train the model.

- Researching the model that will be best for the type of data: The main objective is to train the best performing model possible, using the pre-processed data.

- Training and testing the model : For training a model, we initially splitted the model into 3 three sections which are ‘Training data’ ,‘Validation data’ and ‘Testing data**.** train the classifier using train the classifier using ‘training data set**’,** tune the parameters using ‘validation set**’** and then test the performance of your classifier on unseen ‘test data set’, tune the parameters using ‘validation set’ and then test the performance of your classifier on unseen ‘test data set’.So, during training the classifier only the training and/or validation set is available. The test set will only be available during testing the classifier.

- Evaluation : As model evaluation is an integral part of the model development process. It helps to find the best model that represent our data and how well the chosen model will work. For improving the model, we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

  Below steps have been performed in all the six tasks(Regression and Classification).

- Input data is split into X_train, y_train, X_test and y_test using train_test_split.
- Model is selected and training of data is performed.
- Cross Validation is performed to evaluate the model used earlier.
- Mean Square Error for regression and Accuracy for classification is checked for the model.
- Hyper-Parameter Tuning is performed to estimate the best parameter and optimize the results using either gridsearchcv or randomsearchcv.
- Steps 1-6 are performed for different models of regression/classification. The best model with good results is used to predict the value for given test data (corresponding to the dataset).

4. Data :

Regression:

Dataset1: is provided as cnDieselTrain.mat in task 1.The dataset contains three matrices: cn-TrainX (401 × 133), cnTrainY (1 × 133), and cnTestX (401 × 112). The spectrum has 401 channels (features) and 245 observations. The first matrix (cnTrainX) contains the IR-spectrum for each sample, one column per sample. The second matrix (cnTrainY) contains the output value cetane number for each diesel fuel. The third matrix (cnTestX) contains the input (IR-spectra) for each sample in the test data set.

Dataset2: is provided as ChemTrainNew.mat in task 2. The dataset consist of three matrices: XtrainDS (4466x65), YtrainDS (4466x1) and XtestDS (2971x65). The input matrix (XtrainDS) contains all variables to the process. The first column is time, which is not considered as a feature and the output matrix (YtrainDS) contain the valve opening.

The third matrix (XtestDS) is test data consisting of all inputs from a time period that follow the 19 month training data. The number of features is 65 and datapoints are 7437.

Dataset3: is provided as PowerTrainData.mat in task 3. This dataset contains powerTrainInput (15 × 844), powerTrainOutput (1 × 844), powerTrainDate (1 × 844), and powerTestInput (15 × 115). We have taken the transpose of the matrices and changed them into the following shape, powerTrainInput (844 x 15), powerTrainOutput (844 x 1), powerTrainDate (844 x 1), and powerTestInput (115 x 15). The number of features in the dataset is 15 and total number of observations is 959.

Classification:

Dataset4: is provided as thyroidTrain.mat in task 4. This dataset contains the matrices trainThyroidInput (5000 × 21), trainThyroidOutput (5000 × 3), and testThyroidInput (2200 × 21). The first matrix, trainThyroidInput, contains the input patterns for the training data. The second matrix, trainThyroidOutput, contains the outputs coded in a "1-out-of-3" fashion (i.e. as a one-hot-vector). That is, the outputs are coded as (1,0,0), (0,1,0), or (0,0,1). The third matrix, testThyroidInput, contains the inputs for the test data. There are 7200 observations representing patients. The given 5000 of these, and 2200 are withheld for testing.

Dataset5: is provided as cancerWTrain.mat in task 5. This dataset contains the matrices cancerTrainX (30×400), cancerTrainY (1×400), and cancerTestX (30×169). There are 569 observations, of which 400 are provided to us for training.The given input data is in bad shape and in order to use it for training the model we will reshape it by transposing the input matrix.

Dataset6: is provided as ECGITtrain.mat in task 6. This dataset contains the matrices inputECGITtrain (200 × 312), outputECGITtrain (200 × 1, i.e. a column vector), and inputECGITtest (100 × 312). From inputECGITtrain matrix we have 312 features and we extracted important features, i.e. features 19–26 for each channel. These correspond to inputs 19–26, 45–52, and so on. After that we normalize the data and train the model using kNN and DT methods.

5. RESULTS AND INTERPRETATION:

Regression: Task 1

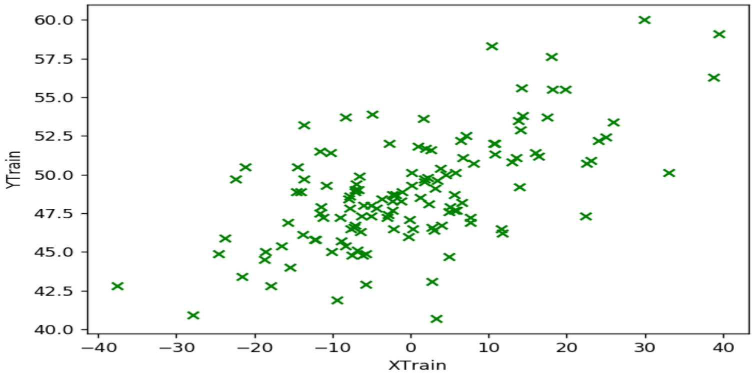


Fig.1. Scatter Plot of Input train and output train Data.

