Assignment 1 Part III

November 13, 2023

[ ]:
```python
import numpy as np
import scipy.optimize as sco
import matplotlib.pyplot as plt
```

0.1 PART III

Data generation

[ ]:
```python
# TRAIN

train_samples_per_class = 30

# Set the variance for each class
variance = 0.3

n1a = np.random.normal(loc=[0, 0], scale=np.sqrt(variance),␣
    ↪size=(train_samples_per_class, 2))
n1b = np.random.normal(loc=[0, 1], scale=np.sqrt(variance),␣
    ↪size=(train_samples_per_class, 2))
n2a = np.random.normal(loc=[1, 1], scale=np.sqrt(variance),␣
    ↪size=(train_samples_per_class, 2))
n2b = np.random.normal(loc=[1, 0], scale=np.sqrt(variance),␣
    ↪size=(train_samples_per_class, 2))

labels = np.array([1]*train_samples_per_class*2 +␣
    ↪[-1]*train_samples_per_class*2)


X_train = np.vstack([n1a, n1b, n2a, n2b])
Y_train = labels

shuffle_idx = np.random.permutation(X_train.shape[0])
X_train = X_train[shuffle_idx]
Y_train = Y_train[shuffle_idx]


# TEST
```

```
test_samples_per_class = 200

n1a = np.random.normal(loc=[0, 0], scale=np.sqrt(variance),␣
↪size=(test_samples_per_class, 2))
n1b = np.random.normal(loc=[0, 1], scale=np.sqrt(variance),␣
↪size=(test_samples_per_class, 2))
n2a = np.random.normal(loc=[1, 1], scale=np.sqrt(variance),␣
↪size=(test_samples_per_class, 2))
n2b = np.random.normal(loc=[1, 0], scale=np.sqrt(variance),␣
↪size=(test_samples_per_class, 2))

labels = np.array([1]*test_samples_per_class*2 + [-1]*test_samples_per_class*2)


X_test = np.vstack([n1a, n1b, n2a, n2b])
Y_test = labels

shuffle_idx = np.random.permutation(X_test.shape[0])
X_test = X_test[shuffle_idx]
Y_test = Y_test[shuffle_idx]
```

[ ]:
```
def soft_margin_SVM_fit(X, Y, C=1.0):
    #zeta_i >= 1 - y_i(w^T x_i + b) and zeta_i >= 0 -> zeta_i = max(0, 1 -␣
↪y_i(w^T x_i + b))
    #MINIMIZE: ||w||^2 + C * sum(zeta_i)

    d = X.shape[1] # d dimensions


    kernel = lambda params: np.linalg.norm(params[:-1])**2 + C * np.sum(np.
↪maximum(0, 1 - Y * (np.matmul(X, params[:-1]) - params[-1])))

    return sco.minimize(kernel, np.zeros(d+1))
```

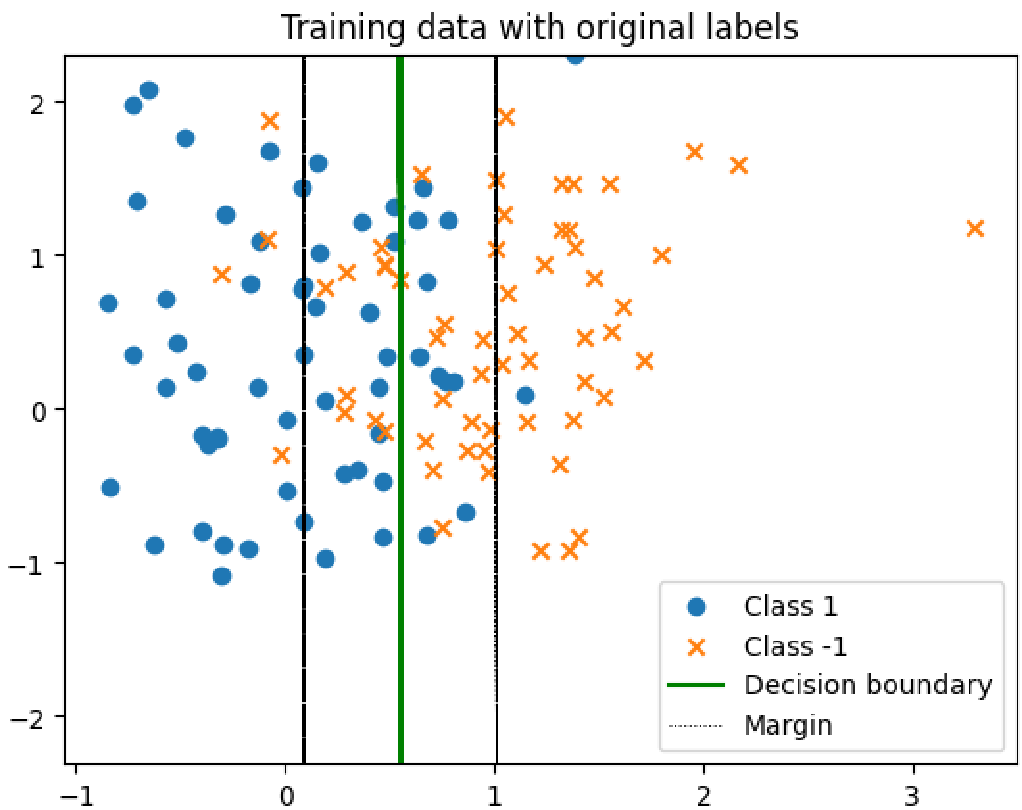
[ ]:
```
params = soft_margin_SVM_fit(X_train, Y_train, C = 3)
params = params.x
```

