Correction - Assignment 1

October 15, 2012

Question a) The accuracy of the unpuned tree is smaller on test data, because of overfiting. The idea was just to explain that the unpruned tree captures all details of the training data, which makes it bigger. On the other hand, the pruned one will have lower accuracy for the training data, but it will be able to better generalize for the unseen data.

Question b) The accuracy of the unpuned tree is smaller on test data, because of overfiting. The idea was just to explain that the unpruned tree captures all details of the training data, which makes it bigger. On the other hand, the pruned one will have lower accuracy for the training data, but it will be able to better generalize for the unseen data.

Wilcoxon test is non-parametrict because it does not make any assumptions about data distribution, and it is more suitable to compare classifiers in multiple domains (using different datasets), but it has some issues. It can be adapted for a multiple trails on the concerned domain. The resulting classifier performance measure over each trail can then be used for comparison in the test. The multiple trail, however violate the independent domain assumption because particularly all the dataset on which the measure are obtained would overlap. This result in a bias in the performance estimates. Wilcoxon test is a signed ranking test, which may create some loss of information.

Question c) The source code for the `crx` dataset is as follows. There are better solutions, than the one presented here:

```
library(DMwR)
library(RWeka)
crx <- read.arff ( file = "crx.arff" )
cv.J48.unpruned<-function(form,train,test,...) {
     classifier1Model <- J48( class ~ . , data=train, control = Weka_control(U=TRUE) )
     classifier1Evaluation <- evaluate_Weka_classifier(classifier1Model ,newdata=test )
     classifier1Accuracy <- as.numeric( substr (classifier1Evaluation$string , 70 ,80) )
     c(classifier1Accuracy)
}
cv.J48<-function(form,train,test,...) {
     classifier1Model <- J48( class ~ ., data=train )
     classifier1Evaluation <- evaluate_Weka_classifier(classifier1Model ,newdata=test )
     classifier1Accuracy <- as.numeric( substr (classifier1Evaluation$string , 70 ,80) )
     c(classifier1Accuracy)
}

res <- experimentalComparison(c(dataset(class ~ .,crx,'crx')),
```

```
                                c(variants('cv.J48.unpruned'),
                                  variants('cv.J48')),
                                cvSettings(3,10,1234))

t.test(getFoldsResults(res, 'cv.J48.unpruned.defaults'),
       getFoldsResults(res,'cv.J48.defaults'),
       paired=TRUE)
wilcox.test(getFoldsResults(res, 'cv.J48.unpruned.defaults'),
            getFoldsResults(res,'cv.J48.defaults'),
            paired=TRUE)
```

T-test Output:

```
data:  getFoldsResults(res, "cv.J48.unpruned.defaults")
       and getFoldsResults(res, "cv.J48.defaults")
t = -5.2718, df = 29, p-value = 1.192e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.983697 -2.197637
sample estimates:
mean of the differences
              -3.590667
```

P-value<0.05, then there is statistical significance, with 95% confidence.

Wilcoxon-test Output:

```
data:  getFoldsResults(res, "cv.J48.unpruned.defaults")
       and getFoldsResults(res, "cv.J48.defaults")
V = 27, p-value = 6.289e-05
alternative hypothesis: true location shift is not equal to 0
```

Once more, the P-value<0.05, then there is statistical significance, with 95% confidence, for the Wilcoxon test.

Question d) The source code for the `titanic` dataset is as follows. There are better solutions, than the one presented here:

```
library(DMwR)
library(RWeka)
titanic <- read.table("titanic.data",header=T)
cv.J48.unpruned<-function(form,train,test,...) {
     classifier1Model <- J48( survived ~ . , data=train, control = Weka_control(U=TRUE) )
     classifier1Evaluation <- evaluate_Weka_classifier(classifier1Model ,newdata=test )
     classifier1Accuracy <- as.numeric( substr (classifier1Evaluation$string , 70 ,80) )
     c(classifier1Accuracy)
}
cv.J48<-function(form,train,test,...) {
     classifier1Model <- J48( survived ~ ., data=train )
     classifier1Evaluation <- evaluate_Weka_classifier(classifier1Model ,newdata=test )
     classifier1Accuracy <- as.numeric( substr (classifier1Evaluation$string , 70 ,80) )
     c(classifier1Accuracy)
}
```

```
res <- experimentalComparison(c(dataset(survived ~ .,titanic,'titanic')),
                              c(variants('cv.J48.unpruned'),
                                variants('cv.J48')),
                              cvSettings(3,10,1234))

t.test(getFoldsResults(res, 'cv.J48.unpruned.defaults'),
       getFoldsResults(res,'cv.J48.defaults'),
       paired=TRUE)
wilcox.test(getFoldsResults(res, 'cv.J48.unpruned.defaults'),
            getFoldsResults(res,'cv.J48.defaults'),
            paired=TRUE)
```

T-Test Output:

```
data:  getFoldsResults(res, "cv.J48.unpruned.defaults")
       and getFoldsResults(res, "cv.J48.defaults")
t = 1.6956, df = 29, p-value = 0.1007
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.08434397  0.90234397
sample estimates:
mean of the differences
                  0.409
```

P-value>0.05, then there is **no** statistical significance, with 95% confidence.

Wilcoxon-test Output:

```
data:  getFoldsResults(res, "cv.J48.unpruned.defaults")
       and getFoldsResults(res, "cv.J48.defaults")
V = 34, p-value = 0.1917
alternative hypothesis: true location shift is not equal to 0
```

Once more, the P-value>0.05, then there is *no* statistical significance, with 95% confidence, for the Wilcoxon test.