- We have used Linear Regression algorithms n this dataset. To reduce the number of features we have used Principal Component Analysis(PCA) .With PCA we have selected 15,30, 40 & 60 features which are the highest varied attributes. We got variance 0.9937 after reducing the PCA to 30. Using GridSearchCV we have performed Hyper Parameter tuning to find the best parameters for the model.

- Best Hyper Parameters with GridSearchCV for Linear Regression: {'alpha': 1.0}

Training				
PCA	15	30	40	60
mean_squared_error	2.59440	1.86706	1.32946	0.60930
mean_absolute_error	1.32820	1.08249	0.91677	0.57430
root_mean_squared_error	1.61071	1.36640	1.15302	0.78058

Validation				
PCA	15	30	40	60
mean_squared_error	6.06835	8.57954	9.58068	30.6511
mean_absolute_error	1.89222	2.17376	2.19437	3.98787
root_mean_squared_error	2.46340	2.92908	3.09526	5.53634

We can see from the above table, as increases number of feature while PCA, the error in training data are decreases but in validation data, error increases. Which means our model leads to over fitting with the increase of features.

Task 2:

- We have used SVR and KNN Regression algorithms for this dataset. To reduce the number of features we have used Principal Component Analysis(PCA) .With PCA we have selected 10 features with the high variance of data. Then we have splitted the training data into training and Validation data, trained our model using training data and performed Cross Validation. Using GridSearchCV we have found the best Hyper parameters for the models.
- Best Hyper Parameters with GridSearchCV for SVM: {'C': 10,'gamma': 0.0001}
- Best Hyper Parameters with GridSearchCV for KNN: {'n_neighbors': 2, 'weights':'distance'}
- After training the models, and finding the best parameters we found that the error rate using SVR method was low in contrast to KNN. So, we selected SVR method to predict the outputs for given test set.

Model	**Error-Training Data**	**Error - Cross Validation**
SV Regression	6.32	2.1976

KNN Regression	7.14	3.2102

Task 3:

These inputs are the result of quite a lot of variable selection so you tried using all the variables but the error were quite higher. In this we have used feature selection using Variable threshold below 0.05, but still could not reduce any features. We have used KNN and Decision Tree algorithms for this dataset. Using GridSerachCV we have found the best parameters for both the KNN and Decision Tree algorithms.

Best Hyper Parameters with GridSearchCV with DT: {'criterion': 'gini', 'max_depth': 8, 'max_features': 3, 'min_samples_leaf': 5, 'min_samples_split': 8}

Best Hyper Parameters with GridSearchCV for KNN: {'n_neighbors': 4, 'weights':

'Distance'}

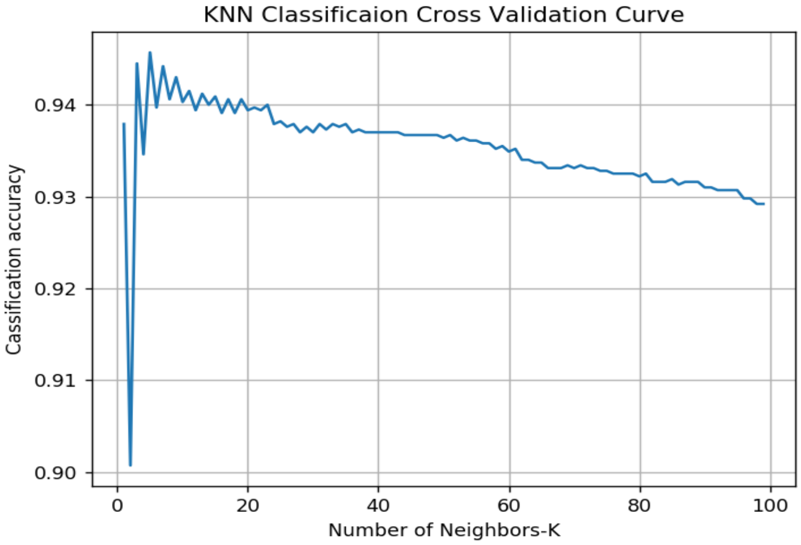
K	**RMS Error**
1	133.2205
2	119.5821
3	106.2412
4	109.5667

Model	**RMS Error-Training Data**	**RMS Error - Cross Validation**
DT Regression	1586.331	137.724534
KNN Regression	1028.007	108.695777

After training with the models, and finding the best parameters using GridSerachCV and after Cross Vaidation we found that the error rate using KNN method has bit low error in contrast to DT. So, we selected KNN method to predict the outputs for given test set.

