Convex Optimization Theory
Notes on Regularization and Support Vector Machines

June 1, 2021

Sonia Gutierrez Luna *
Abinaya Jayakumar *
Juan Francisco Muñoz Elguezabal *

* Msc in Data Science - ITESO

1 Excercise 1

1.1 Part a

Show that the likelihood function for this problem is:

$$L(y|\boldsymbol{X},\boldsymbol{\theta},\sigma^2) \propto (\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}(y-\boldsymbol{X\theta})^T(y-\boldsymbol{X\theta})\right) \tag{1}$$

and find the parameter vector $\boldsymbol{\theta}$ that minimizes L.

$$-\frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(y-X\theta)^T(y-X\theta) \tag{2}$$

Finding $\frac{dl}{d\theta}$, we have $\frac{dl}{d\theta} = -\frac{1}{2\sigma^2} \propto (Y-\boldsymbol{X\theta})^T(-\boldsymbol{X})$. And for $\frac{dl}{d\theta} = 0$

$$\begin{gathered} -\tfrac{1}{2\sigma^2} \propto (Y-\boldsymbol{X\theta})^T(-\boldsymbol{X}) = 0 \\ (Y-\boldsymbol{X\theta})^T = 0 \\ (Y)^T - (\boldsymbol{X\theta})^T = 0 \\ \boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{Y} \end{gathered}$$

1.2 Part b

Propose a prior normal distribution for θ, find the posterior distribution for θ, and expose in a very detailed way the conditions in such formulatio is equivalent to the ridge regularization.

$$\underset{\theta}{argmax} = \left[log\prod_{i=1}^{N}\frac{1}{\alpha\sqrt{2\pi}}e^{-\frac{(y_i-(\theta_0+\theta_1x_1+\ldots+\theta_nx_n))^2}{2\sigma^2}} + log\prod_{j=1}^{P}\frac{1}{\sqrt{\alpha 2\pi}}e^{-\frac{P^2}{\pi^2}}\right] \tag{3}$$

$$\underset{\theta}{argmax} = \left[\sum_{i=1}^{N}\frac{(y_i-(\theta_0+\theta_1x_1+\ldots+\theta_Nx_N))^2}{2\sigma^2} - \sum_{j=1}^{P}\frac{\theta_j^2}{2\pi^2}\right] \tag{4}$$

$$\underset{\theta}{argmin} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{N}\frac{(y_i-(\theta_0+\theta_1x_1+\ldots+\theta_Nx_N))^2}{2\sigma^2} + \frac{\sigma^2}{2\pi^2}\sum_{j=1}^{P}\theta_j^2\right] \tag{5}$$

$$\underset{\theta}{argmin} = \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}(y_i-(\theta_0+\theta_1x_1+\ldots+\theta_Nx_N))^2 + \lambda\sum_{j=0}^{N}\theta_j^2\right] \tag{6}$$

with $\lambda = \frac{\sigma^2}{2\pi^2}$, ridge regularization is given by:

$$\hat{\theta}_{L2} = \underset{\theta}{argmin}\left[\sum_{i=1}^{N}(y_i-(\theta_0+\theta_1x_1+\ldots+\theta_Nx_N))^2 + \lambda\sum_{j=0}^{N}|\theta_j|^2\right] \tag{7}$$

1.3 Part c

Propose a prior Laplace distribution for θ, find the posterior distribution for θ, and expose in a very detailed way the conditions in such formulation is equivalent to the LASSO regularization.

