Exercises for the course
Machine Learning 1
Winter semester 2021/22

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 8

Exercise 1: Dual formulation of the Soft-Margin SVM (5 + 20 + 10 + 5 P)

The primal program for the linear soft-margin SVM is

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to

$$\forall_{i=1}^{N}: \ y_i \cdot (\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

where $\|.\|$ denotes the Euclidean norm, ϕ is a feature map, $\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}$ are the parameter to optimize, and $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ are the labeled data points regarded as fixed constants. Once the hard-margin SVM has been learned, prediction for any data point $\boldsymbol{x} \in \mathbb{R}^d$ is given by the function

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^\top \phi(\boldsymbol{x}) + b).$$

(a) *State* the conditions on the data under which a solution to this program can be found from the Lagrange dual formulation *(Hint: verify the Slater's conditions).*

(b) *Derive* the Lagrange dual and show that it reduces to a constrained quadratic optimization problem. State both the objective function and the constraints of this optimization problem.

(c) *Describe* how the solution $(\boldsymbol{w}, b)$ of the primal program can be obtained from a solution of the dual program.

(d) *Write* a kernelized version of the dual program and of the learned decision function.

Exercise 2: SVMs and Quadratic Programming (10 P)

We consider the `CVXOPT` Python software for convex optimization. The method `cvxopt.solvers.qp` solves quadratic optimization problems given in the matrix form:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^\top P\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} \\ \text{subject to} \quad & G\boldsymbol{x} \preceq \boldsymbol{h} \\ \text{and} \quad & A\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

Here, $\preceq$ denotes the element-wise inequality: $(\boldsymbol{h} \preceq \boldsymbol{h}') \Leftrightarrow (\forall_i : h_i \leq h'_i)$. Note that the meaning of the variables $\boldsymbol{x}$ and $\boldsymbol{b}$ is different from that of the same variables in the previous exercise.

(a) *Express* the matrices and vectors $P, \boldsymbol{q}, G, \boldsymbol{h}, A, \boldsymbol{b}$ in terms of the variables of Exercise 1, such that this quadratic minimization problem corresponds to the kernel dual SVM derived above.

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Kernel Support Vector Machines

In this exercise sheet, we will implement a kernel SVM. Our implementation will be based on a generic quadratic programming optimizer provided in CVXOPT (`python-cvxopt` package, or directly from the website `www.cvxopt.org`). The SVM will then be tested on the UCI breast cancer dataset, a simple binary classification dataset accessible via the `scikit-learn` library.

1. Building the Gaussian Kernel (5 P)

As a starting point, we would like to implement the Gaussian kernel, which we will make use of in our kernel SVM implementation. It is defined as:

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

- **Implement a function `getGaussianKernel` that returns for a Gaussian kernel of scale σ, the Gram matrix of the two data sets given as argument.**

In [1]:

```
import numpy,scipy,scipy.spatial

def getGaussianKernel(X1,X2,scale):
    ### TODO: REPLACE BY YOUR OWN CODE
    import solutions
    K = solutions.getGaussianKernel(X1,X2,scale)
    return K
    ###
```

2. Building the Matrices for the CVXOPT Quadratic Solver (20 P)

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. The dual SVM consists of solving the following quadratic program:
$\max$

We would like to rely on a CVXOPT solver to obtain a solution to our SVM dual. The function `cvxopt.solvers.qp` solves an optimization problem of the type:
\begin{align*} \min_{\boldsymbol{x}} \quad &\frac12 \boldsymbol{x}^\top P \boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x}\\ \text{subject to} \quad & G \boldsymbol{x} \preceq \boldsymbol{h}\\ \text{and} \quad & A \boldsymbol{x} = \boldsymbol{b}. \end{align*}

which is of similar form to our dual SVM (note that \boldsymbol{x} will correspond to the parameters (\alpha_i)_i of the SVM). We need to build the data structures (vectors and matrices) that makes solving this quadratic problem equivalent to solving our dual SVM.