Task 4: In this we have selected the 15 components/attributes using PCA which has highest varied data. We have used KNN Classifier Algorithm for this dataset using train_test_split and the cross validation techniques. In comparing the KNN model with the train_test_split and Cross Validation technique we noticed that Cross Validation technique has the high accuracy.

K Value	Classification Accuracy
1	0.937902
2	0.900701
3	0.944503
4	0.934600



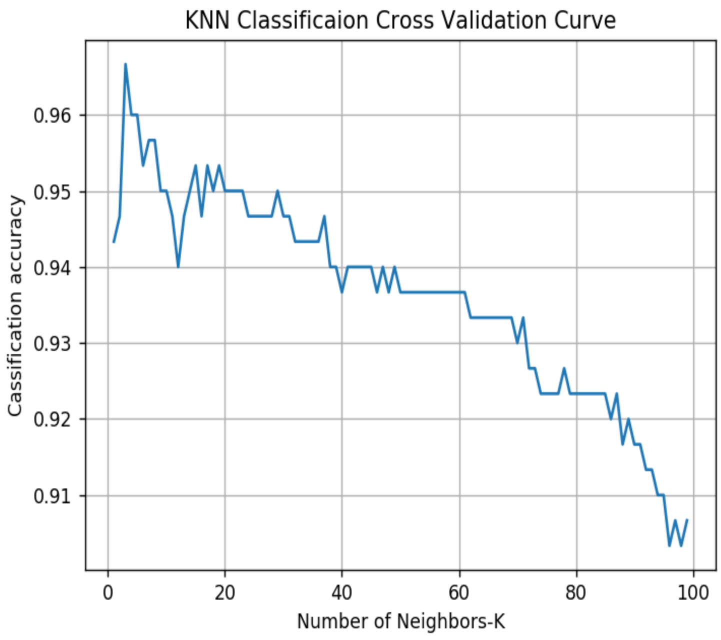
Task 5:

In this we have used KNN and SVM Classifier algorithms and Cross validation technique on the training data.We have normalized the data since the data is in different scales. We have found the best parameters for these models using GridSearchCV.

Best Hyper Parameters for KNN: {'n_neighbors': 3, 'weights': 'uniform'}.

Best Hyper Parameters for SVM: {'C': 1, 'degree': 3, 'gamma': 'auto', 'kernel': 'linear'}

Model	**Accuracy-Training data**	**Accuracy-Cross-Validation**
KNN	**0.923**	**0.976**
SVM	**0.9412**	**0.982**



After training the data with both SVM and KNN, SVM gives the best accuracy so we have used the SVM classifier to predict the given test data.

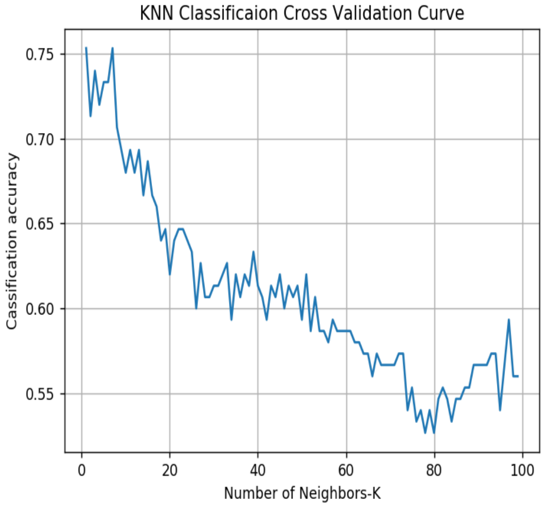
Task 6:

In this we have used Logistic Regression and KNN Classifier methods. We have found the best parameters for these models using GridSearchCV.

Best Hyper Parameters for KNN: {'n_neighbors': 3, 'weights': 'uniform'}.

Best Hyper Parameters for Logistic Regression : {'C': 0.01}

Model	**Accuracy-Training data**	**Accuracy-CV**
Logistic Regression	**0.72**	**0.701**
KNN	**0.73**	**0.743**



So after training the data with both model Logistic Regression and KNN, KNN gives the best accuracy so we have used the KNN classifier to predict the given test data.

6. Conclusion :

We have used different algorithms models for predicting the Test data. In the observation found that, best model depends on the available data , Hyper parameter tuning, feature selection of the given data. Below are the results for both Classification and Regression datasets.

In first Regression model, Linear Regression is the best model for given dataset.

In second Regression Model, SVR method is the best model for given dataset.

In Third Regression Model, KNN method is the best model for given dataset.

In First Classification Model, ANN method is the best model for given dataset.

In Second Classification Model, SVM method is the best model for given dataset.

In Third Classification Model, KNN method is the best model for given dataset.

References:

[1] https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

[2] Asri, H., Mousannif, H., Al Moatassime, H. and Noel, T., 2016. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, pp.1064-1069.

…………………………………..XXXX…………………………………………