Machine Learning Homework 1

2020047029

JunYeong Ahn

Q1.

1)

Decision Tree Acc: 0.7286245353159851

Decision Tree AUC: 0.7036297640653357

2)

Decision Tree Acc: 0.6728624535315985

Decision Tree AUC: 0.6460943257184968

3)

Decision Tree Acc: 0.6691449814126395

Decision Tree AUC: 0.6270472061657033

4)

Decision Tree Acc: 0.7211895910780669

Decision Tree AUC: 0.6849302383010248

5)

Decision Tree Acc: 0.7286245353159851

Decision Tree AUC: 0.7036297640653357

In this question, the decision tree used to predict the presence/absence of diabetes (target class) used Gini(measure of "inequality" of distribution) for its separation criterion. In this case 'Accuracy' is simply the rate of correct classifications and AUC(Area Under the (ROC) Curve) is a measure of how well the classifier distinguishes between classes – how true positive rate and false positive rate trade off. Both are for classification problem.

Accuracy and AUC range from 0 to 1. In most case well distinguished classes lead to high accuracy of model, which means Acc and AUC would have the same trend of increase/decrease. Seeing the five trials above, they do follow such trend. When Acc increases from 0.6691449814126395 to 0.7286245353159851 AUC also increases from 0.6270472061657033 to 0.7036297640653357– but AUC has the smaller range of fluctuation than Acc. Also both stayed around 0.6~0.7 and Acc was always higher than AUC.

Q2.

1)

Decision Tree Acc: 0.7063197026022305

Decision Tree AUC: 0.678412679845596

2)

Decision Tree Acc: 0.6691449814126395

Decision Tree AUC: 0.643970270918245

3)

Decision Tree Acc: 0.7063197026022305

Decision Tree AUC: 0.6812824139556812

4)

Decision Tree Acc: 0.6765799256505576

Decision Tree AUC: 0.6452513966480447

5)

Decision Tree Acc: 0.7026022304832714

Decision Tree AUC: 0.6530051150895141

Unlike the classifier in Q1, entropy(measure of how much "information" each feature contains relative to the target) is used for the splitting (in each nod) criterion option. These five trials also show that Acc and AUC usually have the same trend of increase/decrease – Acc from 0.6691449814126395 to 0.643970270918245 and AUC from 0.643970270918245 to 0.6812824139556812. And still, Acc was always higher than AUC. Considering given 5/5 trials in Q1&2, on average,using Gini index seems to be better than Entropy for this 'Prima_Diabetes' data because Accs and AUCs of the classifier using Gini were higher than them of the classifier using Entropy, although it is a slight difference between two.

Q3.

1)

Decision Tree Acc: 0.71 (+/- 0.08)

Decision Tree AUC: 0.69 (+/- 0.07)

2)

Decision Tree Acc: 0.71 (+/- 0.08)

Decision Tree AUC: 0.69 (+/- 0.07)

3)

Decision Tree Acc: 0.71 (+/- 0.08)

Decision Tree AUC: 0.69 (+/- 0.07)

CV Runtime: 0.03156433619890579

4)

Decision Tree Acc: 0.71 (+/- 0.08)

Decision Tree AUC: 0.69 (+/- 0.07)

5)

Decision Tree Acc: 0.71 (+/- 0.08)

Decision Tree AUC: 0.69 (+/- 0.07)

In all the trials the same result comes out unlike the trials in the train/test split scores in Q 1. Since the train set was randomly and newly sampled in each trial in question 1 and the test set also changed as a result, model becomes somewhat different with changed train set and is tested with different test set, which results in different outcomes of Acc and AUC. In k- fold cross validation (Here, k is 5 but I'll use k to generalize the CV steps), however, the data is divided into k parts (folds). And k-1 of the parts are used for training, and 1 is used for testing. This procedure is repeated k times rotating the test set then those results are averaged. These folds are divided as they are now even whenever CV is conducted again. Thus, the outcome does not change while it changes in Q1. Thus 0.71 of Acc and 0.69 of AUC are averaged version of k (5) cases, which means they give an insight on how the model will generalize to an independent dataset.

