Learning Systems Project Report

Neenu Ajilkumar
Swathi Priya Dudipalli

1. Introduction

Algorithms learn from labelled data in Supervised Learning. Upon understanding the data, the algorithm determines which label should be given to new pattern - based data and associates patterns with unlabelled new data.

Supervised Learning can be divided into 2 categories.

1. Classification
2. Regression

In our project, we have given to study three problems 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. 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:
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). 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.

In this project, We used PCA, KNN Regression, KNN Classifier, SVM Classifier, SV Regression, Linear regression, Ridge regression, Logistic Regression to construct the model, and at last, I using Hyper Parameter tuning with Cross Validation and GridSreachCV method to find the best parameters to improve the result and performance of the model.

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 Ridge 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

3.1.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.

3.2.PCA:

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and in each succeeding component.

3.3. Cross-Validation

This means using a test data set, which is a subset of the available data (typically 25-35%) that is removed before any training is done, and which is not used again until all training is done. The performance on this test data will be an unbiased estimate of the generalization error, provided that the data has not been used in any way during the modeling process. If it has been used, e.g. for model validation when selecting hyper parameter values, then it will be a biased estimate.

3.4 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.

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

- Data has been normalized in order to be in the same scales.
- Reduced the number of components using PCA, and retained highest varied data features.
- 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 Cross Validation techniques.

- 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.

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 Logistic Regression methods.

5. RESULTS AND INTERPRETATION:

Regression: <u>Task 1</u>

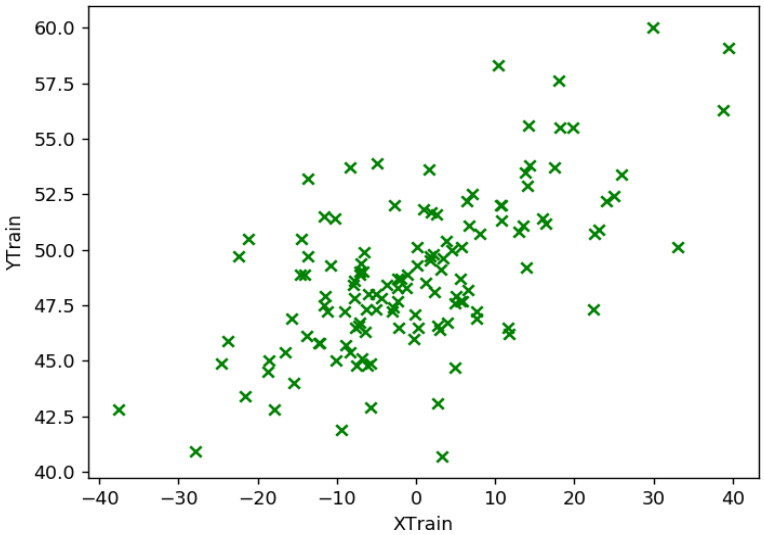


Fig.1. Scatter Plot of Input train and output train Data.

- We have used Linear Regression and KNN Regression algorithms on this dataset.
- To reduce the number of features we have used Principal Component Analysis(PCA) .
- With PCA we have selected 20 features which are the highest varied attributes.

```
Variance retained corresponds to the data after reducing with PCA:
0.9937700162595301
```

- Then we have splitted the training data into training and Validation data, trained our model using the Regression Model.(Linear Regression and KNN)
- Using GridSearchCV we have performed Hyper Parameter tuning to find the best parameters for the model.

K-Value	RMS Error
1	2.858441161
2	2.559787752
3	2.43281553
4	2.428954925
5	2.411119052
6	2.387883157
7	2.361989378
8	2.305807216
9	2.34043019
10	2.375851586

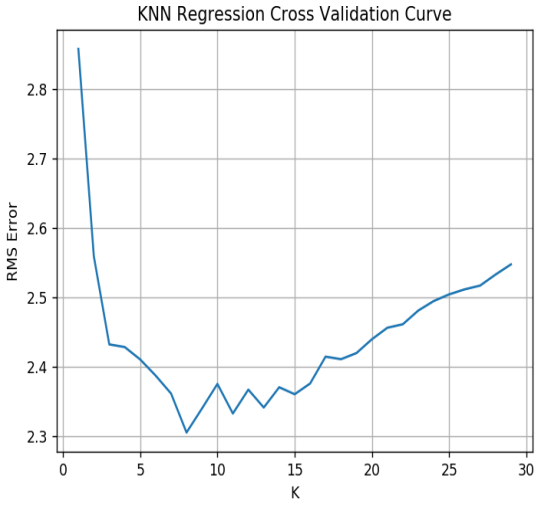


Fig.2.Cross Validation Curve for KNN

- Then we have performed the 10 fold cross validation and taking the average of the 'root_mean_squared_error', which is low on the training data.