$$\underset{\theta}{argmax} = \left[log \prod_{i=1}^{N} \frac{1}{\alpha\sqrt{2\pi}} e^{-\frac{(y_i-(\theta_0+\theta_1 x_1+\ldots+\theta_n x_n))^2}{2\sigma^2}} + log \prod_{j=1}^{P} \frac{e^{-\frac{\theta_j}{2b}}}{2b} \right] \tag{8}$$

$$\underset{\theta}{argmax} = \left[-\sum_{i=1}^{P} \frac{(y_i - (\theta_0 + \theta_1 x_1 + ... + \theta_n x_n))^2}{2\sigma^2} - \sum_{j=0}^{P} \frac{|\theta|}{2b} \right] \tag{9}$$

$$\underset{\theta}{argmax} = \left[-\sum_{i=1}^{N} (y_i - (\theta_0 + \theta_1 x_1 + ... + \theta_n x_n))^2 + \frac{\sigma^2}{b} \sum_{j=0}^{P} |\theta_j| \right] \tag{10}$$

$$\underset{\theta}{argmax} = \left[-\sum_{i=1}^{N} (y_i - (\theta_0 + \theta_1 x_1 + ... + \theta_n x_n))^2 + \lambda \sum_{j=0}^{P} |\theta_j| \right] \tag{11}$$

with $\lambda = {}^{\sigma^2}/_{b}$ we have the LASSO regression.

$$\hat{\theta}_{L1} = \underset{\theta}{argmin} \left[-\sum_{i=1}^{N} (y_i - (\theta_0 + \theta_1 x_1 + ... + \theta_n x_n))^2 + \lambda \sum_{j=0}^{P} |\theta_j| \right] \tag{12}$$

2 Excercise 2

Consider the following optimization problem:

$$\min_{w,b,e} \quad P(w,e) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 \quad s.t. \quad y_k = w^t\varphi(x_k) + b + e_k, \quad k = 1,2,...,N \tag{13}$$

where $\{x_k, y_k\}_{k=1}^N$ represents a training set with input data $x_k \in \mathbb{R}^n$, and the output data given by the labels $y_k \in \mathbb{R}$, $e_k \in \mathbb{R}^n$, and the feature maps have the form $\phi(.) : \mathbb{R}^n \to \mathbb{R}^m$. Then, the models parameters are $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$. Finally, $\gamma > 0$.

2.1 Part a

Show that the Lagrangian of the problem (1) is given by:

$$\mathcal{L}(w,b,e;\alpha) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N}\alpha_k[w^T\varphi(x_k) + b + e_k - y_k] \tag{14}$$

The constrain can be written as the following:

$$e_k = y_k - [w^t\varphi(x_k) + b] \quad k = 1,2,...,N \tag{15}$$

and the function to minimize is:

$$P(w,e) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 \tag{16}$$

Therefore, the Lagrangian formulation of the problem is:

$$\boxed{\mathcal{L}(w,b,e;\alpha) = \frac{1}{2}w^T w + \gamma\frac{1}{2}\sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N}\alpha_k[w^T\varphi(x_k) + b + e_k - y_k]} \tag{17}$$

2.2 Part b

Given the lagrangian formulation of the problem in (ref), we define the first order conditions by:

$$\frac{\partial L}{\partial w} = 0 \quad , \quad \frac{\partial L}{\partial b} = 0 \quad , \quad \frac{\partial L}{\partial e_k} = 0 \quad , \quad \frac{\partial L}{\partial \alpha_k} = 0$$

$$\frac{\partial L}{\partial w} = w - \sum_{k=1}^{N}\alpha_k\varphi(x_k) \quad \to \quad \frac{\partial L}{\partial w} = 0 \quad \to \quad w - \sum_{k=1}^{N}\alpha_k\varphi(x_k) = 0$$

$$\boxed{w = \sum_{k=1}^{N}\alpha_k\varphi(x_k)} \tag{18}$$

$$\frac{\partial L}{\partial b} = \sum_{k=1}^{N}\alpha_k \quad \to \quad \frac{\partial L}{\partial b} = 0$$

$$\boxed{\sum_{k=1}^{N}\alpha_k = 0} \tag{19}$$

$$\frac{\partial L}{\partial e_k} = e_k\gamma - \alpha_k \quad \to \quad \frac{\partial L}{\partial e_k} = 0 \quad \to \quad e_k\gamma - \alpha_k = 0$$

$$\boxed{e_k = {}^{\alpha_k}/_{\gamma} \quad k = 1,2,....,N} \tag{20}$$

$$\frac{\partial L}{\partial \alpha_k} = w^T \varphi(x_k) + b + e_k - y_k \quad \rightarrow \quad \frac{\partial L}{\partial \alpha_k} = 0$$