[ ]:
```
params
```

[ ]:
```
array([-2.16704405e+00, -1.28356783e-03, -1.19236064e+00])
```

[ ]:
```

```

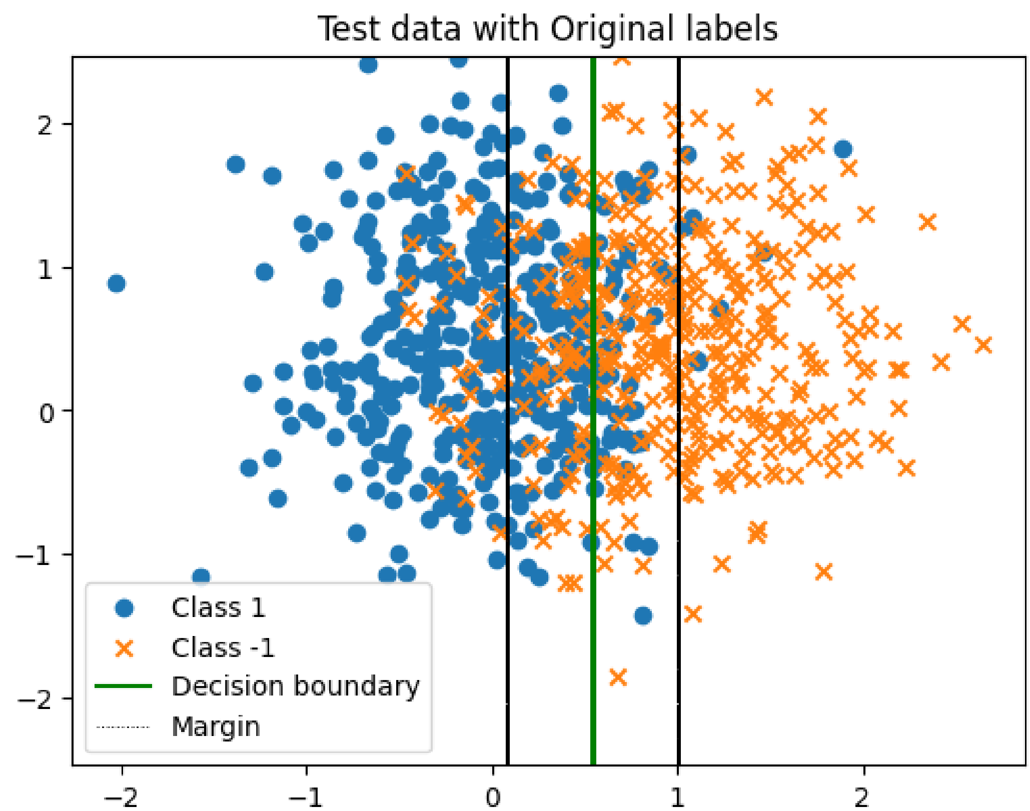
```
plt.scatter(X_train[Y_train == 1][:, 0], X_train[Y_train == 1][:, 1],␣
 ↪label='Class 1', marker='o')
plt.scatter(X_train[Y_train == -1][:, 0], X_train[Y_train == -1][:, 1],␣
 ↪label='Class -1', marker='x')
plt.plot(X_train[:, 0], -params[0]*X_train[:, 0]/params[1] + params[-1]/
 ↪params[1], label='Decision boundary', color='green')
plt.plot(X_train[:, 0], -params[0]*X_train[:, 0]/params[1] + params[-1]/
 ↪params[1] + 1/params[1], color='black', ls=':', lw=0.7)
plt.plot(X_train[:, 0], -params[0]*X_train[:, 0]/params[1] + params[-1]/
 ↪params[1] - 1/params[1], color='black', ls=':', lw=0.7, label='Margin')
plt.ylim(-max(abs(X_train[:, 1])), max(abs(X_train[:, 1])))
plt.legend()
plt.title('Training data with original labels')
plt.show()
```