- The below table shows the best parameters and the different values for errors on Training data and Test data.
- Best Hyper Parameters with GridSearchCV for Linear Regression: {'alpha': 1.0}
- With KNN, we have found the best parameters as below.
- Best Hyper Parameters with GridSearchCV for KNN: {'n_neighbors': 8, 'weights': 'uniform'}
- After training the models, and finding the best parameters and then performing the cross validation we have selected Linear Regression model as the best and using the same we have predicted the final output of the test data.

Model	Error-Training Data	Error - Cross Validation
Linear Regression	2.075807	0.039197654
KNN Regression	2.305807	0.040245809

Task 2:

- In this dataset there are 4466 samples of these are for training and validation, and 2971 samples are kept for testing.
- 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	5.221539	1.039197654
KNN Regression	7.779866	3.040245809

Task 3: In this dataset there are 844 samples of these are for training and validation, and 115 samples are kept for testing.

- 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.

K	RMS Error
1	131.172205
2	113.285582
3	109.060242
4	108.695777
5	110.839594
6	113.325271

- Using GridSerachCV we have found the best parameters for both the KNN and Decision Tree algorithms.

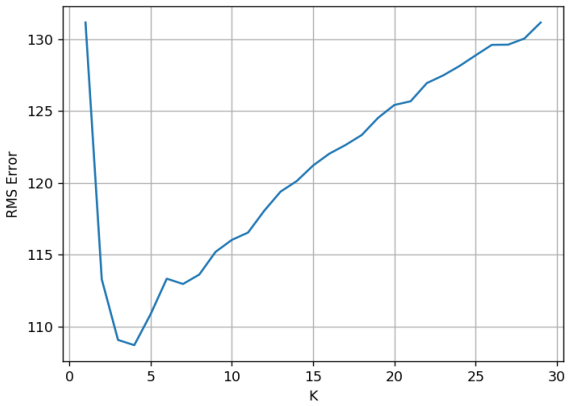


Fig.3.Cross Validation Curve for KNN.

- 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'}

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 Validation 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.

Classification:

Task 4:

- The data for this project are 7200 observations representing patients. 5000 of them are training data, and the other 2200 sample are testing data, and there are 21 variables.

- 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.