$$\boxed{w^T \varphi(x_k) + b + e_k - y_k = 0 \quad k = 1, 2, ..., N} \tag{21}$$

2.3 Part c

Define adequate vector variables, such that the optimization problem reduces to a set of linear equations which must be solved for α and b. if we replace 18 and 19 in 20, we have:

$$\sum_{k,l=1}^{N} \alpha_k \varphi(x_k)^T \varphi(x_l) + b + \frac{x_l}{\gamma} - y_l = 0 \tag{22}$$

If we define $k, l = 1, 2, ..., N, \quad y = [y_1, y_2, ..., y_N]^T, \quad 1_v = [1, 1, ..., 1]^T, \quad \alpha[\alpha_1, \alpha_2, ..., \alpha_N]^T$. And K such that:

$$K = \begin{bmatrix} K(x_1, x_1) & ... & K(x_N, x_1) \\ \vdots & & \vdots \\ K(x_1, x_N) & ... & K(x_N, x_N) \end{bmatrix}$$

if we define, $y^T \alpha = 0$, $by + (K + \frac{I}{\gamma})\alpha = 1_v$, $1_v \alpha = 0$ we have the following equation system:

$$K\alpha + 1_v b + \frac{\alpha}{\gamma} - y = 0 \quad \rightarrow \quad (K + \frac{I}{\gamma})\alpha + 1_v y = y \tag{23}$$

which can be expressed as the following:

$$\boxed{\left[\begin{array}{c|c} 0 & 1_v \\ \hline 1_v & K + I/\gamma \end{array}\right] \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_v \end{bmatrix}}$$

2.4 Part d

Finally, explain (with math) how the optimization problem (1) is related to the regression problem.

2.4.1 KKT conditions

- Stationary condition:

  $\nabla_w L(w, b; \alpha) = 0 \quad \rightarrow \frac{\partial L}{\partial b} = 0$

- Primal feasibility condition:

  $w^T \varphi(x_k) + b + e_k - y_k = \ 0, \quad k = 1, 2, ..., N$

- Dual feasibility condition:

  $\alpha_k \geq 0, \quad k = 1, 2, ..., N$

- Complementary slackness condition:

  $\alpha_k[(w^T \varphi(x_k) + b + e_k - y_k) - 1] = 0 \quad k = 1, 2, ..., N$

From the last three KKT conditions it is observed that almost all α will be support vectors.

3 Excercise 3

Consider the following optimization problem:

$$\underset{w,b,e}{min} \quad P(w,e) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 \quad s.t. \quad y_k \left[w^t \varphi(x_k) + b\right] = 1 - e_k, \quad k = 1, 2, ..., N \tag{24}$$

where $y_k \in \{-1, +1\}$, is the response (target) variable.

3.1 Lagrangian

the function to minimize is:

$$P(w,e) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 \tag{25}$$

we can observe that the expresion for the restriction is the following

$$y_k \left[w^t \varphi(x_k) + b\right] = 1 - e_k, \quad k = 1, 2, ..., N \quad \rightarrow \quad \sum_{k=1}^{N} \alpha_k (y_k [w^T \varphi(x_k) + b] - 1 + e_k)$$

and that is defined given that the problem is subjeto to N restrictions. therefore, the lagrangian formulation of the problem is the following:

$$\boxed{\mathcal{L}(w,b,e,\alpha) = \frac{1}{2}w^T w + \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 - \sum_{k=1}^{N} \alpha_k (y_k [w^T \varphi(x_k) + b] - 1 + e_k)} \tag{26}$$