[ ]:
```
Y_pred = np.sign(np.dot(X_test, params[:-1]) - params[-1])
```

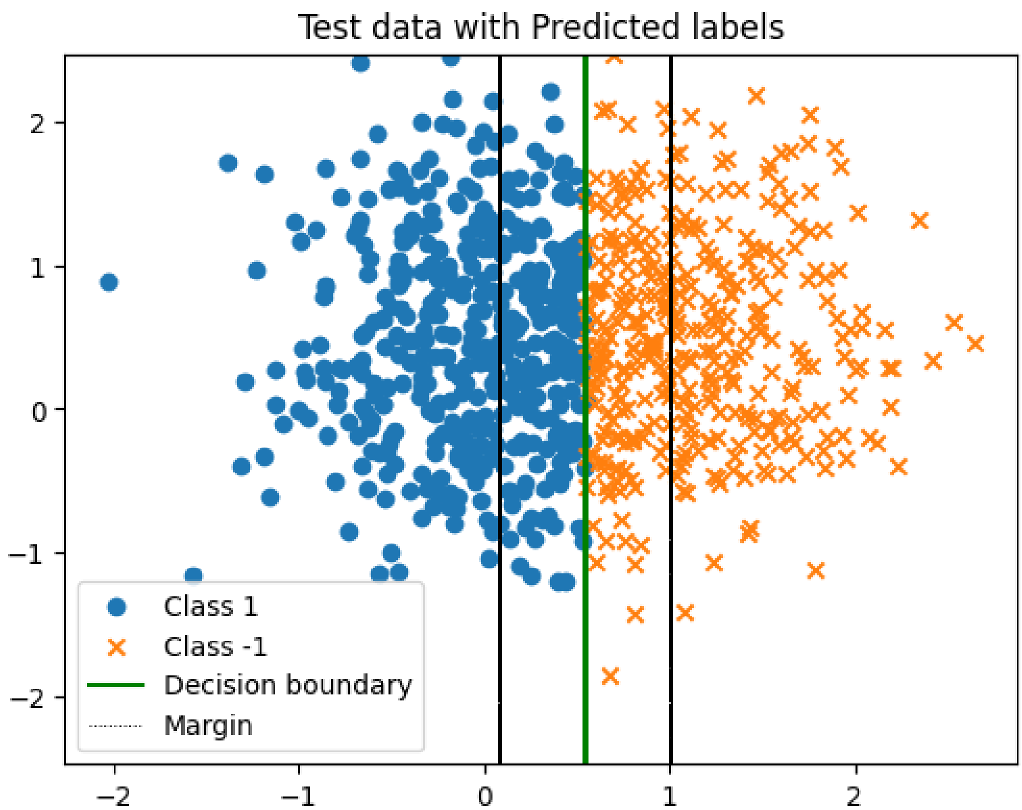
[ ]:
```
plt.scatter(X_test[Y_test == 1][:, 0], X_test[Y_test == 1][:, 1], label='Class␣
 ↪1', marker='o')
```

```
plt.scatter(X_test[Y_test == -1][:, 0], X_test[Y_test == -1][:, 1],␣
↪label='Class -1', marker='x')
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/
↪params[1], label='Decision boundary', color='green')
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/params[1]␣
↪+ 1/params[1], color='black', ls=':', lw=0.7)
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/params[1]␣
↪- 1/params[1], color='black', ls=':', lw=0.7, label='Margin')
plt.ylim(-max(abs(X_test[:, 1])), max(abs(X_test[:, 1])))
plt.title('Test data with Original labels')
plt.legend()
plt.show()
```