- **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`.**

In [2]:

```
import cvxopt,cvxopt.solvers
cvxopt.solvers.options['show_progress'] = False

def getQPMatrices(K,T,C):
    ### TODO: REPLACE BY YOUR OWN CODE
    import solutions
    P,q,G,h,A,b = solutions.getQPMatrices(K,T,C)
    return P,q,G,h,A,b
    ###
```

3. Computing the Bias Parameters (10 P)

Given the parameters (\alpha_i)_i the optimization procedure has found, the prediction of the SVM is given by:
f(x) = \text{sign}\Big(\sum_{i=1}^N \alpha_i y_i k(x,x_i) + \theta\Big)

Note that the parameter \theta has not been computed yet. It can be obtained from any support vector that lies exactly on the margin, or equivalently, whose associated parameter \alpha is not equal to 0 or C. Calling one such vector "x_M", the parameter \theta can be computed as:
\theta = y_M - \sum_{j=1}^N \alpha_j y_j k(x_M,x_j)

- **Implement a function `getTheta` that takes as input the Gram Matrix used for training, the label vector, the solution of our quadratic program, and the hyperparameter C. The function should return the parameter \theta.**

In [3]:

```
def getTheta(K,T,alpha,C):
    ### TODO: REPLACE BY YOUR CODE
    import solutions
    theta = solutions.getTheta(K,T,alpha,C)
    return theta
    ###
```

4. Implementing a class `GaussianSVM` (15 P)

All functions that are needed to learn the SVM have now been built. We would like to implement a SVM class that connects them and make the SVM easily usable. The class structure is given below and contains two functions, one for training the model, and one for applying it to test data.

- **Implement the function `fit` that makes use of the functions `getGaussianKernel`, `getQPMatrices`, `getTheta` you have already implemented. The function should learn the SVM model and store the support vectors, their label, (\alpha_i)_i and \theta into the object (`self`).**
- **Implement the function `predict` that makes use of the stored information to compute the SVM output for any new collection of data points**

In [4]:

```
class GaussianSVM:

    def __init__(self,C=1.0,scale=1.0):

        self.C, self.scale = C, scale

    def fit(self,X,T):

        ### TODO: REPLACE BY YOUR CODE
        import solutions
        solutions.fit(self,X,T)
        ###

    def predict(self,X):

        ### TODO: REPLACE BY YOUR CODE
        import solutions
        Y = solutions.predict(self,X)
        return Y
        ###
```

5. Analysis

The following code tests the SVM on some breast cancer binary classification dataset for a range of scale and soft-margin parameters. For each combination of parameters, we output the number of support vectors as well as the train and test accuracy averaged over a number of random train/test splits. Running the code below should take approximately 1-2 minutes.