3.2 First order conditions

Given the lagrangian formulation of the last problem , we define the first order conditions by

$$\frac{\partial L}{\partial w} = 0 \quad , \quad \frac{\partial L}{\partial b} = 0 \quad , \quad \frac{\partial L}{\partial e_k} = 0 \quad , \quad \frac{\partial L}{\partial \alpha_k} = 0$$

the first order conditions are the following:

$$\frac{\partial L}{\partial w} = w - \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) \quad \rightarrow \quad \frac{\partial L}{\partial w} = 0 \quad \rightarrow \quad w - \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k) = 0$$

$$\boxed{w = \sum_{k=1}^{N} \alpha_k y_k \varphi(x_k)} \tag{27}$$

$$\frac{\partial L}{\partial b} = \sum_{k=1}^{N} \alpha_k y_k \quad \rightarrow \quad \frac{\partial L}{\partial b} = 0$$

$$\boxed{\sum_{k=1}^{N} \alpha_k y_k = 0} \tag{28}$$

$$\frac{\partial L}{\partial e_k} = e_k \gamma - \alpha_k \quad \rightarrow \quad \frac{\partial L}{\partial e_k} = 0 \quad \rightarrow \quad e_k \gamma - \alpha_k = 0$$

$$\boxed{e_k = \alpha_k / \gamma} \tag{29}$$

$$\frac{\partial L}{\partial \alpha_k} = y_k [w^T \varphi(x_k) + b] - 1 + e_k \quad \rightarrow \quad \frac{\partial L}{\partial \alpha_k} = 0$$

$$\boxed{y_k [w^T \varphi(x_k) + b] - 1 + e_k = 0} \tag{30}$$

3.3 Vector variables

Given that $k(x_k, x_l) = \varphi^T(x_k)\varphi(x_l) \quad k = 1, 2, ..., N$, and that we can replace (2), (3), and (4) in (5), we have the following:

$$y_k \left[\sum_{l=1}^{N} \alpha_l y_l \varphi^T(x_l)\varphi(x_k) + b \right] - 1 + \frac{\alpha_k}{\gamma} = 0 \quad k = 1, 2, ..., N$$

$$\sum_{k,l=1}^{N} [y_k \alpha_l y_l \varphi^T(x_l)\varphi(x_k) + b y_k] - 1 + \frac{\alpha_k}{\gamma} = 0$$

So we can define Ω as the following matrix:

$$\Omega = \begin{bmatrix} y_1\varphi^T(x_1)\varphi(x_1)y_1 & ... & y_N\varphi^T(x_N)\varphi(x_1)y_N \\ \vdots & & \vdots \\ y_1\varphi^T(x_1)\varphi(x_N)y_N & ... & y_N\varphi^T(x_N)\varphi(x_N)y_N \end{bmatrix}$$

If we define $k, l = 1, 2, ..., N, \quad y = [y_1, y_2, ..., y_N]^T, \quad 1_v = [1, 1, ..., 1]^T, \quad \alpha[\alpha_1, \alpha_2, ..., \alpha_N]^T$. We can rewrite as:

3.4 KKT Matrix System

$$\Omega\alpha + by + \frac{\alpha}{\gamma} - 1_v = 0 \tag{31}$$

which leads to

$$\left(\Omega + \frac{I}{\gamma}\right)\alpha + by - 1_v = 0$$

$$y^T\alpha = 0$$

and can finally be written as the following:

$$\boxed{\left[\begin{array}{c|c} 0 & y^T \\ \hline y & \Omega + I/\gamma \end{array}\right] \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_v \end{bmatrix}}$$

with the KKT conditions:

- Stationary condition:

  $\nabla_w L(w, b; \alpha) = 0 \quad \rightarrow \frac{\partial L}{\partial b} = 0$

- Primal feasibility condition:

  $y_k[w^T\varphi(x_k) + b] - 1 + e_k = 0, \quad k = 1, 2, ..., N$

- Dual feasibility condition:

  $\alpha_k \geq 0, \quad k = 1, 2, ..., N$

- Complementary slackness condition:

  $\alpha_k[(y_k[w^T\varphi(x_k) + b] - 1 + e_k) - 1] = 0 \quad k = 1, 2, ..., N$

It can be seen that in the classification problems, the vectorization is in terms of $\Omega = Ky$ (Hadamard's product) and, that in the optimization problems the vectorization is in terms of K.