[ ]:
```
plt.scatter(X_test[Y_pred == 1][:, 0], X_test[Y_pred == 1][:, 1], label='Class␣
↪1', marker='o')
plt.scatter(X_test[Y_pred == -1][:, 0], X_test[Y_pred == -1][:, 1],␣
↪label='Class -1', marker='x')
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/
↪params[1], label='Decision boundary', color='green')
```

```
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/params[1]␣
 ↪+ 1/params[1], color='black', ls=':', lw=0.7)
plt.plot(X_test[:, 0], -params[0]*X_test[:, 0]/params[1] + params[-1]/params[1]␣
 ↪- 1/params[1], color='black', ls=':', lw=0.7, label='Margin')
plt.ylim(-max(abs(X_test[:, 1])), max(abs(X_test[:, 1])))
plt.title('Test data with Predicted labels')
plt.legend()
plt.show()
```



[ ]:
```
accuracy = np.count_nonzero(Y_pred == Y_test) / Y_test.shape[0]

print('Accuracy for test data after training over training data: ',␣
 ↪accuracy*100)
```

```
Accuracy for test data after training over training data:  81.125
```

(b) Repeat Q2(a) and Q2(b) of Part II to obtain best C for the current problem (instead of k in kNN classifier) and the current dataset.

[ ]:
```
# Training SVM over multiple values of C for holout of 0.3 and 10 repetitions
hold_out_rho = 0.3
```

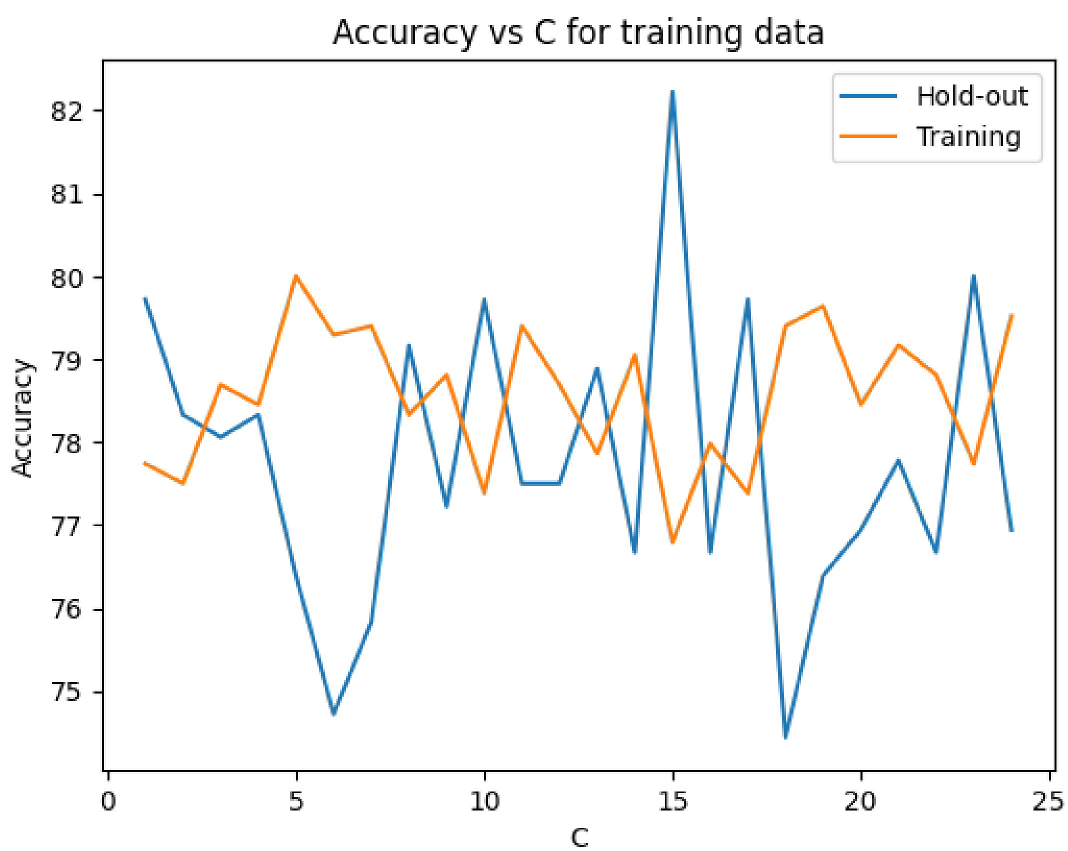
```python
n_rep = 10
def outCVSVM(X, Y, C, n_rep, hold_out_rho):
    n = len(X)
    n_hold_out = int(n*hold_out_rho)
    hold_out_accuracy = np.zeros(n_rep)
    train_accuracy = np.zeros(n_rep)
    for i in range(n_rep):
        shuffle_idx = np.random.permutation(n)
        X_train = X[shuffle_idx]
        Y_train = Y[shuffle_idx]
        X_hold_out = X_train[:n_hold_out]
        Y_hold_out = Y_train[:n_hold_out]
        X_train = X_train[n_hold_out:]
        Y_train = Y_train[n_hold_out:]
        # Y_pred_hold_out = soft_margin_SVM_fit(X_train, Y_train, X_hold_out, k)
        params = soft_margin_SVM_fit(X_train, Y_train, C).x
        Y_pred_hold_out = np.sign(np.dot(X_hold_out, params[:-1]) - params[-1])
        hold_out_accuracy[i] = np.count_nonzero(Y_hold_out == Y_pred_hold_out)/
 ↪len(Y_hold_out)
        # Y_pred_train = soft_margin_SVM_fit(X_train, Y_train, X_train, k)
        Y_pred_train = np.sign(np.dot(X_train, params[:-1]) - params[-1])
        train_accuracy[i] = np.count_nonzero(Y_train == Y_pred_train)/
 ↪len(Y_train)


    return np.round([np.mean(hold_out_accuracy)*100, np.
 ↪mean(train_accuracy)*100], 2)
```

