DMG2 Assignment : Problem 3

Naive Bayes Classifier, Decision Tree Classifier

In [1]:

```
import numpy as np
import pandas as pd
import os
import scipy
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn import tree
from sklearn.feature_extraction import DictVectorizer
from sklearn.preprocessing import LabelEncoder
from sklearn.naive_bayes import MultinomialNB

sns.set_style('whitegrid')
```

In [2]:

```
DATA_DIR = '/home/jishnu/Documents/ISB/Term3/dmg2/assignments/hw_assignment
1/dmg2/datasets/mushroom'
train = pd.read_csv(os.path.join(DATA_DIR,'train.csv'),usecols=['V{0}'.forma
t(i) for i in range(1,24)])
test = pd.read_csv(os.path.join(DATA_DIR,'test.csv'),usecols=
['V{0}'.format(i) for i in range(1,24)])

train.columns
```

Out[2]:

```
Index(['V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11',
       'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21',
       'V22', 'V23'],
      dtype='object')
```

In [3]:

```
# Vectorizing categorical data
X_dict = train.iloc[:,1:].T.to_dict().values()
X_vector = DictVectorizer(sparse=False).fit_transform(X_dict)

X_test_dict = test.iloc[:,1:].T.to_dict().values()
X_test_vector = DictVectorizer(sparse=False).fit_transform(X_test_dict)

# Vectorizing class labels
le = LabelEncoder()
Y_train = le.fit_transform(train.iloc[:,0])
Y_test = le.fit_transform(test.iloc[:,0])
```

Decision Tree Classifier

In [4]:

```
dt_clf = tree.DecisionTreeClassifier(max_depth=10).fit(X_vector,Y_train)
```

In [5]:

```
dt_clf.score(X_vector,Y_train)
```

Out[5]:

```
1.0
```

In [6]:

```
dt_clf.score(X_test_vector,Y_test)
```

Out[6]:

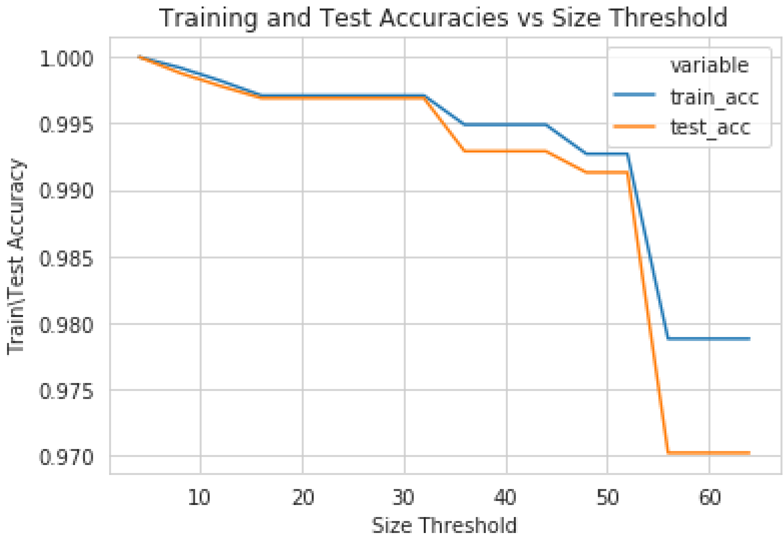
```
1.0
```

In [20]:
```
dt_accuracies = pd.DataFrame(columns=['size_threshold','train_acc','test_acc'])
for size_threshold in range(4,65,2):
    dt_clf = tree.DecisionTreeClassifier(min_samples_leaf=size_threshold,criterion='entropy').fit(X_vector,Y_train)
    train_acc = np.round(dt_clf.score(X_vector,Y_train),4)
    test_acc = np.round(dt_clf.score(X_test_vector,Y_test),4)
    dt_accuracies = dt_accuracies.append({'size_threshold' : size_threshold,'train_acc' : train_acc,'test_acc' : test_acc},ignore_index=True)
dt_accuracies.head()
```

Out[20]:

	size_threshold	train_acc	test_acc
0	4.0	1.0000	1.0000
1	6.0	0.9996	0.9994
2	8.0	0.9992	0.9988
3	10.0	0.9992	0.9988
4	12.0	0.9982	0.9978

In [8]:
```
sns.lineplot(x='size_threshold',y='value',hue='variable',
             data=dt_accuracies.melt(id_vars=['size_threshold'],value_vars=['train_acc','test_acc']),
             ci=0)
plt.xlabel('Size Threshold')
plt.ylabel('Train\Test Accuracy')
plt.title('Training and Test Accuracies vs Size Threshold')
plt.show();
```



The test accuracies start decreasing at around size threshold of 32.