4 Excercise 4

Consider the following optimization problem:

$$\underset{w,b,e}{min} P(w,e) = \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \quad s.t. \quad e_k = w^t x_k, \quad k = 1,2,...,N \tag{32}$$

4.1 Part a

Calculate the Lagrangian for the problem (3).

$$L(w,e;\alpha) = \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k (e_k - w^t x_k) \tag{33}$$

4.2 Part b

Show tha the KKT matrix system (the dual problem) is given by the eigenvalue problem:

$$\begin{bmatrix} x_1^T x_1 & ... & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & ... & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}$$

with $\lambda = 1/\gamma$

$$\frac{\partial L}{\partial w} = w - \sum_{K=1}^{N} \alpha_k x_k \quad \rightarrow \quad \frac{\partial L}{\partial w} = 0 \quad \rightarrow \quad 0 = w - \sum_{K=1}^{N} \alpha_k x_k$$

$$\boxed{w = \sum_{K=1}^{N} \alpha_k x_k} \tag{34}$$

$$\frac{\partial L}{\partial e_k} = \gamma e_k - \alpha_k \quad \rightarrow \quad \frac{\partial L}{\partial e_k} = 0 \quad \rightarrow \quad 0 = \gamma e_k - \alpha_k$$

$$\boxed{e_k = \frac{\alpha_k}{\gamma}} \tag{35}$$

$$\frac{\partial L}{\partial \alpha_k} = e_k - w^T x_k \quad \rightarrow \quad \frac{\partial L}{\partial \alpha_k} = 0 \quad \rightarrow \quad 0 = e_k - w^T x_k$$

$$\boxed{e_k = w^T x_k} \tag{36}$$

Replacing 35 in 36, and 34 in 36 we have:

$$\frac{\alpha_k}{\gamma} = w^T x_k \quad \rightarrow \quad \frac{\alpha_k}{\gamma} = \left(\sum_{K=1}^{N} \alpha_k x_k \right)^T x_k \quad \rightarrow \quad \frac{\alpha_k}{\gamma} = \sum_{K=1}^{N} \alpha_k^T x_k^T x_k$$

$$\sum_{K=1}^{N} x_k^T x_k \alpha_k^T = \frac{1}{\gamma} \alpha_k \tag{37}$$

since $\alpha_k \in \mathbb{R} \rightarrow \alpha_k^T = \alpha_k$, and with $\lambda = 1/\gamma$, we have that 37 can be expressed in the form:

$$\begin{bmatrix} x_1^T x_1 & ... & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & ... & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix}$$

4.3 Part c

Finally, explain (with math) how the optimization problem (3) is related to the principal component analysis (PCA) method.

To show the link between Least Squares Support Vector Machines (LS-SVM) and PCA, we first consider we have the data $\{X_k\}_{k=1}^N$ as the input space, a data with zero mean as points in space, for which one tries to find the projected variables $w^T x$ that captures the most amount of variation (maximum variance). Therefore the constrained optimization problem for this, is the following:

$$\max_w Var(w^T x) = Cov(w^T x, w^T x) \tag{38}$$

where:
w^T = vector of weights (parameters)
x = data with zero mean
we state 38 because $Var(x) = Cov(x, x)$, and with $\mu_x = 0$. With this we can state that:

$$cov(w^T x, w^T x) \simeq \frac{1}{N}\sum_{k=1}^{N}(w^T x_k)^2 \simeq \frac{1}{N}\sum_{k=1}^{N} w^T w x_k^T x_k \tag{39}$$

with 39 we can reformulate 38 like the following:

$$\max_w Var(w^T x) = w^T \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T w \tag{40}$$