[ ]:
```python
aaa = []
for C in range(1, 25):
    svmout = outCVSVM(X_train, Y_train, C, n_rep, hold_out_rho)
    print("C = ", C, ": ", svmout)
    aaa.append(svmout)
```

```
C =  1 :  [79.72 77.74]
C =  2 :  [78.33 77.5 ]
C =  3 :  [78.06 78.69]
C =  4 :  [78.33 78.45]
C =  5 :  [76.39 80.  ]
C =  6 :  [74.72 79.29]
C =  7 :  [75.83 79.4 ]
C =  8 :  [79.17 78.33]
C =  9 :  [77.22 78.81]
C =  10 :  [79.72 77.38]
C =  11 :  [77.5 79.4]
C =  12 :  [77.5  78.69]
C =  13 :  [78.89 77.86]
```

```
C =  14 :  [76.67 79.05]
C =  15 :  [82.22 76.79]
C =  16 :  [76.67 77.98]
C =  17 :  [79.72 77.38]
C =  18 :  [74.44 79.4 ]
C =  19 :  [76.39 79.64]
C =  20 :  [76.94 78.45]
C =  21 :  [77.78 79.17]
C =  22 :  [76.67 78.81]
C =  23 :  [80.   77.74]
C =  24 :  [76.94 79.52]
```

[ ]:
```python
plt.plot(range(1, 25), [x[0] for x in aaa], label="Hold-out")
plt.plot(range(1, 25), [x[1] for x in aaa], label="Training")
plt.xlabel("C")
plt.ylabel("Accuracy")
plt.title("Accuracy vs C for training data")
plt.legend()
plt.show()
```



[ ]:
```python
C_best = np.argmax([x[0] for x in aaa]) + 1
print('C =', C_best, 'is the best C for hold-out')
```

