Report

Group ID: B2
Lab Group No.: 03

Members:
Anindya Brata Choudhury (14.02.04.098)
Mehjabin Nowshin (14.02.04.101)

1 Problem Description

Sea level is an average level of the surface of one or more of Earth's oceans from which heights such as elevations may be measured. Sea levels can be affected by many factors and are known to have varied greatly over geological time scales. The careful measurement of variations in MSL can offer insights into ongoing climate change, and sea level rise has been widely quoted as evidence of ongoing global warming.
In this problem, we have a data-set containing some environmental data spanning over more than a hundred years. Our aim is to predict the sea level based on the gathered data.

2 Data-set Description

The data-set consists of 4 columns among which 3 are features and the last column is the target class. The features are various weather and environmental data. Each sea level is predicted on the basis of the environmental data from the corresponding year 1.

3 Description of the Models

We have used 5 models in this problem to see which one gives a better result. Firstly, the target is separated from the features. Then it is split for training and testing. $\frac{4}{5}^{th}$ data is used for training and the rest $\frac{1}{5}^{th}$ is used for testing purpose. The models are:

Table 1: **Data Definition**

Variable	**Definition**
Carbon Concentration	Concentration of Carbon in parts per million by volume
Temperature	Temperature in degrees fahrenheit
Land-Ocean Temperature Anomaly	Land-Ocean Temperature Anomaly degrees celcius
Adjusted Sea Level	Adjusted Sea Level in inchs

1. Random Forest Regressor
2. Extra Trees Regressor
3. Decision Tree Regressor
4. Support Vector Regression (SVR)
5. Ada Boost Regressor

3.1 Random Forest Regressor

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.
Library: sklearn.ensemble.RandomForestRegressor
A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

3.2 Extra Trees Regressor

Extra trees regressors are a lot like Random Forest Regressors. With respect to random forests, the method drops the idea of using bootstrap copies of the learning sample, and instead of trying to find an optimal cut-point for each one of the K randomly chosen features at each node, it selects a cut-point at random.
This idea is rather productive in the context of many problems characterized by a large number of numerical features varying more or less continuously: it leads often to increased accuracy thanks to its smoothing and at the same time significantly reduces computational burdens linked to the determination of optimal cut-points in standard trees and in random forests.
Library: sklearn.ensemble.ExtraTreesRegressor
This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

3.3 Decision Tree Regressor

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.
Library: sklearn.tree.DecisionTreeRegressor
A decision tree regressor.
Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.

3.4 Support Vector Regression (SVR)

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.
Library: sklearn.svm.SVR
The model produced by support vector classification (as described above) depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by Support Vector Regression depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

3.5 Ada Boost Regressor

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.
Every learning algorithm tends to suit some problem types better than others, and typically has many different parameters and configurations to adjust before it achieves optimal performance on a dataset, AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier. When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder-to-classify examples.
Library: sklearn.ensemble.AdaBoostRegressor

Table 2: Comparison of performance metrics

	Exp. variance	Mean Sq. Error	Mean Sq. log Error	R^2
Random Forest Regressor	0.986	0.069	0.005	0.986
Extra Trees Regressor	0.99	0.042	0.004	0.99
Decision Tree Regressor	0.983	0.083	0.008	0.983
Support Vector Machine	0.832	1.037	0.026	0.807
Ada Boost Regressor	0.986	0.069	0.005	0.986

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.

4 Comparison of the performance scores

For testing the performance we have chosen 4 performance matrices which are - explained variance score, mean squared error, mean squared log error, and R^2 score. The following scores were found: 2

5 Discussion

After trying out five different regression algorithms to try and predict sea levels from our collected data we observe that Extra Trees Regressor achieves the best metrics with an explained variance score of nearly 99%. On the other hand, Support Vector Regressor of the Support Vector Machine library performs the worst with explained variance score of 82%. The rest of the approaches all gave values around 98% for explained variance score.