We define as a restriction $w^T w = 1$ in order to express that the variance of the data is constant. With that we have a reformulation of the constrained optimization problem:

$$L(w; \lambda) = \frac{1}{2} w^T \sum_{k=1}^{N} \frac{x_k x_k^T}{N} w - \lambda(w^T w - 1) \tag{41}$$

we the proceed to find $\frac{\partial L}{\partial w} = 0$, $\frac{\partial L}{\partial \lambda} = 0$, by using $C = \frac{1}{N}\sum_{k=1}^{N} x^T x$ with which we have:

$$Cw = \lambda w \tag{42}$$

where:
C = a symmetric and positive semidefinite matrix
w = eigenvector that corresponds to the largest eigenvalue
λ = a scalar value

Therefore, we can express the original data in terms of vectors instead its original dimensions, because if C is a diagonalizeable nxn matrix, then it has n eigenvectors, and we know eigenvectors are ortogonal to each other. So we can express the original data in terms of vectors instead in terms of its original dimensions and those will be ortogonal to each other, and also, every eigenvector will capture the max variance.

with 42 and the following formulation will help us to stablish a direct link between (LS-SVM) and PCA.

$$\max_w \sum_{k=1}^{N} [0 - z]^2 \tag{43}$$

where:
$z = w^T x$: The projected variable, x, to a target space z.
0: Is considered a single target value.

Since with PCA the main interest is to find the direction for which the variance is maximal, and, with the formulation of LS/SVM, this leads to the following primal optimization problem:

$$\underset{w,e}{max} L(w,e) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w \quad s.t. \quad e_k = w^T x_k - 0, \quad k = 1, 2, ..., N \tag{44}$$

We can observe in 45 that, since in 42 we defined a value 0 as target, the error variables in this formulation, LS-SVM for PCA, will be the difference between the projected data points, $\{X_k\}_{k=1}^{N}$ to the target space z, and the value of 0. These error variables, E_k, are maximized for the given N data points. And this is done while keeping the norm of w small with the regularization term γ, which is a positive real constant.

Now, we have the next lagrangian formulation:

$$L(w, e; \alpha_k) = \frac{1}{2} \gamma \sum_{k=1}^{N} e_k^2 - \frac{1}{2} w^T w - \sum_{k=1}^{N} \alpha_k \left(e_k - w^T x_k \right) \tag{45}$$

The first order conditions are:

$$\frac{\partial L}{\partial w} = w - \sum_{K=1}^{N} \alpha_k x_k \quad \rightarrow \quad \frac{\partial L}{\partial w} = 0 \quad \rightarrow \quad 0 = w - \sum_{K=1}^{N} \alpha_k x_k$$

$$\boxed{w = \sum_{K=1}^{N} \alpha_k x_k} \tag{46}$$

$$\frac{\partial L}{\partial e_k} = \gamma e_k - \alpha_k \quad \rightarrow \quad \frac{\partial L}{\partial e_k} = 0 \quad \rightarrow \quad 0 = \gamma e_k - \alpha_k$$

$$\boxed{e_k = \frac{\alpha_k}{\gamma} \quad k = 1, 2, ..., N} \tag{47}$$

$$\frac{\partial L}{\partial \alpha} = e_k - w^T x_k \quad \rightarrow \quad \frac{\partial L}{\partial \alpha} = 0 \quad \rightarrow \quad 0 = e_k - w^T x_k$$

$$\boxed{e_k = w^T x_k \quad k = 1, 2, ..., N} \tag{48}$$

If we replace 52 in 48 and 47 also in 48, and if we consider that $\lambda = \frac{1}{\gamma}$, we have:

$$\begin{bmatrix} x_1^T x_1 & ... & x_1^T x_N \\ \vdots & & \vdots \\ x_N^T x_1 & ... & x_N^T x_N \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} = \lambda \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{bmatrix} \quad \rightarrow \quad Cw = \lambda w$$

where:
$C = x_k^T x_l, \quad x, l = 1, 2, ..., N$
$w = \alpha, \quad \alpha = 1, 2, ..., N$