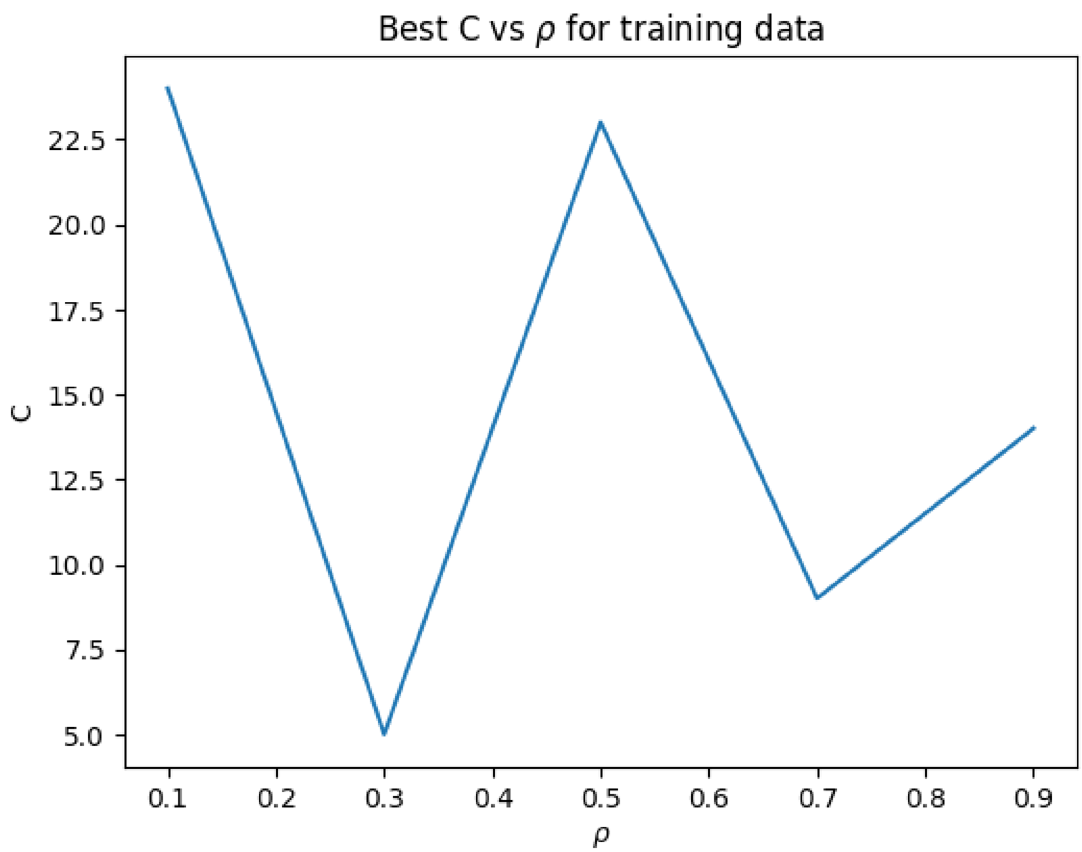
```
C = 15 is the best C for hold-out
```

(b) How is the value of C affected by (percentage of points held out) and number of repetitions? What does a large number of repetitions provide?

[ ]:
```python
# Rho vs C
best_C = []
for rho in [0.1, 0.3, 0.5, 0.7, 0.9]:
    # print("rho = ", rho, ": ", outCVSVM(X_train, Y_train, 5, 10, rho))
    aaa = []
    for C in range(1, 25):
        svmout = outCVSVM(X_train, Y_train, C, n_rep, hold_out_rho)
        aaa.append(svmout)
    best_C.append(1+np.argmax([x[0] for x in aaa]))
    print('Rho = ' + str(rho) + ': C =', np.argmax([x[0] for x in aaa]), 'is the best C for hold-out')
```

```
Rho = 0.1: C = 23 is the best C for hold-out
Rho = 0.3: C = 4 is the best C for hold-out
Rho = 0.5: C = 22 is the best C for hold-out
Rho = 0.7: C = 8 is the best C for hold-out
Rho = 0.9: C = 13 is the best C for hold-out
```

[ ]:
```python
plt.plot([0.1, 0.3, 0.5, 0.7, 0.9], best_C)
plt.xlabel("$\\rho$")
plt.ylabel("C")
plt.title("Best C vs $\\rho$ for training data")
plt.show()
```



[ ]:
```
for n_rep in [1, 5, 10, 20, 30]:
    print("n_rep = ", n_rep, ": ", outCVSVM(X_train, Y_train, 5, n_rep, 0.3))
```

```
n_rep =  1 :  [75.   79.76]
n_rep =  5 :  [73.89 79.76]
n_rep =  10 :  [75.56 80.24]
n_rep =  20 :  [78.75 77.62]
n_rep =  30 :  [77.69 78.41]
```