In [5]:

```
import numpy,sklearn,sklearn.datasets,numpy

D = sklearn.datasets.load_breast_cancer()
X = D['data']
T = D['target']
T = (D['target']==1)*2.0-1.0

for scale in [30,100,300,1000,3000]:
    for C in [10,100,1000,10000]:

        acctrain,acctest,nbsvs = [],[],[]

        svm = GaussianSVM(C=C,scale=scale)

        for i in range(10):

            # Split the data
            R = numpy.random.mtrand.RandomState(i).permutation(len(X))
            Xtrain,Xtest = X[R[:len(R)//2]]*1,X[R[len(R)//2:]]*1
            Ttrain,Ttest = T[R[:len(R)//2]]*1,T[R[len(R)//2:]]*1

            # Train and test the SVM
            svm.fit(Xtrain,Ttrain)
            acctrain += [(svm.predict(Xtrain)==Ttrain).mean()]
            acctest  += [(svm.predict(Xtest)==Ttest).mean()]
            nbsvs += [len(svm.X)*1.0]

        print('scale=%9.1f  C=%9.1f  nSV: %4d  train: %.3f  test: %.3f'%(
            scale,C,numpy.mean(nbsvs),numpy.mean(acctrain),numpy.mean(acctest)))
    print('')
```

```
scale=     30.0  C=     10.0  nSV:  183  train: 0.997  test: 0.921
scale=     30.0  C=    100.0  nSV:  178  train: 1.000  test: 0.918
scale=     30.0  C=   1000.0  nSV:  184  train: 1.000  test: 0.918
scale=     30.0  C=  10000.0  nSV:  182  train: 1.000  test: 0.918

scale=    100.0  C=     10.0  nSV:  117  train: 0.965  test: 0.935
scale=    100.0  C=    100.0  nSV:   97  train: 0.987  test: 0.940
scale=    100.0  C=   1000.0  nSV:   85  train: 0.998  test: 0.932
scale=    100.0  C=  10000.0  nSV:   71  train: 1.000  test: 0.926

scale=    300.0  C=     10.0  nSV:   88  train: 0.939  test: 0.924
scale=    300.0  C=    100.0  nSV:   48  train: 0.963  test: 0.943
scale=    300.0  C=   1000.0  nSV:   36  train: 0.978  test: 0.946
scale=    300.0  C=  10000.0  nSV:   32  train: 0.991  test: 0.941

scale=   1000.0  C=     10.0  nSV:   66  train: 0.926  test: 0.916
scale=   1000.0  C=    100.0  nSV:   55  train: 0.935  test: 0.929
scale=   1000.0  C=   1000.0  nSV:   49  train: 0.956  test: 0.946
scale=   1000.0  C=  10000.0  nSV:   38  train: 0.971  test: 0.951

scale=   3000.0  C=     10.0  nSV:   87  train: 0.912  test: 0.903
scale=   3000.0  C=    100.0  nSV:   68  train: 0.926  test: 0.919
scale=   3000.0  C=   1000.0  nSV:   58  train: 0.934  test: 0.929
scale=   3000.0  C=  10000.0  nSV:   49  train: 0.953  test: 0.943
```

We observe that the highest accuracy is obtained with a scale parameter that is neither too small nor too large. Best parameters are also often associated to a low number of support vectors.

Sheet 8

1

a) The solution can be found from the Lagrange dual formulation if strong duality holds for this convex optimization problem. Strong duality can be confirmed by verifying Slater's conditions:

I Slater condition 1:

All convex inequality constraints are fulfilled for at least one point $\tilde{\underline{x}} \in \mathbb{R}^d$:

$$\forall_{i=1}^{N} \quad y_i (\underline{w}^T \phi(\underline{\tilde{x}}) + b) \geq 1 - \xi_i$$

II Slater condition 2:

All affine equality and inequality constraints are fulfilled for at least one point $\tilde{\underline{x}}$:

$$\forall_{i=1}^{N} \quad \xi_i \geq 0$$

II is easy to verify because we can simply choose all ξ_i such that $\xi_i \geq 0$ for all i.

I Similarly, $y_i (\underline{w}^T \phi(\underline{\tilde{x}}) + b) \geq 1 - \xi_i$ always holds if we choose all ξ_i very large: $\forall_{i=1}^{N} \quad \xi_i >> 1$.

4) The primal is given by $\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$

We can add the constraints by making them zero:

$\forall_{i=1}^{N}\ y_i(\underline{w}^T \phi(\underline{x}_i) + b) \geq 1 - \xi_i$

$\Leftrightarrow \quad 0 \geq 1 - \xi_i - y_i(\underline{w}^T \phi(\underline{x}_i) + b)$

$\forall_{i=1}^{N}\ \xi_i \geq 0 \quad \Leftrightarrow \quad 0 \geq -\xi_i$

Therefore, the dual is obtained maximizing the Lagrange multiplier λ, λ' for both constraints on the primal:

$$\text{Dual} = \max_{\lambda,\lambda' \geq 0} \quad \text{Primal} = \max_{\lambda,\lambda' \geq 0} \min_{w,\beta,\xi} \underbrace{f(w,\beta,\xi)}_{\text{objective}}$$

$$= \max_{\lambda,\lambda' \geq 0} \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \lambda_i \left(1 - \xi_i - y_i(\underline{w}^T \phi(\underline{x}_i) + b)\right) - \sum_{i=1}^{N} \lambda_i' \xi_i$$