Since we know that $k(x_k, x_l) = x_k^T x_l$ is a linear kernel, we can observe that C is the $Gram$ matrix of such kernel. Also, α is a vector of the dual variables and also an eigenvector of the problem, with λ as its corresponding value. So in order to obtain the maximal variance we must select the eigenvector corresponding to the match of the largest eigenvalue. Therefore, the score value, that is, the projected variable to a target space will be:

$$z(x) = w^T x = \sum_{l=1}^{N} \alpha_l x_l^T x_l \tag{49}$$

5 Excercise 5

Consider the following optimization problem:

$$\min_{w,b,\xi} \quad P(w,\xi) = \frac{1}{2}w^T w + c\sum_{k=1}^{N}\xi_k \quad s.t. \quad y_k\left[w^T\phi(x_k)+b\right] \geq 1-\xi_k, \quad \xi_k \geq 0, \quad k=1,2,...,N. \tag{50}$$

where $y_k \in \{-1,1\}$ is the response (target) variable, $\xi_k \in \mathbb{R}^n$ are slack variables, and $c>0$. Where $\{x_k,y_k\}_{k=1}^N$ represents a training set with input data $x_k \in \mathbb{R}^n$, the output data given by $y_k \in \mathbb{R}$, $e_k \in \mathbb{R}^n$ are slack variables, and the feature maps have the form $\phi(.) : \mathbb{R}^n \to \mathbb{R}^m$. Then, the model's parameters are $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$. Finally, $\gamma > 0$.

5.1 Part a

The lagrangian for the problem 50 is the following:

$$L(w,\xi;\alpha) = \frac{1}{2}w^T w + c\sum_{k=1}^{N}\xi_k - \sum_{k=l}^{N}\alpha_k(y_k[w^T\phi(x_k)+b]-1+\xi_k) - \sum_{k=1}^{N}V_k\xi_k \tag{51}$$

The above equationm is similar to lagrangian from where $\lambda_k = V_k$

5.2 Part b

Calculate the dual cost function (wolfe lagrangian):

$$D(\alpha) = \min_{w,b} = L(w,b,\alpha)$$

$$\frac{\partial L}{\partial w} = w - \sum_{K=1}^{N}\alpha_k y_k \varphi(x_k) \quad \to \quad \frac{\partial L}{\partial w} = 0$$

$$\boxed{w = \sum_{K=1}^{N}\alpha_k y_k \varphi(x_k)} \tag{52}$$

$$\frac{\partial L}{\partial b} = -\sum_{k=1}^{N}\alpha_k y_k \quad \to \quad \frac{\partial L}{\partial b} = 0$$

$$\boxed{\sum_{k=1}^{N}\alpha_k y_k = 0} \tag{53}$$

$$\frac{\partial L}{\partial \xi} = c - \alpha_k - \lambda_k \quad \to \quad c = \alpha_k - \lambda_k, \quad k=1,2,...,N$$

$$\boxed{0 \leq \alpha_k \leq c} \tag{54}$$

So the final formulation will be:

$$L(w,b,\xi,\alpha,\lambda) = \frac{1}{2}w^T w - \sum_{k=1}^{N}\alpha_k[y_k(w^T\varphi(x_k)+b)-1] - \sum_{k=1}^{N}\alpha_k\xi_k + c\sum_{k=1}^{N}\xi_k - \sum_{k=1}^{N}\lambda_k\xi_k \tag{55}$$

5.3 Part c

Show that the dual problem in lagrange multiplication α_k (The Wolfe dual problem) as a quadratic programming formulation.

by using $(x_k, x_l) = \varphi(x_N)^T \varphi(x_l)$ we can formulate the problem like the following:

$$maxD(\alpha) = -\frac{1}{2}\sum_{k=1}^{N} \alpha_k \alpha_l y_k y_l K(x_k, x_l) + \sum_{k=1}^{N} \alpha_k \tag{56}$$

with $\sum_{k=1}^{N} \alpha_k y_k = 0, \quad k = 1, 2, ..., N, \quad 0 \leq \alpha_k \leq c$

5.4 Part d

For the problem (4), derive the KKT conditions:

- Stationary condition

  $w_l(w, b, \alpha) = 0 \quad \rightarrow \quad w = \sum_{k=1}^{N} \alpha_k y_k x_k$