[ ]:
```
best_C_nrep = []
for n_rep in [1, 5, 10, 20, 30]:
    # print("rho = ", rho, ": ", outCVSVM(X_train, Y_train, 5, 10, rho))
    aaa = []
    for C in range(1, 25):
        svmout = outCVSVM(X_train, Y_train, C, n_rep, 0.3)
        aaa.append(svmout)
    best_C_nrep.append(1 + np.argmax([x[0] for x in aaa]))
    # print('N_rep = ' + str(rho) + ': k = ', np.argmax([x[0] for x in␣
 ↪aaa])*2+1, 'is the best k for hold-out')
```

```
    print('For n = ' + str(n_rep) + ' repetitions: C =',1 + np.argmax([x[0] for␣
↪x in aaa]), 'is the best')
```

```
For n = 1 repetitions: C = 21 is the best
For n = 5 repetitions: C = 9 is the best
For n = 10 repetitions: C = 7 is the best
For n = 20 repetitions: C = 11 is the best
For n = 30 repetitions: C = 6 is the best
```

[ ]:
```
#C vs n_rep
plt.plot([1, 5, 10, 20, 30], best_C_nrep)
plt.xlabel('n_rep')
plt.ylabel('C')
plt.title('Best C vs No.of repetitions for training data')
plt.show()
```

(c) Make a comparison of the performance of the soft-margin SVM with best value of C and kNN with best k (for the current modified dataset), that is, which one performs better.

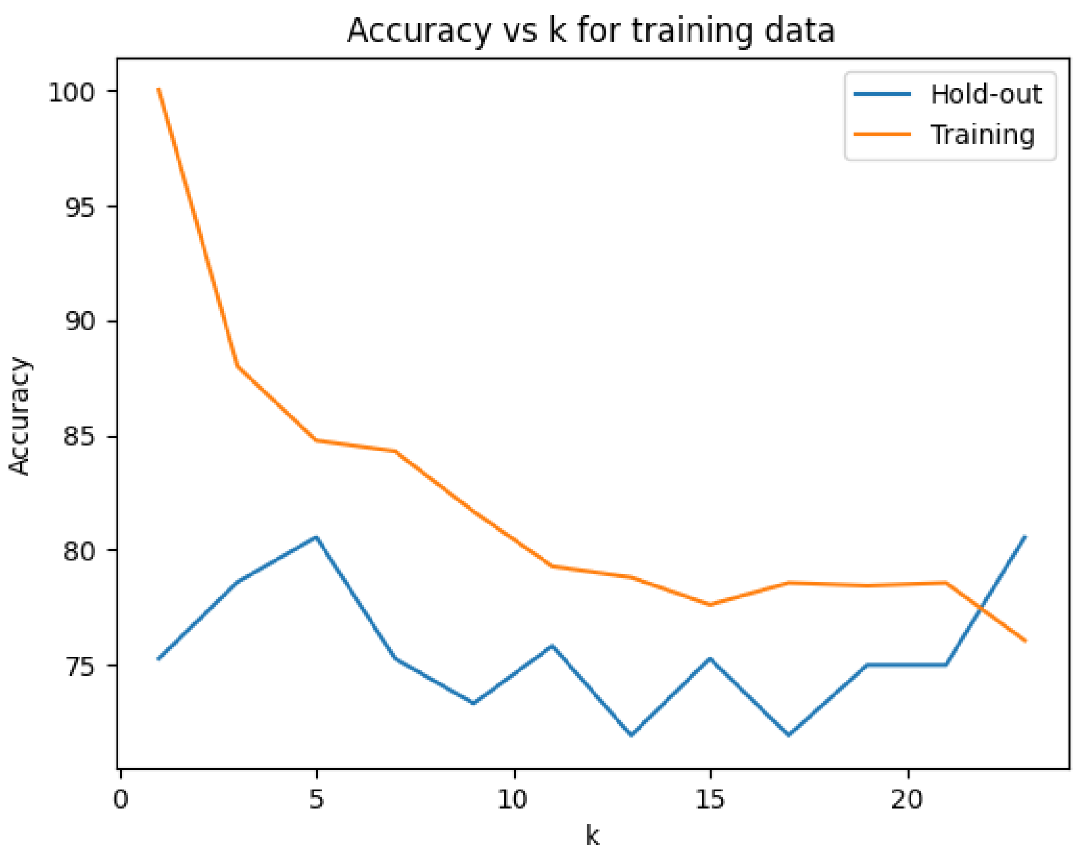
[ ]:
```
def kNNClassify(X_train, Y_train, X_test, n_neighbors):

    distances = np.sqrt(np.sum((X_test[:, np.newaxis, :] - X_train[np.newaxis, :
 ↪, :])**2, axis=2))

    kNN_loc = np.argsort(distances, axis=1)[:, :n_neighbors]

    k_NN_labels = Y_train[kNN_loc]

    Y_pred = np.array([np.argmax(np.bincount(x+1)) for x in k_NN_labels]) # +1␣
 ↪as bincount needs non-negative integers


    return Y_pred-1
```