Also (+/- '0.08') and (+/- 0.07) on the back of Acc(0.71) and AUC(0.69) mean they guarantee each 0.63~0.79(Acc) and 0.62~0.78(AUC) of error range with 95% confidence level – for example, '0.08' is 2 (originally 1.96 that stands for 95% confidence level) * standard deviation. As a result, standard deviation of Accs and AUCs for each fold case of CV would be 0.04 and 0.035.

Q4.

Fold = 3)

Decision Tree Acc: 0.69 (+/- 0.06)

Decision Tree AUC: 0.67 (+/- 0.05)

CV Runtime: 0.01994633674621582

Fold = 8)

Decision Tree Acc: 0.69 (+/- 0.06)

Decision Tree AUC: 0.66 (+/- 0.07)

CV Runtime: 0.052850961685180664

Fold = 10)

Decision Tree Acc: 0.71 (+/- 0.14)

Decision Tree AUC: 0.67 (+/- 0.16)

CV Runtime: 0.09465167999267578

Acc didn't change when No. of folds change from 3 to 8 but has increased from 0.69 to 0.71 when No. of folds change from 8 to 10. But AUC has slightly decreased (to 0.66) and increased to get back to its original value(0.67) when No. of folds changed from 3 to 8 and 10. However, standard deviations become larger in both scores as k increases – from 0.3 to 0.7 in Acc and from 0.25 to 0.8 in AUC – since diversity (I'm not sure you will get what I want to mean though) of values that will be averaged has increased.

Run time is somewhat proportional to k (the number of fold), although it is not a perfect proportionality. This make sense since when the number of fold is k, the data is divided into k parts (folds). And k-1 of the parts are used for training, and 1 is used for testing. This procedure is repeated k times rotating the test set, which roughly means there is k times of calculation. With k being larger from 3 to 8 and 10, run times increase from about 0.02 to 0.05 and 0.1.

Comparing to the CV performance above in Q3 (k=5), three cases are worse than when No. of folds is 5 in Acc and AUC: the highest Acc and AUC when k = 5. However, the runtime might be the shortest in k=3 case, not in k=5 case because of the calculation load. So there could be something like trade-off between runtime and Acc&AUC.

+Actually there's no perfect answer to how many folds to use. We must ensure that the training set and testing set are drawn from the same distribution. And that both sets contain sufficient variation such that the underlining distribution is represented – we should consider the size of observations.

Q5.

1)

Decision Tree RMSE: 0.7814821083775916

Decision Tree Expl Var: 0.04443299536312206

2)

Decision Tree RMSE: 0.8548704998001577

Decision Tree Expl Var: -0.06320685838650286

3)

Decision Tree RMSE: 0.8153110712219863

Decision Tree Expl Var: -0.0824721254264924

4)

Decision Tree RMSE: 0.8081614937622307

Decision Tree Expl Var: -0.007775906142411859

5)

Decision Tree RMSE: 0.7945124291035351

Decision Tree Expl Var: 0.03477818637039343

MSE is a sum of differences between values predicted by a model and the values observed and RMSE is a root form of MSE. Unlike the other metrics, for RMSE lower is better. Explained Variance measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. Both are for regression

problem.

RMSE of the decision tree stayed around 0.8 (from 0.7814821083775916 to 0.8548704998001577). And explained variance almost stayed around 0 (from -0.0824721254264924 to 0.04443299536312206). Both have something in common in that they deal with the model and actual data. The lower RMSE and the higher Explained Variance, the better prediction of the regression model. However, we cannot identify the trend that RMSE grows as Expl Var decreases and that one decrease as one grows as well. Even with the same RMSE, Expl Var can de different and with higher RMSE Expl Var can be bigger as well. Seeing the trial 2) and 3), RMSE increases in 2) but Expl Var increases in 3) as well. So Expl Var seems to has nothing to do with the value of RMSE.

Q6.

