Kaggle Competition: Titanic: Machine Learning from Disaster[1]

Kaijun HOU[1], Qiurui MA[2] {khouab, qmaai}@ust.hk

[1]Department of Computer Science and Engineering [2]Department of Computer Science and Engineering (for Dual Degree)

MATH 4432 Project 2

Dataset and Task

Variable	Definition	Type
survival	Survival	Boolean
pclass	Ticket class	Categorical
sex	Sex	Categorical
Age	Age in years	Numerical
sibsp	# of siblings / spouses aboard	Numerical
parch	# of parents / children aboard	Numerical
ticket	Ticket number	Literal
fare	Passenger fare	Numerical
cabin	Cabin number	Literal
embarked	Port of Embarkation	Categorical

- Learn from the training dataset and predict “survival” attribute in test dataset.

Data Preprocessing

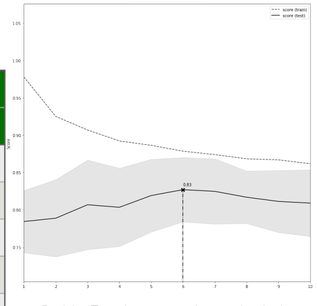
- Convert all categorical variables to discrete numerical variables
- Remove “ticket” and “cabin” variables since they are literal variables
  - For “ticket”, after removing all numbers, we notice that test set contains unseen ticket type in training set.
- Only “Age” and “Embarked” attributes have missing values:
  - Age: Fill missing values will 0
  - Embarked: Fill with (number_of_distinct_values_in_original_Embarked + 1)

[1]https://www.kaggle.com/c/titanic

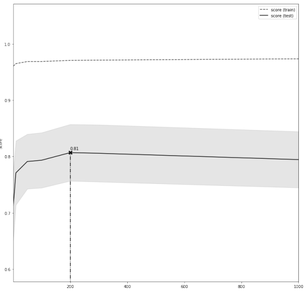
[2]Penalty parameter C of the error term

Experiment

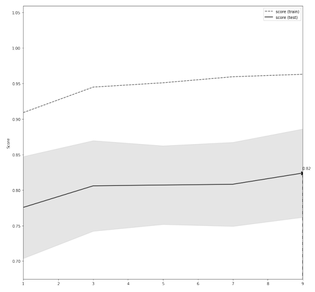
Classifier	Best Accuracy		
	Train	Validation	Test
Decision Tree	0.98	0.82	0.78
Random Forest	0.98	0.83	0.75
Bagging for Decision Tree	0.97	0.83	0.74
Bagging for Random Forest	0.93	0.83	0.74
AdaBoost for Decision Tree	0.98	0.80	0.73
AdaBoost for Random Forest	0.98	0.82	0.75
SVM	0.98	0.80	0.78
LDA		0.78	0.76
QDA		0.79	0.76
Gaussian Naive Bayes		0.79	0.75
Multinomial Naive Bayes		0.79	0.75
Multilayer Perceptron	0.71	0.70	0.76
kNN	0.98	0.72	0.64
Gradient Boosting	0.86	0.82	0.77



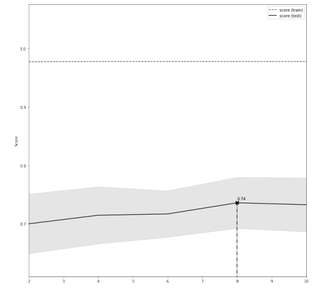
Decision Tree: Accuracy ~ min_samples_leaf



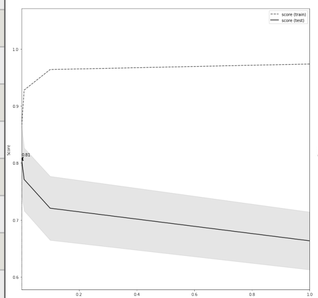
SVM: Accuracy ~ C[2]



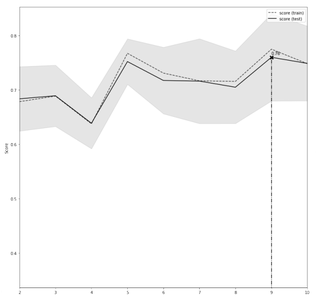
Decision Tree (Bagging): Accuracy ~ number of estimators



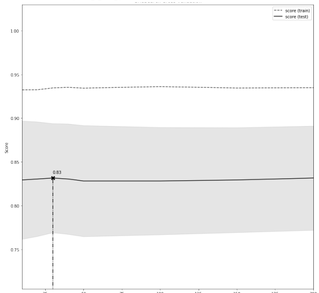
kNN: Accuracy ~ k



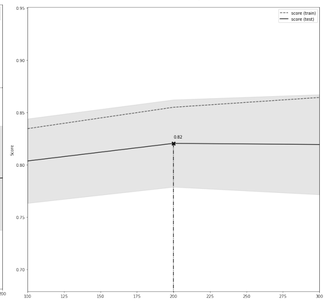
SVM: Accuracy ~ gamma



Multilayer Perceptron: Accuracy ~ Hidden layer size



Random Forest (Bagging):
Accuracy ~ number of trees in on random forest



Gradient Boosting: Accuracy ~ number of estimators

Try different classifiers. Do grid search on their params. Use highest cross-validation score as the validation accuracy and the corresponding parameters to do submission. Use the submission score as the test accuracy. Use the highest training accuracy over all parameter spaces and cross-validation folds as training accuracy. Grid search is not used on bayesian related classifiers. Thus only cross-validation score is given. All scores are shown in table above. Some classifiers provide feature importance ranking, the score is shown on the left.

Variable	Decision Tree	Gradient Boosting
pclass	0.11	0.10
sex	0.49	0.12
Age	0.12	0.26
sibsp	0.02	0.04
parch	0.04	0.05
fare	0.21	0.41
embarked	0.01	0.02

Analysis

Tree methods are very fitted for solving this problem, as most of its predictors are qualitative. The explainability of trees also favours the problem. However, its poor ability on extracting the interaction between features may severely hinder its better performances. Seven out of nine predictors are chose, as cabin and ticket have too much NaN or too much classes to be made use of. An overall importance of the predictors ranks the following: gender, age, fare, class, etc. The result does not differs much when the other three predictors are removed. This strongly aligns with Titanic's scenarios, as chances of survival is given based on gender and age regardless of kinsman-ship.

1. decision tree and bagging:
 max depth of 6 for 4 predictors indicates a possible interaction between gender, fare and gender, age. As the two is most likely. Because of its relative low level of interaction, decision tree in this case does not suffer from high variance. Furthermore, the optimal level of max depth is only 6, meaning a relatively low interaction as well. The two points all indicates low variance, rendering bagging less useful. The result also support this. But the reason why bagging methods underperform a single decision tree is unexplained somehow.
2. random forest:
 random forest shall not perform better than the decision tree approach, as the total number of predictors is few. The only possible correlation between all predictor is class and fare. Removing 'class' yields result similar to the decision tree. This may explain why random forest has approximately similar result with decision tree.

Contribution

Kaijun HOU: Codes and Poster
Qiurui MA: Analysis