Project Proposal

Yuhan Meng,Yuhang Lan,Shirley Zhang

1.Introduction

In recent years, cars play a really important role in our daily life. With them, it is convenient for us to go outside no matter what the weather is. Also, cars greatly reduce the amount of time people spend on the road. As university students, it won't be so long for us to buy our first cars, but how can we find the suitable cars? In order to give some recommendations to those people who are not sure which car to buy, we try to use K-NN,decision trees,bagging with k-NN and random forests to label cars. We use "Car Evaluation " dataset to generate our algorithms. There are 1728 observations，6 variables，4 class labels in this dataset. The features we use to classify the cars are "the buying price", "price of the maintenance", "number of doors", "capacity in terms of person to carry", "the size of the luggage boot" and "estimated safety of the car". Besides, the labels of the cars are the evaluation results which are "unacceptable", "acceptable", "good" and "very good".

2.Goals

The basic goals of our project are as following:

- Construct suitable machine learning model based on K-NN,decision trees,bagging with k-NN and random forests.
- Compare and evaluate the performance of these models .
- Optimize models.
- Choose the best model for application.

3. Model construction

The process of building the model :
First, we divide the dataset into three parts, the training dataset, the validation dataset and the testing dataset.
Then we use four learning methods to get the models and the learning methods are k-NN, decision trees and ensemble methods (bagging and random forest).
Next, we predict the labels of testing dataset.
Finally, since the labels of these observations are already given, we can evaluate models and compare the performance of k-NN and decision trees on this dataset.
What's more,we compare the performance of a single base classifier with ensemble methods--comparing the performance of decision tree with random forest and comparing the performance of k-NN with bagging(with k-NN). For instance, we will

calculate the accuracy score of the prediction on the training and testing dataset to compare the performance of the random forest to the performance of a single unpruned decision tree. Then, we can select model based on the performance of different models.

4. Model evaluation

In real life, when we decide to buy some produces we always try to compare different products of the same kind. The method we used to evaluate cars can also apply to evaluate other products.Once we get a dataset including a lot of variables and their features, we can evaluate the products and provide some suggestions.
We may evaluate the model by following methods:
(1)Basic methods.

- Bias and variance.If bias and variance are both at a low level, the model will be good. We may check whether using the ensemble method will decrease the variance or bias of the whole algorithm.The variance provides an estimate of how much the estimate varies as we vary the training data. Hence, when we apply bagging method, we may compare the ensembing model with the basic model (like k-NN).
- Overfitting and underfitting. The model shouldn't be too overfitting or underfitting. For instance, if we choose k-NN model, the number of the features may influence the model's accuracy. Too small or too large number of the features will lead to underfitting and overfitting of the model.Then we need to find models that perfectly reduce the error of both training set and testing set, meanwhile, decrease the gap between traing and test error.
- Confidence interval. First get the point estimate for the accuracy.If we have N test data, and X predicts the correct number of records, then we can estimate the accuracy by X / N. However, this is not very scientific because we use sample estimation, which has high probability of bias. Therefore, the more scientific method is to estimate the interval of accuracy. Here, the confidence interval in statistics may be used.Next, we may get the statistical 95% confidence interval based on the distribution assumption of the dataset.

(2)Resampling methods (repeated holdout,empirical confidence interval).
(3)Cross validation (hyperparameter tuning,model selection, algorithm selection).
(4)Statistical tests
(5)Evaluation metrics

5.Application

In the scope of machine learning, classification is a crucial topic. Classification refers to identifying to which of a set of categories a new observation belongs, on the basis of a training dataset. In many real life problems, people are facing classification objects in order to make choices leading to optimizing some specific goals. According to our experience, classification is always applied in social science study and retrieving information from some sources. For example, judge whether a patient is diagnosed with breast cancer or not based on her lab results.

There are many issues which can share the same algorithm with this subject. At the commercial level, classification as the approach to distinguish merchandise is one of the major topics in the discussion to optimize the customer experience. It is expected not just to provide reference and standards when customers make selections on purchase, but also make the industry standardized and specialized.

By solving the algorithm of car evaluation, we could give some basic rules to complete classification procedures. The process can be easily ensembled together, made visible, and generalized to the outside world with the same cores. Based on some artificial but typical features, our model will be appropriate to popularized. For example, a customer who wishes to buy a car can evaluate the condition of it by entering basic information like number of doors and capacity in terms of person to carry to give a grade and have some intuition about the price. Or, socialists who would be interested in classifying the result of a questionnaire into several classes will get benefits from this procedure.

6.Resources

The dataset we are planning to use is Car Evaluation from UCL machine learning repository, which is derived from simple hierarchical decision model. According to the website, because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods. This is also why the k-NN and decision trees will be extremely appropriate for this data.The softwares we plan to use are Python and R. We use Python as the computation and programming tool. And we use R as a visualization tool to make some descriptive analysis.

7.Contributions

Three of us would share the work of collecting data, researching the references, and discuss what model we would use. Then we plan to establish 4 models. Yuhan Meng establishes 2 basic models of the k-NN and decision tree. Xueqian Zhang establishes the bagging with k-NN model. Yuhang Lan establishes the random forest model. After that we compare and optimize the models together. And choose the best model for our data. Then Yuhan Meng and Xueqian Zhang will complete the project report and Yuhang Lan will be in charge of the slides. And all of us will revise our work mutually. After all, three of us will give the speech together. We will meet each other on every Monday to discuss the problems we meet in that week.