Jonathan Pedoeem REU UNCW 2018

Professor Chen 06/01/2018-Week 2

**Assessing Regression and Classification Techniques on the MORPH-II Database**

1. **Introduction**

The MORPH-II (Craniofacial Longitudinal Morphological Face Database) is a database of police mug-shots collected over eight years by the Face Aging Group at University of North Carolina Wilmington [1]. The academic use database contains over 55,000 unique images of 13,000 different individuals [2]. The database also contains other pieces of meta-data such as race, gender, date of birth, and date of acquisition [1].

This size of this database, the style of photos (mug-shots that all follow a very similar format), and the meta-data that comes along with it, makes it a prime choice for data mining. The MORPH-II database can be used to train machine learning models to predict the age and gender of a new image with high accuracy.

Applications of such a model are far reaching. It can be used to check if a customer is over the legal drinking age, for example. In addition such a model can be the stepping stone for more complicated machine learning models like facial recognition.

Regression techniques are used and compared to model numerical data such as age, while classification techniques are compared to model categorical data such as gender. Five Fold Cross Validation (5- Fold CV) and Leave One Out Cross Validation (LOOCV) were used to predict accuracy of these models in a general setting.

1. **Methodology**

The methods used to train regression and classification models can be split into non-tree and tree methods. The non-tree methods include various methods to create a fit in the case of regression, or a boundary in the case of classification. Tree-methods “involve *stratifying* or *segmenting* the predictor space into a number of simple regions” [3]. Tree-methods build a tree where at each node is a terminal node where the value held at that leaf is the prediction, or the node contains a decision point where the branch corresponding to the correct value for the entry is selected. The non-tree methods include Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (K-NN). The tree methods include, Decision Trees, Bootstrapping Aggregation (Bagging), Random Forests, and Boosting.

Classification methods are tasked with taking either continuous or discrete inputs and outputting a discrete result or a result that can be easily interpreted to discrete values. Subsequently, standard regression methods are not the best choice. Standard polynomial regression maps an input to a continuous set of outputs. At times, it is not clear how to take that continuous output and map it to a discrete boundary. Logistic regression addresses this issue. The logistic function (seen here in the two-dimensional form)

Allows for the probability of the output being true (Y=1) given an input X. The probability of an output being false is . While the output of the logistic function is still continuous, the fact that it outputs a probability makes it convenient for the purposes of classification. Logistic Regression is the method to find the best choices for the coefficients of the logistic function.

There are several instances where Logistic Regression is not a good choice for categorical data. If the classes are well separated, Logistic Regression is highly unstable [3]. In the cases where Logistic Regression is not a good choice, LDA offers an alternative. LDA differs from Logistic Regression as it models and then uses Bayes’ theorem to predict .

Where is the prior probability that an observation comes from the k*­*th class. LDA predicts a normal density function for each class k, and then creates a boundary between them. QDA is very similar to LDA, however, it doesn’t assume all the classes share the same covariance matrix.

K-NN offers a different approach to classification. K-NN predicts the category by taking the majority of the K- nearest entries to the entry that needs to be classified in the space of predictors.

The goal of a Decision Tree is to split the predictor space into J p-dimensional rectangular boxes. A regression Decision Tree minimizes the Residual Sum Squares (RSS).

Each element in a region has the same prediction . The prediction is determined average when dealing with regression problems. For a classification tree the Classification Error rate is minimized.

Where is the portion of observations that are in the mth region from the kth class.

Decision trees truly take a continuous predictor space and map it to a discrete output space. Trees offer an advantage over linear models as the regions that are produced are rectangular which can not be easily achieved with a linear model.

Boosted Aggregation, or Bagging, offers improvement to a simple decision tree. One issue with a simple decision tree is that it has high variance. If the training set was sampled a few different ways from the dataset the resulting trees are likely to be much different. As the goal is to have the best model, having high variance makes it ambiguous as to which one is the best fit. A solution to this is to use to bootstrap to create many different trees and then to use the average of the created trees. Bootstrap is the idea of sampling with replacement from the test set. Each bootstrapped test set is used to create a Decision Tree, which are then averaged together.

While Bagging does offer improvements over a simple Decision Tree, it has some of its own drawbacks. First it is computationally intensive as the number of trees can be over 500 which would make its run time 500 times longer than a Decision Tree. More importantly, each of the trees are highly correlated. As a result, if there is a dominant predictor most trees in the Bagging method will have the same splits, especially at the beginning of the tree. Using trees that are uncorrelated will lead to less variance, which was the goal of Bagging [3]. Random Forest addresses this issue by only selecting m of the p predictors. In classification usually a choice of m such that is a good one, while in regression a choice of is better. This limit means that most of the predictors will not be used, resulting in each tree to be not as correlated to each other as in Bagging.

Boosting is yet another tree method. At its core Boosting differs from Bagging and Random Forest, in the fact that it grows trees sequentially, instead of growing each tree independently and averaging them out. Boosting learns at a much slower rate but allows for the learning to be more directed at areas where the tree does not do well.