1)

Decision Tree RMSE: 0.7656113336972712

Decision Tree Expl Var: 0.03240245681867271

2)

Decision Tree RMSE: 0.7691019809770724

Decision Tree Expl Var: 0.004950854207083966

3)

Decision Tree RMSE: 0.7936691466131932

Decision Tree Expl Var: -0.014336075507353385

4)

Decision Tree RMSE: 0.8390577197921828

Decision Tree Expl Var: -0.009922545674664551

5)

Decision Tree RMSE: 0.7416198487095663

Decision Tree Expl Var: 0.11435922983293167

RMSE of the decision tree became lower than 0.8 in most case (from 0.7416198487095663 to 0.8390577197921828). And most of explained variances became positive, which means they get larger than 0 on average (from -0.014336075507353385 to 0.11435922983293167), although there is such a subtle difference that it is hard to say that the change of option brought a prominent change in values with only 5 trials.

These five trials show that RMSE and explained variance are not related. Considering given 5/5 trials in Q5&6, on average, using ‘friedman_mse’ seems to be better than ‘mse’ for this ‘WineQuality_Red’ data because Explained Variance became slightly larger than that of Q5 and RMSE got lower in that the minimum and maximum of RMSE both decreased and them of Expl Var both increased in Q6 compared to Q5.

Q7.

1)

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

2)

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

3)

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

4)

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

5)

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

Similar to the answer for Q3, in all the trials the same result comes out unlike the trials in the train/test split scores in Q5. In k- fold cross validation, the data is divided into k parts (folds). And k-1 of the parts are used for training, and 1 is used for testing. This procedure is repeated k times rotating the test set then those results are averaged. The folds are divided as they are now even whenever CV is conducted. Thus the 0.9 of RMSE and -0.31 of Expl Var are averaged ones of k (5) cases, which means they give an insight on how the model will generalize to an independent dataset.

Also (+/- '0.10') and (+/- 0.17) on the back of RMSE(0.90) and Expl Var(-0.31) mean they guarantee each 0.80~1.00(RMSE) and -0.49~-0.14(Expl Var) of error range with 95% confidence level – for example, '0.10' is 2 (originally 1.96 that stands for 95% confidence level) * standard deviation. Thus, standard deviation of RMSEs and Expl Vars for each fold case of CV would be 0.05 and 0.085.

Q8.

Fold=3)

Decision Tree RMSE:: 0.98 (+/- 0.09)

Decision Tree Expl Var: -0.46 (+/- 0.27)

CV Runtime: 0.13962149620056152

Fold=8)

Decision Tree RMSE:: 0.93 (+/- 0.11)

Decision Tree Expl Var: -0.50 (+/- 0.64)

CV Runtime: 0.21343350410461426

Fold=10)

Decision Tree RMSE:: 0.91 (+/- 0.16)

Decision Tree Expl Var: -0.47 (+/- 0.85)

CV Runtime: 0.28024744987487793

RMSE decreased from 0.98 to 0.93 and 0.91 as No. of folds change from 3 to 5 and 8. But Expl Var has slightly decreased and increased (-0.46 > -0.50 > -0.47) as No. of folds change from 3 to 5 and 8; actually Expl Var would have more to do with the distribution of data rather than with the number of folds. And standard deviations become larger in both scores as k increases – from 0.9 to 0.16 in RMSE and from 0.27 to 0.85 in Expl Var – since diversity(?) of values that will be averaged has increased.

Run time is still somewhat proportional to k (the number of fold), although it is not a perfect proportionality. This make sense since when the number of fold is k, the data is divided into k parts (folds). And k-1 of the parts are used for training, and 1 is used for testing. This procedure is repeated k times rotating the test set, which roughly means there is k times of calculation. With k being larger from 3 to 8 and 10, run times increase from about 0.14 to 0.21 and finally 0.28.

