Report

Group ID: B1
Lab Group No.: 02

Members:
Md. Sabbir Hassan (14.02.04.058)
Nowshin Nawar Arony (14.02.04.067)

1 Problem Description

Different types of food contain various nutritional elements. Some foods contain more calories than the other. Some are rich in Vitamin A or C. On the contrary, few items contain lots of fat.
In this problem, we have a data-set containing some nutritional facts of different foods. We have 4 classes of food type : Fruit, Vegetable, Meat and Dairy. Our aim is to predict the class of food based on their nutritional elements.

2 Dataset Description

The data-set consists of 9 columns among which 8 are features and the last column is the target class. The features are nutritional elements of different types of food. Each class of the food is predicted on the basis of their nutritional facts.

Data Dictionary:

Variable	Definition
Calories	The amount of calorie
Total Fat	Total amount of fat content
Na	Amount of Sodium
K	Amount of Potassium
Protein	Amount of protein
Vitamin A	Percent Daily Value of Vitamin A
Vitamin C	Percent Daily Value of Vitamin C
Iron	Percent Daily Value of Iron
Item	Type of the Food item

3 Description of the Models

We have used 5 models in this problem to see which one gives the better accuracy. Score for each model is tested individually. The data-set is first standardized by using scalar transform so that no feature has more priority over the other. Then it is split for training and testing. 75% data is used for training and the rest 25% is used for testing purpose.

3.1 SGD (Stochastic Gradient Descent) Classifier

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression. SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features. The advantages and disadvantage of Stochastic Gradient Descent are as follows:

Advantages:

1. Efficiency.
2. Ease of implementation (lots of opportunities for code tuning).

Disadvantages:

1. SGD requires a number of hyper parameters such as the regularization parameter and the number of iterations.
2. SGD is sensitive to feature scaling.

3.2 Decision Tree Classifier

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.
A decision tree has 2 kinds of nodes:

1. Each leaf node has a class label, determined by majority vote of training examples reaching that leaf.
2. Each internal node is a question on features. It branches out according to the answers.

3.3 Random forest Classifier

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 over fitting to their training set.

- Random forest classifier will handle the missing values.
- When we have more trees in the forest, random forest classifier wont over fit the model.
- In the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

```
RandomForestClassifier(n_estimators=200)
```

n_estimators: integer, optional (default = 10)
The number of trees in the forest.

3.4 AdaBoost Classifier

Ada-boost, like Random Forest Classifier is another ensemble classifier. Ensemble classifier are made up of multiple classifier algorithms and whose output is combined result of output of those classifier algorithms.
Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier.

1. Retrains the algorithm iteratively by choosing the training set based on accuracy of previous training.
2. The weight-age of each trained classifier at any iteration depends on the accuracy achieved.

3.5 SVM Classifier

Support vector machines (SVMs) are a set of supervised learning methods used for classification and regression. The advantages and disadvantages of support vector machines are given below:

Advantages:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

- Versatile: Different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

Disadvantages:

- If the number of features are much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

C-Support Vector Classification:

The implementation is based on libsvm. The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples.

`svm.SVC( kernel=rbf )`

kernel : string, optional (default =rbf)

Specifies the kernel type to be used in the algorithm. It must be one of linear, poly, rbf, sigmoid, precomputed or a callable. If none is given, rbf will be used.

4 Comparison of the performance scores

For testing the performance we have chosen 4 performance matrices which are - precision, recall, F1 and accuracy score. The comparison of the performance scores for each model is shown below:

	Precision Score	Recall Score	F1_Score	Accuracy Score For Training	Accuracy Score For Testing
SGD Classifier	0.8634	0.87	0.8867	0.9175	0.87
DecisionTree Classifier	0.9092	0.87	0.8627	1.0	0.87
RandomForest Classifier	0.9258	0.92	0.9202	1.0	0.92
Adaboost Classifier	0.5636	0.61	0.6685	0.7275	0.61
SVM Classifier	0.8451	0.79	0.7910	0.91	0.79

5 Discussion

Amongst the 5 models that we have used in this project, RandomForest Classifier gives the best result. RandomForest Classifier works with randomly selected 2 or 3 the features and fit them in the model which results better prediction. There is no guarantee that standardization will improve the classification performance. Standardization will change the distribution of data. Ensemble learning helps improve machine learning results by combining several models.