Naive Bayes Classifier

In [9]:

```
nb_accuracies = pd.DataFrame(columns=
['lap_sm_param','train_acc','test_acc'])
for lap_sm_param in range(0,51):
    nb_clf = MultinomialNB(alpha=lap_sm_param).fit(X_vector,Y_train)
    train_acc = np.round(nb_clf.score(X_vector,Y_train),4)
    test_acc = np.round(nb_clf.score(X_test_vector,Y_test),4)
    nb_accuracies = nb_accuracies.append({'lap_sm_param' : lap_sm_param,'tra
in_acc' : train_acc,'test_acc' : test_acc},ignore_index=True)
nb_accuracies.head()
```

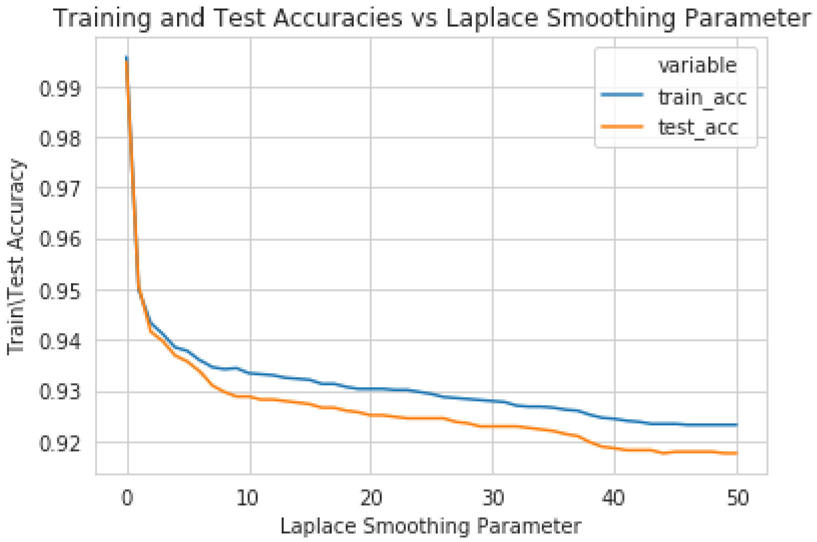
```
/home/jishnu/anaconda3/lib/python3.6/site-packages/sklearn/naive_bayes.py:47
2: UserWarning: alpha too small will result in numeric errors, setting alpha
= 1.0e-10
  'setting alpha = %.1e' % _ALPHA_MIN)
```

Out[9]:

	lap_sm_param	train_acc	test_acc
0	0.0	0.9957	0.9947
1	1.0	0.9499	0.9506
2	2.0	0.9433	0.9416
3	3.0	0.9411	0.9397
4	4.0	0.9385	0.9369

In [10]:

```
sns.lineplot(x='lap_sm_param',y='value',hue='variable',
             data=nb_accuracies.melt(id_vars=['lap_sm_param'],value_vars=['tra
in_acc','test_acc']),
             ci=0)
plt.xlabel('Laplace Smoothing Parameter')
plt.ylabel('Train\Test Accuracy')
plt.title('Training and Test Accuracies vs Laplace Smoothing Parameter')
plt.show();
```



The best value of test accuracy is achieved when setting smoothing parameter to zero.

The decision tree classifier gives much better accuracies when compared to naive bayes classifier.

Google Form Answers

1) What's the training accuracy for Naive Bayes classifier at lambda = 10?

```
In [12]: nb_accuracies.loc[nb_accuracies['lap_sm_param'] == 10]['train_acc']
```

```
Out[12]: 10    0.9334
         Name: train_acc, dtype: float64
```

2) Whats the test accuracy for Naive Bayes classifier at lamda = 30?

```
In [16]: nb_accuracies.loc[nb_accuracies['lap_sm_param'] == 30]['test_acc']
```

```
Out[16]: 30    0.9229
         Name: test_acc, dtype: float64
```

3) What's the training accuracy of decision tree classifier at SizeThreshold = 30?

```
In [21]: dt_accuracies.loc[dt_accuracies['size_threshold'] == 30]['train_acc']
```

```
Out[21]: 13    0.9971
         Name: train_acc, dtype: float64
```

4) What's the test accuracy of decision tree classifier at SizeThreshold = 10?

```
In [22]: dt_accuracies.loc[dt_accuracies['size_threshold'] == 10]['test_acc']
```

```
Out[22]: 3    0.9988
         Name: test_acc, dtype: float64
```