[ ]:
```
hold_out_rho = 0.3
n_rep = 10
def outCVkNN(X, Y, k, n_rep, hold_out_rho):
    n = len(X)
    n_hold_out = int(n*hold_out_rho)
    hold_out_accuracy = np.zeros(n_rep)
    train_accuracy = np.zeros(n_rep)
    for i in range(n_rep):
        shuffle_idx = np.random.permutation(n)
        X_train = X[shuffle_idx]
        Y_train = Y[shuffle_idx]
        X_hold_out = X_train[:n_hold_out]
        Y_hold_out = Y_train[:n_hold_out]
        X_train = X_train[n_hold_out:]
        Y_train = Y_train[n_hold_out:]
        Y_pred_hold_out = kNNClassify(X_train, Y_train, X_hold_out, k)
        hold_out_accuracy[i] = np.count_nonzero(Y_hold_out == Y_pred_hold_out)/
 ↪len(Y_hold_out)
        Y_pred_train = kNNClassify(X_train, Y_train, X_train, k)
        train_accuracy[i] = np.count_nonzero(Y_train == Y_pred_train)/
 ↪len(Y_train)


    return np.round([np.mean(hold_out_accuracy)*100, np.
 ↪mean(train_accuracy)*100], 2)
```

[ ]:
```
# Training kNN over multiple values of k for holout of 0.3 and 10 repetitions
aaa = []
for k in range(1, 25, 2):
    knnout = outCVkNN(X_train, Y_train, k, n_rep, hold_out_rho)
    print("k = ", k, ": ", knnout)
    aaa.append(knnout)
```

```
k =  1 :  [ 75.28 100.  ]
k =  3 :  [78.61 87.98]
k =  5 :  [80.56 84.76]
k =  7 :  [75.28 84.29]
k =  9 :  [73.33 81.67]
k =  11 :  [75.83 79.29]
k =  13 :  [71.94 78.81]
k =  15 :  [75.28 77.62]
k =  17 :  [71.94 78.57]
k =  19 :  [75.   78.45]
k =  21 :  [75.   78.57]
k =  23 :  [80.56 76.07]
```

[ ]:
```
plt.plot(range(1, 25, 2), [x[0] for x in aaa], label="Hold-out")
plt.plot(range(1, 25, 2), [x[1] for x in aaa], label="Training")
plt.xlabel("k")
plt.ylabel("Accuracy")
plt.title("Accuracy vs k for training data")
plt.legend()
plt.show()
```



[ ]:
```python
k_best = np.argmax([x[0] for x in aaa])*2+1
print('k =', k_best, 'is the best k for hold-out')
```

```
k = 5 is the best k for hold-out
```

[ ]:
```python
#Run kNN on Test data with best k
Y_pred_kNN = kNNClassify(X_train, Y_train, X_test, k_best)
kNN_accuracy = np.count_nonzero(Y_pred_kNN == Y_test)*100 / Y_test.shape[0]
```

[ ]:
```python
#Run SVM on Test data with best C
params = soft_margin_SVM_fit(X_train, Y_train, C = C_best)
params = params.x
Y_pred_SVM = np.sign(np.dot(X_test, params[:-1]) - params[-1])
SVM_accuracy = np.count_nonzero(Y_pred_SVM == Y_test)*100 / Y_test.shape[0]
```

[ ]:
```python
print('kNN accuracy: ', kNN_accuracy)
print('SVM accuracy: ', SVM_accuracy)

if kNN_accuracy > SVM_accuracy:
    print('kNN is better')

else:
    print('SVM is better')
```

```
kNN accuracy:  77.0
SVM accuracy:  80.625
SVM is better
```