Predicting the accuracy of a trained model int a real-world setting is not a simple task. Given that in most situations the dataset available is small and expensive, splitting up the database into a training set and a testing set can be both expensive and inaccurate. Two cross validation techniques are K-Fold Cross Validation and Leave One Out Cross Validation (LOOCV). K-Fold CV is a CV method where the dataset is split into K groups. The model is trained on of the dataset and tested on elements. This is done for each of the k fold of the dataset. The predicted accuracy would be the average of each of the folds’ accuracies. This allows for a prediction without keeping a big portion of the dataset only for testing purposes. LOOCV is another method that offers a more realistic prediction. LOOCV is a method where the model is trained on all entries but one and then tested on that entry. This allows both for most of the entries to be used in assessing the trained model and mimics a real-world scenario where the model is tested on one new entry. One drawback of LOOCV is that it is more computationally expensive than K-Fold CV.

1. **Results**

One of the tasks attempted with the MORPH-II database is to correctly categorize the gender of an entry based on a portion of their Bio-Inspired Features (BIFs). Gender having only two classes is a form of categorical data, therefore classification methods are used for this task.

|  |  |  |
| --- | --- | --- |
| Assessing Gender Classification Using Five-Fold Cross Validation | | |
|  | **Accuracy** | **Standard Deviation** |
| Logistic Regression | 82.7% | 2.9% |
| LDA | 86.1% | 1.3% |
| QDA | 84.3% | 0.0% |
| K-NN (With K=3) | 84.3% | 1.3% |
| Decision Tree | 79.3% | 1.6% |
| Bagging | 85.3% | 0.8% |
| Random Forest | 85.3% | 2.0% |
| Boosting | 85.6% | 2.1% |

Table 1 Comparison of Classification Techniques Using 5-Fold CV

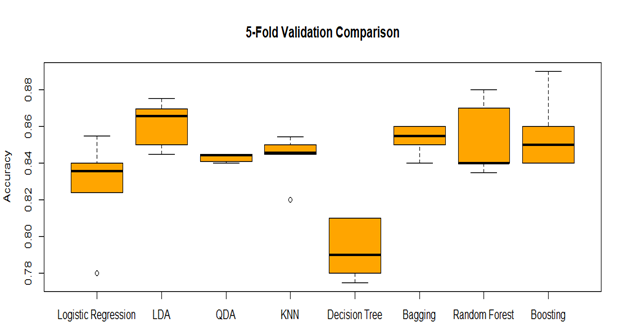
Table 1 shows a comparison of all the methods tested at this task, using 5-Fold CV as an assessment of accuracy. The methods are relatively close, with LDA having the highest accuracy and Boosting coming in second place. Interestingly, Random Forest and Bagging have the same 5-Fold accuracy even though Random Forest is supposed to show an improvement.

Figure 1 Box Plot Comparing Accuracy of Classification Techniques using 5-Fold CV

Figure 1 shows the box plot comparison of the different methods. Decision Tree has the lowest median. Given the simplicity of the Decision Tree model, this is not a surprise.

Table 2 shows that Leave One Out Cross Validation agrees with the top two highest ranked methods as 5-Fold Cross validation; however, there are some differences. Random Forest pulls ahead of Bagging, which is what one would expect given the motivation for Random Forests. Looking at the run times, Random Forest is significantly quicker than bagging as it does not train on all the predictors.

|  |  |  |  |
| --- | --- | --- | --- |
| Assessing Gender Classification Using Leave One Out Cross Validation | | | |
|  | **Accuracy** | **Standard Error** | **Time** |
| Logistic Regression | 83.9% | 36.8 | — |
| LDA | 86.2% | 34.5 | — |
| QDA | 84.3% | 36.4 | — |
| K-NN (K=3) | 84.4% | 36.3 | — |
| Decision Tree | 79.8% | 40.2 | 00:03:39 |
| Bagging | 84.3% | 36.4 | 03:28:43 |
| Random Forest | 85.1% | 35.6 | 01:54:40 |
| Boosting | 85.7% | 35.0 | 01:12:44 |

Table 2 Comparison of Classification Techniques Using LOOCV

The next task was to create a model that predicts the age of an entry using 150 of her BIF Values. As age is numerical data a regression model is used. The error metric is Mean Absolute Error which is defined as:

This can be interpreted as the average amount of years the model is off by. Boosting performs the best with a MAE of five and a half years, which Bagging coming in second place.

|  |  |  |
| --- | --- | --- |
| Assessing Regression Models of Age Using 5-Fold Cross Validation and Mean Absolute Error | | |
| Method | **Mean Absolute Error** | **Standard Deviation** |
| Decision Tree | 6.10 | 0.27 |
| Bagging | 5.59 | 0.25 |
| Random Forest | 5.77 | 0.24 |
| Boosting | 5.48 | 0.32 |

Table 3 Comparison of Regression Techniques using 5-Fold Cross Validation

1. **Conclusion**

The results show that LDA performs the best for gender classification while Boosting is the top tree method for both gender classification and age regression. These are only preliminary results; different optimizations can be used to enhance the performance of each of these models.

Dimension reduction can be used to reduce the amount of dimensions needed to represent all of the 2,569 BIF features allowing for more information to be used to train each model.

The tree models can also be pruned to lessen the chances of overfitting. In addition, other forms of boosting, such as Ada-boosting can be tested and compared to the current results.

1. **References**

[1] <https://www.faceaginggroup.com/morph/>

[2] <http://people.uncw.edu/vetterr/MORPH-NonCommercial-Stats.pdf>

[3] G. James et al., *An introduction to statistical learning: with applications in R*. New York: Springer, 2014.