KNN Model	Train_test_split data	Cross Validation
Accuracy	0.92	0.94

- On 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
5	0.945702
6	0.939698

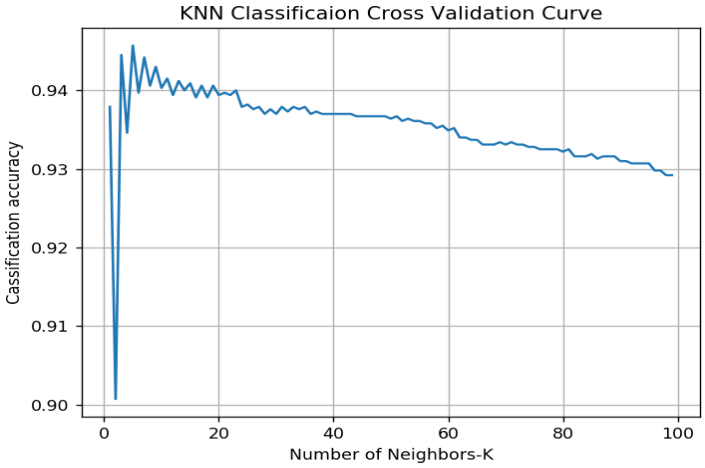


Fig 4: KNN Cross Validation Curve

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	0.92	0.97
SVM	0.94	0.98

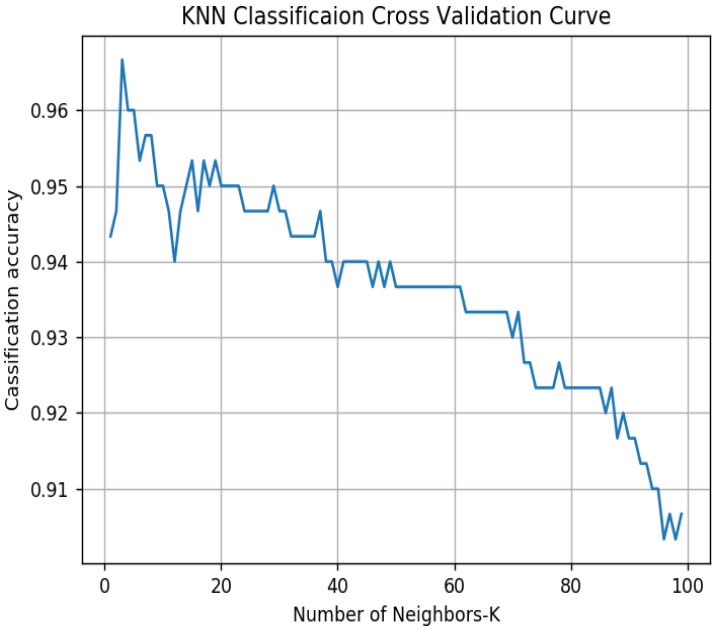


Fig 5: Cross Validation Curve KNN

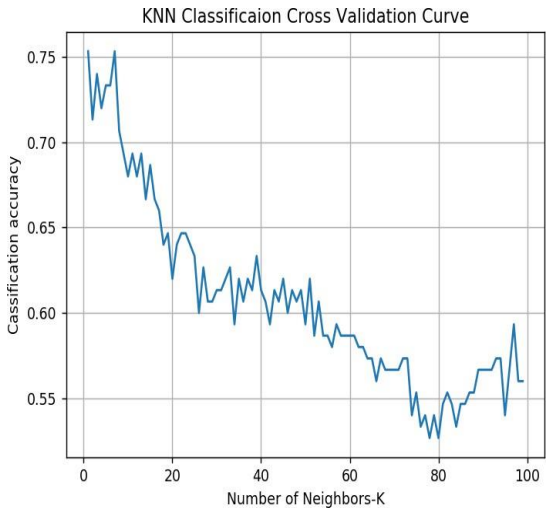
After training the data with both SVM and KNN, SVM gives the best accuracy so we have used the SVM classifier to predict the given test data.

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}

Fig 5: Cross Validation Curve KNN

Model	Accuracy-Training data	Accuracy-CV
Logistic Regression	0.7	0.72
KNN	0.73	0.76



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

6. Discussion

- From this Project, we can get conclusion from the result that the Cross-validation technique plays an important role in assessing the model performance.
- In our opinion, finding best model depends on the available data , Hyper parameter tuning, feature selection of the given data.
- In Regression,
  For First problem, Linear Regression ; For Second problem, SVR Model ; For Third problem, KNN Models are the best model for given dataset.
- In Classification,
  For First problem, KNN Model; For Second problem, SVM Model ; For Third problem, KNN Model are the best model for given dataset.
- In other research work researchers have used different number of techniques to train better model for more reliable outputs. They are as follows:
  1. Outlier Detection which merges feature extraction, clustering analysis and generalized extreme deviate to remove abnormal consumption of power during certain days.
  2. Recursive feature elimination which removes recurring features to optimise the training algorithms.
  3. Genetic Algorithms can be used to optimize the weights of predictive models for better estimation.
  4. Grey prediction with rolling algorithm can be used because of its high accuracy, low computational power and applicability to limited data situations.
  5. ARIMA(Autoregressive integrated moving average) can be used for forecasting future energy consumption.
- There is more scope of improvement as the research increases further and we will need to train with more precise and robust predictive models.

References

[1] 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.

[2] D. Selvathi., V.S. Sharnitha., 2011. Thyroid classification and segmentation in ultrasound images using machine learning algorithms.

[3] Machine Learning Methods in Electrocardiography Classification Iryna Mykoliuk1 , Daniel Jancarczyk1 , Mikolaj Karpinski1 , Viktor Kifer2

[4] Principal component analysis -http://en.wikipedia.org/wiki/Principal_component_analysis

[5] Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning Sebastian Raschka - https://arxiv.org/abs/1811.12808