Comparing to the CV performance above in Q7 (k=5), three cases are worse than when No. of folds is 5 in RMSE and Expl Var: the lowest RMSE, the highest Expl Var and even the lowest standard variance when k = 5. However, the runtime might be the shortest in k=3 case, not in k=5 case because of the calculation load. So there could still be something like trade-off between runtime and Acc&AUC.

Q9.

Decision Tree RMSE:: 0.96 (+/- 0.04)

Decision Tree Expl Var: -0.48 (+/- 0.23)

In Q7.

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

Among total 10 -originally 11 but since 'Class' variable is our target class it is excluded in my counts of features – features; 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', only five features ('fixed acidity', 'residual sugar', 'free sulfur dioxide', 'total sulfur dioxide', 'alcohol') have survived and the other five features ('volatile acidity', 'citric acid', 'chlorides', 'density', 'pH', 'sulphates') have been abandoned after feature selection.

The mean of RMSE and error range (calculated in Q9. by 5-fold CV) are respectively 0.96 and 0.92~1.00. And the mean of Expl Var and error range (calculated in Q9. by 5-fold CV) are respectively -0.48 and -0.71~-0.25. Compared to the record in Q7, RMSE has increased by 0.06 and Expl Var has decreased by 0.17, which intuitively means predictions of our regression model become worse after feature selection.

Q10.

Decision Tree RMSE:: 0.91 (+/- 0.08)

Decision Tree Expl Var: -0.35 (+/- 0.26)

In Q7.

Decision Tree RMSE:: 0.90 (+/- 0.10)

Decision Tree Expl Var: -0.31 (+/- 0.17)

Among total 10 – as I mentioned at Q9. target variable 'Class' is excluded – features; 'fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', only three features ('volatile acidity', 'sulphates', 'alcohol') have survived and the other seven features ('fixed acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH') have been abandoned after Wrapper-Based feature selection.

The mean of RMSE and error range (calculated in Q10. by 5-fold CV) are respectively 0.91 and 0.83~0.99. And the mean of Expl Var and error range (calculated in Q10. by 5-fold CV) are respectively -0.35 and -0.61~-0.09. Compared to the record in Q7, RMSE has increased by 0.01 and Expl Var has decreased by 0.04. Seeing the figures themselves, still the performance gets worse after Wrapper-Based feature selection, although there can be a tradeoff between such subtle loss of score and something like interpretation simplicity.

Q11.

We can in certain situation but there's no absolute answer of it.

We cannot say that the Decision Tree model performed better on one of them (a classification problem and a regression problem). Because a classifier and a regressor uses different performance/evaluation score, we are not able to directly compare those two situations. For example, I've recorded Accuracy and AUC for our decision tree classifier, which both range from 0 to 1. And I've recorded RMSE whose value has no limitation and Expl Var for our regressor – different scale of value might be one of problem. Plus, RMSE represented 'bad prediction' as it getting large while other scores such as AUC or Expl Var usually represents 'nice prediction' of models as getting larger. And in the first place, the characteristics of scores differ based on problem type (classification/regression).

There are many more reason we cannot compare the performances of a classifier and a regressor than these. Thus,

Q12.

Never. What I mainly obtained was Accs & AUCs for classification, RMSEs & Expl Vars for regression and some feature selection results for classification. I need multiple performance scores (at least more than two) in a single probelm to rationally determine whether the Decision Tree is good for those two datasets. So it seems to be okay to decide the model is good or not - however, what figure of performance score indicates good? The answer on this question might be quite subject with a sole model. With those datasets some other models such as Logistic Regression, QDA and SVM have a chance to outperform Decision tree, but I didn't made any comparisons. Also, I should inpect something more like our csv datasets more carefully so that I can fully understand the problem and find the best answer for it before making such a decision.

Likewise, when my boss or a customer ask me to build a decision tree for one of two datasets, I would like to say;

'If you ask me to build 'a' decision tree, I can do so but cannot know whether there would be nicer models than it. I know 'generally' it has some advantages like simplicity in interpretation but do not know whether the decision tree will the best for this special case. *Children think 1,000 won is a lot of money until they get 10,000 won.'*