We minimize the objective by taking the partial derivatives:

$$\frac{\partial f}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i \phi(x_i) \overset{!}{=} 0 \quad \Leftrightarrow w = \sum_{i=1}^{N} \lambda_i y_i \phi(x_i)$$

$$\frac{\partial f}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i \overset{!}{=} 0$$

$$\frac{\partial f}{\partial \xi} = C - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i'$$

$$\Leftrightarrow \frac{\partial f}{\partial \xi_i} = C - \lambda_i - \lambda_i' \rightarrow 0 \leq \lambda_i \leq C$$

(box constraint)

Now we obtain the dual in the simplified form

$$\text{Dual} = \max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \phi(x_i)^T\phi(x_j) + \sum_{i=1}^{N}\lambda_i$$

with the constraints $\sum_{i=1}^{N}\lambda_i y_i = 0$ and $\forall_{i=1}^{N}\ 0 \le \lambda_i \le C$

c) From b) we have $w = \sum_{i=1}^{N}\lambda_i y_i \phi(x_i)$ and we know

$$\lambda_i(1-\xi_i-y_i(w^T\phi(x_i)+b)) = 0 \quad \text{and}$$

$$(C-\lambda_i)(-\xi_i) = 0$$

Since $\lambda_i \neq C \Rightarrow \xi_i = 0$

and $\lambda_i \neq 0 \Rightarrow 1 - w^T\phi(x_i) - b = 0$ } if we define $y_i := 1$

Using $y_i := 1$ and the constraint of b) $0 \le \lambda_i \le C$ we obtain

$$b = 1 - w^T\phi(x_i)$$

d) We are given the function $f(x)$ and must w, b, and the kernel:

$$f(x) = \text{sign}(w^T\phi(x) + b)$$

$$= \text{sign}\left(\sum_{i=1}^{N}\lambda_i y_i k(x_i, x) + 1 - \sum_{i=1}^{N}\lambda_i y_i k(x_i, x_{SV})\right)$$

2)

a) In 1.b) we found

$$\max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j \phi(x_i)^T\phi(x_j) + \sum_{i=1}^{N}\lambda_i$$

with $\sum_{i=1}^{N}\lambda_i y_i = 0$ and $0 \le \lambda_i \le C$

We can rewrite it such that

$$\min_{\lambda} \frac{1}{2}\sum \lambda_i y_i y_j k(x_i,x_j)\lambda_j - \sum_{i=1}^{N}\lambda_i$$

$$= \min_{\lambda} \frac{1}{2}\underline{\lambda}^T P_{ij}\underline{\lambda} - \underline{1}\cdot\underline{\lambda} \iff \min_{x} \frac{1}{2}\underline{x}^T\underline{\underline{P}}\underline{x} + \underline{q}^T\underline{x}$$

Therefore $\boxed{\underline{x} \hat{=} \underline{\lambda},\ \underline{\underline{P}} \hat{=} y_i y_j k(x_i,x_j),\ \underline{q}^T \hat{=} -\underline{1}}$

The constraint $\sum_{i=1}^{N}\lambda_i y_i = 0$ simplifies to

$$\underline{\lambda}\,\underline{y}^T = \underline{y}^T\underline{\lambda} = 0 \iff A\underline{x} = \underline{b}$$

Therefore $\boxed{A \hat{=} \underline{y}^T,\ \underline{b} \hat{=} \underline{0}}$.

The constraint $0 \le \lambda_i \le C$ corresponds to

$$\left.\begin{matrix} -\lambda_i \le 0 \\ \lambda_i \le C \end{matrix}\right\} = \begin{pmatrix} -\underline{1} \\ \underline{1} \end{pmatrix}\cdot\lambda \le \begin{pmatrix} 0 \\ C \end{pmatrix}\underline{1}$$

Therefore, $\boxed{G \hat{=} \begin{pmatrix} -\underline{1} \\ \underline{1} \end{pmatrix},\ h \hat{=} \begin{pmatrix} 0 \\ C \end{pmatrix}\underline{1}}$