Support Vector Machines

In this exercise sheet, you will experiment with training various support vector machines on a subset of the MNIST dataset composed of digits 5 and 6. First, download the MNIST dataset from http://yann.lecun.com/exdb/mnist/, uncompress the downloaded files, and place them in a `data/` subfolder. Install the optimization library *CVXOPT* (`python-cvxopt` package, or directly from the website `www.cvxopt.org`). This library will be used to optimize the dual SVM in part A.

Part A: Kernel SVM and Optimization in the Dual

We would like to learn a nonlinear *SVM* by optimizing its dual. An advantage of the dual SVM compared to the primal SVM is that it allows to use nonlinear kernels such as the Gaussian kernel, that we define as:

$$k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{\sigma^2}\right)$$

The dual SVM consists of solving the following quadratic program:

$$\max_{\alpha} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{ij} a_i a_j y_i y_j k(x_i, x_j)$$

subject to:

$$0 \leq a_i \leq C \qquad \text{and} \qquad \sum_{i=1}^{N} a_i y_i = 0.$$

Then, given the alphas, the prediction of the SVM can be obtained as:

$$f(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} a_i y_i k(x, x_i) + \theta > 0 \\ -1 & \text{if } \sum_{i=1}^{N} a_i y_i k(x, x_i) + \theta < 0 \end{cases}$$

where

$$\theta = \frac{1}{\#SV} \sum_{i \in SV} \left(y_i - \sum_{j=1}^{N} a_j y_j k(x_i, x_j) \right)$$

and `SV` is the set of indices corresponding to the unbound support vectors.

Implementation (25 P)

We will solve the dual SVM applied to the MNIST dataset using the *CVXOPT* quadratic optimizer. For this, we have to build the data structures (vectors and matrices) to must be passed to the optimizer.

- *Implement* a function `gaussianKernel` that returns for a Gaussian kernel of scale σ, the Gram matrix of the two data sets given as argument.
- *Implement* a function `getQPMatrices` that builds the matrices `P, q, G, h, A, b` (of type `cvxopt.matrix`) that need to be passed as argument to the optimizer `cvxopt.solvers.qp`.
- *Run* the code below using the functions that you just implemented. (It should take less than 3 minutes.)

In [27]:
```
import utils.numpy.cvxopt.cvxopt.solvers
import numpy as np
from cvxopt import matrix
#import solutions
```

In [82]:
```
def gaussianKernel(x, xp, scale):

    # build kernel matrix (probably there is a faster way to do this, but not sure how
    K = np.zeros((x.shape[0],xp.shape[0]))

    for i in range(x.shape[0]):
        for j in range(xp.shape[0]):

            K[i,j] = np.exp(-(numpy.linalg.norm(x[i,:]-xp[j,:]))**2/(scale**2))


    return K


def getQPMatrices(Xtrain,Ttrain,C):

    # get n
    n = Ttrain.shape[0]

    # To get p, we can take the Kernel matrix and multiply el
    # q makes sure all alphas are added up,
    # therefore is a matrix consisting of ones with size 1xn
    P = np.outer(Ttrain,Ttrain)*Xtrain

    #print(P)

    q = np.ones(n)*-1
    q = np.expand_dims(q,1)

    #print("P:",P.shape)
    #print("q:",q.shape)

    P = matrix(P,tc='d')
    q = matrix(q,tc='d')

    # h is the constrain matrix. since each alpha needs to be bigger 0 and
    # smaller than C, there are two inequalities for each alpha
    h = np.hstack((np.zeros(n),np.ones(n)*C))
    G = np.vstack((np.eye(n)*-1,np.eye(n)))

    h = np.expand_dims(h,1)

    #print('G:',G.shape)
    #print('h:',h.shape)

    h = matrix(h,tc='d')
    G = matrix(G,tc='d')

    # only the sum of each alpha with the label is a equality condition,
    # that has to be met. Therefore b is 1x1 "matrix" and A is 1xn matrix,
    # with A*alpha being equal to the dot product of A and alpha
    Ttrain = np.expand_dims(Ttrain,0)

    #print("A:",Ttrain.shape)
    A = matrix(Ttrain,tc='d')
    b = matrix(0.0,tc='d')

    return P,q,G,h,A,b


Xtrain,Ttrain,Xtest,Ttest = utils.getMNIST56()

#print(Xtrain.shape)
#print(Ttrain.shape)

cvxopt.solvers.options['show_progress'] = False

for scale in [10,30,100]:
    for C in [1,10,100]:

        # Prepare kernel matrices

        ### TODO: REPLACE BY YOUR OWN CODE
        Ktrain = gaussianKernel(Xtrain,Xtrain,scale)
        Ktest  = gaussianKernel(Xtest,Xtrain,scale)
        ###

        #print(Ktrain.shape)
        #print(Ktest.shape)

        # Prepare the matrices for the quadratic program

        ### TODO: REPLACE BY YOUR OWN CODE
        P,q,G,h,A,b = getQPMatrices(Ktrain,Ttrain,C)
        ###


        # Train the model (i.e. compute the alphas)
        alpha = numpy.array(cvxopt.solvers.qp(P,q,G,h,A,b)['x']).flatten()

        # Get predictions for the training and test set
        SV = (alpha>1e-6)
        uSV = SV*(alpha<C-1e-6)
        theta = 1.0/sum(uSV)*(Ttrain[uSV]-numpy.dot(Ktrain[uSV,:],alpha*Ttrain)).sum()
        Ytrain = numpy.sign(numpy.dot(Ktrain[:,SV],alpha[SV]*Ttrain[SV])+theta)
        Ytest  = numpy.sign(numpy.dot(Ktest [:,SV],alpha[SV]*Ttrain[SV])+theta)

        # Print accuracy and number of support vectors
        Atrain = (Ttrain==Ttrain).mean()
        Atest  = (Ytest ==Ttest ).mean()
        print('Scale=%3d  C=%3d  SV: %4d  Train: %.3f  Test: %.3f'%(scale,C,sum(SV),Atrain,Atest))
```

```
Scale= 10  C=  1  SV: 1000  Train: 1.000  Test: 0.937
Scale= 10  C= 10  SV: 1000  Train: 1.000  Test: 0.937
Scale= 10  C=100  SV: 1000  Train: 1.000  Test: 0.937
Scale= 30  C=  1  SV:  254  Train: 1.000  Test: 0.985
Scale= 30  C= 10  SV:  274  Train: 1.000  Test: 0.986
Scale= 30  C=100  SV:  256  Train: 1.000  Test: 0.986
Scale=100  C=  1  SV:  317  Train: 0.973  Test: 0.971
Scale=100  C= 10  SV:  159  Train: 0.990  Test: 0.975
Scale=100  C=100  SV:  136  Train: 1.000  Test: 0.975
```

Analysis (10 P)

- *Explain* which combinations of parameters σ and C lead to good generalization, underfitting or overfitting?

Good generalization: Scale = 30, C = 10. Because we have good learning (Train: 1.000) and best predictions in the test set (Test: 0.986).

Underfitting: Scale = 100, C = 1. Because we have the lowest prediction rate in the train set (Train: 0.973) and low prediction rate in the test set.

Overfitting: Scale = 10, C = 100. Because we have good flexibility in the train set and lower prediction rate in the test set.

- *Explain* which combinations of parameters σ and C produce the fastest classifiers (in terms of amount of computation needed at prediction time)?

Scale = 100, C = 100 produce the fastest classifier because it uses only 136 support vectors, the least of all parameter sets.

Good generalization: Scale = 30, C = 10. Because we have good learning (Train: 1.000) and best predictions in the test set (Test: 0.986).

Underfitting: Scale = 100, C = 1. Because we have the lowest prediction rate in the train set (Train: 0.973) and low prediction rate in the test set.

Overfitting: Scale = 10, C = 100. Because we have good flexibility in the train set and lower prediction rate in the test set.

Part B: Linear SVMs and Gradient Descent in the Primal

The quadratic problem of the dual SVM does not scale well with the number of data points. For large number of data points, it is generally more appropriate to optimize the SVM in the primal. The primal optimization problem for linear SVMs can be written as

$$\min_{w,\theta} \; ||w||^2 + C\sum_{i=1}^{N} \zeta_i \qquad \text{where} \qquad \forall_{i=1}^{N} : y_i(w \cdot x_i + \theta) \geq 1 - \zeta_i \qquad \text{and} \qquad \zeta_i \geq 0.$$

It is common to incorporate the constraints directly into the objective and then minimizing the unconstrained objective

$$\mathcal{A}(w,\theta) = ||w||^2 + C\sum_{i=1}^{N} \max\left(0, 1 - y_i(w \cdot x_i + \theta)\right)$$

using simple gradient descent.

Implementation (15 P)

- *Implement* the function J computing the objective $\mathcal{A}(w,\theta)$
- *Implement* the function ΔJ computing the gradient of the objective $\mathcal{A}(w,\theta)$ with respect to the parameters w and θ.
- *Run* the code below using the functions that you just implemented. (It should take less than 1 minute.)

In [3]:
```
import utils.numpy as np

def J(w, theta, C, Xtrain, Ttrain):
    tmp = 0

    for i in range(Ttrain.shape[0]):

        tmp = tmp + max(0, 1-Ttrain[i]*(np.dot(Xtrain[i,:],w)+theta))

    return numpy.linalg.norm(w)**2+C*tmp

def DJ(w, theta, C, Xtrain, Ttrain):

    tmp = np.zeros(Xtrain.shape[1])

    for i in range(Ttrain.shape[0]):

        if(1-(Ttrain[i]*(np.dot(Xtrain[i,:],w)+theta)) > 0):

            tmp = tmp - Ttrain[i]*Xtrain[i,:]

    dw = 2 * w + C * tmp

    tmp = 0

    for i in range(Ttrain.shape[0]):

        if(1-(Ttrain[i]*(np.dot(Xtrain[i,:],w)+theta)) > 0):

            tmp = tmp -Ttrain[i]

    dtheta = C*tmp

    return dw, dtheta
```

In [4]:
```
import utils.numpy



C = 10.0
lr = 0.001

Xtrain,Ttrain,Xtest,Ttest = utils.getMNIST56()

n,d = Xtrain.shape

w = numpy.zeros([d])
theta = 1e-9

for it in range(0,101):

    # Monitor the training and test error every 5 iterations
    if it%5==0:
        Ytrain = numpy.sign(numpy.dot(Xtrain,w)+theta)
        Ytest  = numpy.sign(numpy.dot(Xtest ,w)+theta)

        ### TODO: REPLACE BY YOUR OWN CODE
        Obj    = J(w,theta,C,Xtrain,Ttrain)
        ###

        Atrain = (Ytrain==Ttrain).mean()
        Atest  = (Ytest ==Ttest ).mean()
        print('It=%3d   J: %9.3f  Train: %.3f  Test: %.3f'%(it,Obj,Atrain,Atest))

    ### TODO: REPLACE BY YOUR OWN CODE
    dw,dtheta = DJ(w,theta,C,Xtrain,Ttrain)
    ###

    w = w - lr*dw
    theta = theta - lr*dtheta
```

```
It=  0   J: 10000.000  Train: 0.471  Test: 0.482
It=  5   J: 68520.417  Train: 0.961  Test: 0.958
It= 10   J: 49918.674  Train: 0.973  Test: 0.961
It= 15   J: 37473.229  Train: 0.973  Test: 0.963
It= 20   J: 28590.129  Train: 0.974  Test: 0.965
It= 25   J: 21746.877  Train: 0.977  Test: 0.967
It= 30   J: 16987.200  Train: 0.980  Test: 0.968
It= 35   J: 13646.095  Train: 0.986  Test: 0.967
It= 40   J: 11187.127  Train: 0.986  Test: 0.967
It= 45   J:  9182.940  Train: 0.991  Test: 0.967
It= 50   J:  7692.273  Train: 0.990  Test: 0.968
It= 55   J:  6437.609  Train: 0.988  Test: 0.966
It= 60   J:  5253.071  Train: 0.995  Test: 0.966
It= 65   J:  4515.520  Train: 0.992  Test: 0.967
```