Lab3_Q2

March 29, 2020

ComS 573

Lab 3

Kanak Choudhury

1 Problem 2

[1]:
```
import numpy as np
import sklearn
from sklearn import tree
from sklearn.impute import SimpleImputer
from sklearn.model_selection import cross_val_score
import sys
print('python ' +sys.version)
print('numpy '+ np.__version__)
print('sklearn '+ sklearn.__version__,'\n\n')

data = open('house-votes-84.data','r').read().splitlines();
dt_size = np.shape(data);
dt_x = np.zeros([dt_size[0],16]);
dt_y = [];

for i in range(0,dt_size[0]):
    aa = data[i].split(',')
    dt_y.append('republican' if aa[0]=='republican' else 'democrat')
    dt_x[i,:] = [-1 if aa[x+1]=='?' else 1 if aa[x+1]=='y' else 0 for x in␣
 ↪range(0,16)]

dt_y = np.asarray(dt_y)
# impute = SimpleImputer(missing_values=-1, strategy='most_frequent')
# impute.fit(dt_x)
# dt_x = impute.transform(dt_x)


ctree = tree.DecisionTreeClassifier()
acc = cross_val_score(ctree, dt_x, dt_y, cv=5)
print("Accuracies for 5-fold classification:")
```

```
for i in range(5):
    print('Accuracy for fold %d: %.2f%%' %(i+1,acc[i]*100))

print("\n")

ci = np.array([acc.mean()-acc.std()*1.96, acc.mean()+acc.std()*1.96])
print("Confidence interval is given as following:")
print('CI: lower: %.4f,     upper: %.4f' %(ci[0], ci[1]))
print("\n")
```

```
python 3.6.9 |Anaconda, Inc.| (default, Jul 30 2019, 14:00:49) [MSC v.1915 64
bit (AMD64)]
numpy 1.16.5
sklearn 0.21.3


Accuracies for 5-fold classification:
Accuracy for fold 1: 95.45%
Accuracy for fold 2: 95.45%
Accuracy for fold 3: 96.55%
Accuracy for fold 4: 93.02%
Accuracy for fold 5: 90.70%


Confidence interval is given as following:
CI: lower: 0.9009,     upper: 0.9838
```

[ ]: