Machine Learning 1 WS18/19

Submission for Exercise Sheet 9

Michael Hoppe 362514
Wai Tang Victor Chan 406094
Jonas Piotrowski 399334
Aki Saksala 399293
Sourabh Raj 397371
Farnoush Rezaei Jafari 398708

Exercise Sheet 9

1a) $\Lambda(w,\theta,\alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{N} \alpha_i \left(y_i \left(w^T x_i + \theta\right) - 1\right)$

b) Primal Problem is $\min_{w,\theta} \max_{\alpha \geq 0} \Lambda(w,\theta,\alpha)$

Dual Problem is $\max_{\alpha \geq 0} \min_{w,\theta} \Lambda(w,\theta,\alpha)$

Express $\min_{w,\theta} \Lambda(w,\theta,\alpha)$ in terms of α:

$$\begin{cases} \nabla_w \Lambda(w,\theta,\alpha) = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 & \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i \quad (1) \\ \frac{\partial}{\partial \theta} \Lambda(w,\theta,\alpha) = \sum_{i=1}^{N} \alpha_i y_i = 0 & \quad (2) \end{cases}$$

Putting (1) into $\Lambda(w,\theta,\alpha)$:

$\min_{w,\theta} \Lambda(w,\theta,\alpha) =$

$$= \Lambda(w^*(\alpha), \theta^*(\alpha), \alpha)$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i y_i x_i\right)^T \left(\sum_{j=1}^{N} \alpha_j y_j x_j\right) - \sum_{i=1}^{N} \alpha_i y_i \left(\sum_{j=1}^{N} \alpha_j y_j x_j\right)^T x_i - \theta^*(\alpha) \underbrace{\sum_{i=1}^{N} \alpha_i y_i}_{=0 \text{ by } (2)} + \sum_{i=1}^{N} \alpha_i$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

The Lagrange dual is

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

subject to $\alpha_i \geq 0$ for $1 \leq i \leq N$

and $\sum_{i=1}^{N} \alpha_i y_i = 0$

Suppose α^* is the optimal solution of the dual.

Then optimal solution of the primal is $w^* = \sum_{i=1}^{N} \alpha_i y_i x_i$.

In primal problem, we require $w^T x_i + \theta \geq 1$ for $y_i = 1$
$w^T x_i + \theta \leq -1$ for $y_i = -1$

In particular we wanted $\theta = 1 - \min_{i: y_i = 1} w^T x_i$

and $\theta = -1 - \max_{i: y_i = -1} w^T x_i$

so we pick $\theta^* = -\frac{1}{2}\left(\min_{i: y_i = 1} w^T x_i + \max_{i: y_i = -1} w^T x_i \right)$

(c) Primal:

$$\min_{w, \theta} \|w\|^2$$

s.t. $y_i \left(w^T \Phi(x_i) + \theta \right) \geq 1 \quad \forall\ 1 \leq i \leq N$

Dual:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to $\alpha_i \geq 0$ for $1 \leq i \leq N$

and $\sum_{i=1}^{N} \alpha_i y_i = 0$

2) From ①: Dual:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to $\alpha_i \geq 0$ for $1 \leq i \leq N$

and $\sum_{i=1}^{N} \alpha_i y_i = 0$

Since

$$\operatorname{argmax}_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$= \operatorname{argmin}_{\alpha} \underbrace{\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)}_{(\alpha_1 \cdots \alpha_N)\begin{pmatrix} y_1 y_1 K(x_1,x_1) & y_1 y_2 K(x_1,x_2) & \cdots \\ y_2 y_1 K(x_2,x_1) & \ddots & \\ \vdots & & y_N y_N K(x_N, x_N) \end{pmatrix}\begin{pmatrix}\alpha_1 \\ \vdots \\ \alpha_N\end{pmatrix}} - \sum_{i=1}^{N} \alpha_i$$

$$[P]_{ij} := y_i y_j K(x_i, x_j) \qquad q := -\mathbb{1} = (-1, -1, \ldots, -1)^T$$

$$G := -I \qquad h := \vec{0} = (0, \ldots, 0)^T$$

$$A := (y_1, y_2, \ldots, y_N) \qquad b := 0$$

The output $\vec{x}$ is exactly the $\vec{\alpha}$ we are looking for

Programming Part

December 17, 2018

1 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.

1.1 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:

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

The dual SVM consists of solving the following quadratic program:

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

subject to:

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

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

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

where

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

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

1.1.1 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 [8]:
```
import utils,numpy,cvxopt,cvxopt.solvers

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

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

def gaussianKernel(X, XP, Scale):
    distance = -2 * numpy.dot(X, XP.T) + numpy.sum(XP ** 2, axis=1) + numpy.sum(X ** 2

    K = numpy.exp(-distance / (Scale ** 2))
    return K

def getQPMatrices(Ktrain, Ttrain, C):
    n = len(Ttrain)
    P = cvxopt.matrix(numpy.outer(Ttrain, Ttrain) * Ktrain)
    q = cvxopt.matrix(numpy.zeros(n)-1)
    A = cvxopt.matrix(Ttrain, (1, n))
    b = cvxopt.matrix(0.0)
    G = cvxopt.matrix(numpy.vstack((numpy.diag(numpy.repeat(-1, n)), numpy.identity(n)
    h = cvxopt.matrix(numpy.hstack((numpy.zeros(n), numpy.repeat(C, n))))
    return P, q, G, h, A, b

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)
        ###


        # 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 = (Ytrain==Ttrain).mean()
        Atest  = (Ytest ==Ttest ).mean()
        print('Scale=%3d  C=%3d  SV: %4d  Train: %.3f  Test: %.3f'%(scale,C,sum(SV),At
    print('')
```

```
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
```

1.1.2 Analysis (10 P)

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

With large σ or variance of the gaussian kernel, the border curve we obtain has less curvature, while small C may also prevents the function from reaching optimal value when the optimal alpha would have been bigger. Therefore, σ and small C lead to underfitting. On the other hand, the border curves have greater versatility with small σ, which leads to overfitting. With $\sigma = 30$ and $C = 10$ or 100, we seem to get the best accuracy in test data set and it is a sign of good generalization.

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

The number of computations for θ depends on the number of support vectors.We can see that the number of support vectors decreases when σ or C increases. It is because we tend to have more samples lying close to the border line in better fitted model.

1.2 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

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

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

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

using simple gradient descent.

1.2.1 Implementation (15 P)

- *Implement* the function `J` computing the objective $J(w,\theta)$
- *Implement* the function `DJ` computing the gradient of the objective $J(w,\theta)$ 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 [5]: import utils,numpy

        C = 10.0
        lr = 0.001

        def J(w, theta, C, X, T):
            Xi = numpy.matmul(w, numpy.transpose(X))
            Xi = [xi + theta for xi in Xi]
            Xi = T * Xi
            Xi = [max(0, 1 - xi) for xi in Xi]
            J = sum (w * w) + C * sum (Xi)
            return J

        def DJ(w, theta, C, X, T):
            dw = 2 * w
            dtheta = 0
            for i in range (len(T)):
                if T[i] * (sum(w * X[i]) + theta) <= 1:
                    dw = dw - C * T[i] * X[i]
                    dtheta = dtheta - C * T[i]
```

```
        return dw, dtheta


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)
        ###

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

    ### 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
It= 70   J:  4016.851  Train: 0.996  Test: 0.966
It= 75   J:  3647.983  Train: 0.997  Test: 0.965
It= 80   J:  3497.204  Train: 0.998  Test: 0.966
```

```
It= 85   J:  3404.280  Train: 1.000  Test: 0.966
It= 90   J:  3336.804  Train: 1.000  Test: 0.966
It= 95   J:  3270.665  Train: 1.000  Test: 0.966
It=100   J:  3205.837  Train: 1.000  Test: 0.966
```

```
In [ ]:
```