In [68]:

```
import os
os.chdir("/Users/wangkewei/Desktop/machine learning/ML-DSBA-AI-Assignment_1")
import scipy as sp
import numpy as np
from sklearn.cross_validation import train_test_split
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import roc_auc_score
import timeit
```

In [69]:

```
os.getcwd()
```

Out[69]:

```
'/Users/wangkewei/Desktop/machine learning/ML-DSBA-AI-Assignment_1'
```

In [70]:

```
import pandas as pd
# Load the data set
data = pd.read_csv('data.csv', delimiter=',')
#load 1st column
Y = data.iloc[:,0:1]
# load columns 2 - end
X = data.iloc[:,1:data.shape[1]]
test = pd.read_csv('test.csv', delimiter=',')
Y_test = test.iloc[:,0:1]
X_test = test.iloc[:,1:data.shape[1]]
```

In [71]:

```
#for question1
model = LogisticRegression()
start = timeit.default_timer()
model.fit(X, Y)
stop = timeit.default_timer()
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [72]:

```
Y_pred = model.predict(X_test)
```

In [73]:

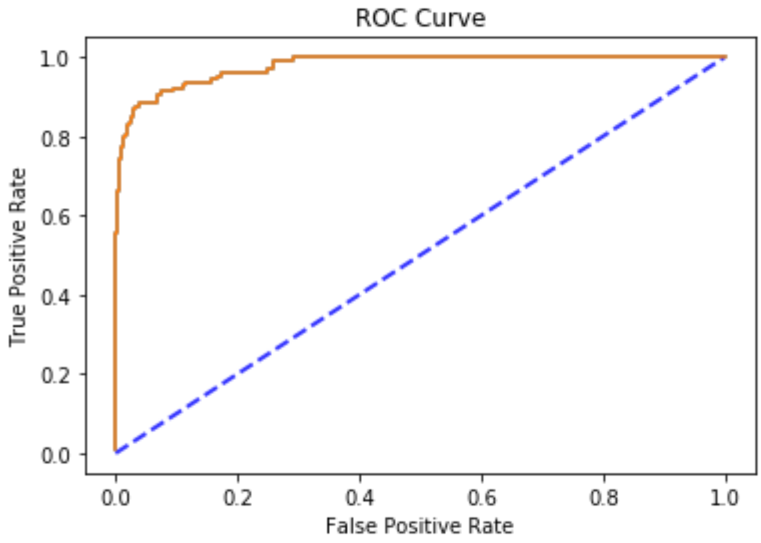
```
Y_pred_proba = model.predict_proba(X_test)[:,1]
```

In [74]:

```
fpr, tpr, thresholds = roc_curve(Y_test, Y_pred_proba)
```

In [75]:

```
plt.plot(fpr, tpr)
plt.plot([0, 1], [0, 1], linestyle='--', lw=2, color='b',
         label='Chance', alpha=.8)
plt.plot(fpr,tpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.show()
```



In [76]:

```
AUC = roc_auc_score(Y_test,Y_pred_proba)
print(AUC)
```

```
0.9776601239669421
```

In [77]:

```
print('Time: ', stop-start)
```

```
Time:  1.1959112840122543
```

In [78]:

```
#for question2
from sklearn.feature_selection import RFE
```

In [79]:

```
runtime = []
AUC2 = []
```

In [80]:

```
#20
selection_20 = RFE(model, 20, step = 1)
start_20 = timeit.default_timer()
selection_20.fit(X,Y)
stop_20 = timeit.default_timer()
runtime.append(stop_20-start_20)
Y_pred_prob_20 = selection_20.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_20))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [81]:

```
#40
selection_40 = RFE(model, 40, step = 1)
start_40 = timeit.default_timer()
selection_40.fit(X,Y)
stop_40 = timeit.default_timer()
runtime.append(stop_40-start_40)
Y_pred_prob_40 = selection_40.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_40))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [82]:

```
#60
selection_60 = RFE(model, 60, step = 1)
start_60 = timeit.default_timer()
selection_60.fit(X,Y)
stop_60 = timeit.default_timer()
runtime.append(stop_60-start_60)
Y_pred_prob_60 = selection_60.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_60))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [83]:

```
#80
selection_80 = RFE(model, 80, step = 1)
start_80 = timeit.default_timer()
selection_80.fit(X,Y)
stop_80 = timeit.default_timer()
runtime.append(stop_80-start_80)
Y_pred_prob_80 = selection_80.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_80))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [84]:

```
#100
selection_100 = RFE(model, 100, step = 1)
start_100 = timeit.default_timer()
selection_100.fit(X,Y)
stop_100 = timeit.default_timer()
runtime.append(stop_100-start_100)
Y_pred_prob_100 = selection_100.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_100))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [85]:

```
#150
selection_150 = RFE(model, 150, step = 1)
start_150 = timeit.default_timer()
selection_150.fit(X,Y)
stop_150 = timeit.default_timer()
runtime.append(stop_150-start_150)
Y_pred_prob_150 = selection_150.predict_proba(X_test)[:,1]
AUC2.append(roc_auc_score(Y_test,Y_pred_prob_150))
```

```
/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:5
78: DataConversionWarning: A column-vector y was passed when a 1d ar
ray was expected. Please change the shape of y to (n_samples, ), for
example using ravel().
  y = column_or_1d(y, warn=True)
```

In [86]:

```
runtime
```

Out[86]:

```
[90.31326198700117,
 86.0984163400135,
 83.33518329798244,
 76.05272869498003,
 63.91755524999462,
 19.882781848020386]
```

In [87]:

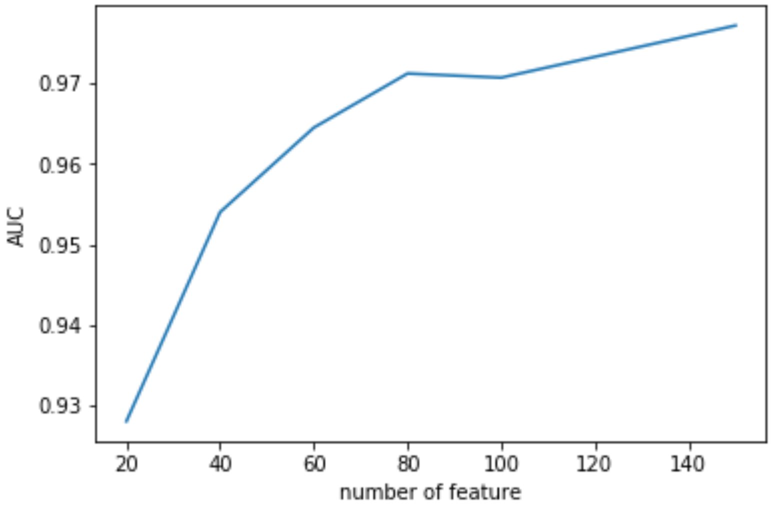
```
AUC2
```

Out[87]:

```
[0.9280272284533648,
 0.9540012544273908,
 0.964497860094451,
 0.9712219598583236,
 0.9707054309327037,
 0.9771804899645808]
```

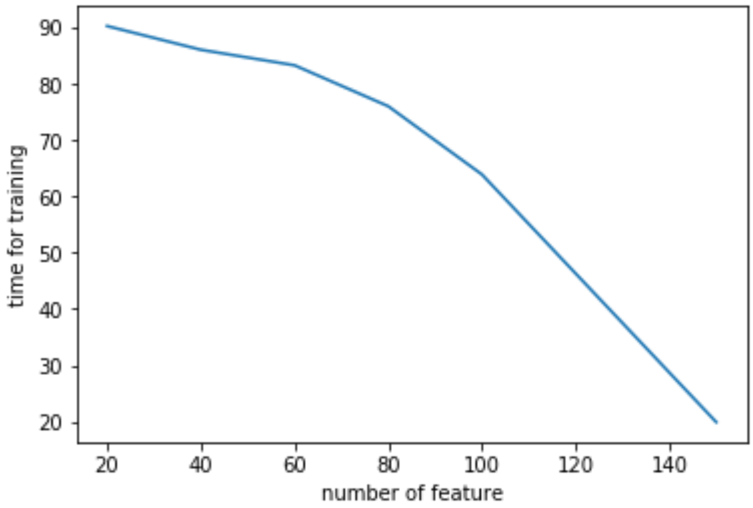
In [88]:

```
num_feature = [20,40,60,80,100,150]
plt.plot(num_feature,AUC2)
plt.xlabel('number of feature')
plt.ylabel('AUC')
plt.show()
```



In [89]:

```
plt.plot(num_feature,runtime)
plt.xlabel('number of feature')
plt.ylabel('time for training')
plt.show()
```



In [90]:

```
#conclusion: feature selection can't improve the accuracy compared with the prev
ious question's method whihc used all
#of the training data. However, we can find that the accuracy will increase with
 more features selected but the speed
#will slow down after 80. The time for training will decrease with more features
 selected.
```