  $\frac{\partial L}{\partial b}(w, b, \alpha) = 0 \quad \rightarrow \quad \sum_{k=1}^{N} \alpha_k y_k = 0$

- Primal feasibility condition

  $y_k[w^T x_k + b] \geq 1 - \xi_k \quad \rightarrow \quad y_k[w^T x_k + b] + \xi_k \geq 1$

- Dual feasibility condition

  $y_k > 0, \alpha_k > 0$

- Complementary slackness condition

  $\alpha_k[y_k(w^T x_k + b) - 1 + \xi_k]$

5.5 Part e

Use the complementary slackness condition to prove the existence of some $\alpha_k = 0$. Then, confirm or refute the following conditions related to α_k:

$$\begin{aligned} \alpha_k = 0 &\implies y_k[w^T \varphi(x_k) + b] \geq 1 \\ \alpha_k = c &\implies y_k[w^T \varphi(x_k) + b] \leq 1 \\ 0 < \alpha_k < c &\implies y_k[w^T \varphi(x_k) + b] = 1 \end{aligned}$$

from slackness condition, we have that: $\alpha_k[y_k(w^T x_k + b) - 1 + \xi_x] = 0$

$$\begin{aligned} \alpha_k &= 0 \\ (w^T \varphi(x_k) + b) + \xi_k &\geq 1 - 1, y = 1 \\ y_k[w^T \varphi(x_k) + b] &\geq 1 \end{aligned}$$

$$\begin{aligned} \alpha_k &= c \\ y_k[w^T \varphi(x_k) + b] - 1 + \xi_k &= 0 \\ y_k[w^T \varphi(x_k) + b] &= 1 \end{aligned} \tag{57}$$

$$\begin{aligned} 0 \leq \alpha_k &\leq c \\ y_k[w^T \varphi(x_k) + b] - 1 + \xi_k &= 0 \\ y_k[w^T \varphi(x_k) + b] &= 1 \end{aligned}$$

5.6 Part f

Present a rigorous method to find the value of b.

$$\begin{gathered}\alpha_k[y_k(w^T x_k + b) - (1 + \xi_k)] = 0 \\ \alpha_k y_k(w^T x_k) + \alpha_k y_k b - \alpha_k(1 + \xi_k) = 0 \\ \alpha_k y_k b = \alpha_k(1 - \xi_k) - \alpha_k y_k(w^T x_k) \\ b = \frac{\alpha_k(1-\xi_k) - \alpha_k y_k(w^T x_k)}{\alpha_k y_k}\end{gathered} \tag{58}$$

$$\boxed{b = \frac{1 - \xi_k}{y_k} - w^T x_k}$$

5.7 Part g

Finally provide the following refinements and explanations.

- Define $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$ for $k, l = 1, 2, ..., N$ and write the dual problem in terms of $K(x_k, x_l)$.

$$maxD(\alpha) = \frac{1}{2}\sum_{k=1}^{N} \alpha_k \alpha_l y_k y_l K(x_k, x_l) + \sum_{k=1}^{N} \alpha_k y_k = 0, \quad k = 1, 2, ..., N \quad 0 \leq \alpha_k \leq c \tag{59}$$

- Show that the solution vector w has an explansion in terms of the training vectors x_k with $k = 1, 2, ..., N$. Then, explain the effect of the sparsness on w.

$$w = \sum_{k=1}^{N} \alpha_k y_k x_k \tag{60}$$

- Those Lagrange multipliers such that $\alpha_k > 0$ are called *support values*, and those vectors x_k corresponding to the *support values* in the expansaion for the solution vector w are called *support vectors*. Firstly, notice that, by definition, only the *support values* and *support vectors* are relevant to provide a solution for the optimization problem.

$$\alpha_k[y_k(w^T x_k + b) - 1 + \xi_k] \quad \rightarrow \quad \alpha > 0, \quad y_k(w^T x_k + b) + \xi = 1 \tag{61}$$

- Explain (with math) how the optimization problem (4) is related to the binary classification problem.

$$\begin{gathered}\alpha_k[y_k(w^T x_k + b) - 1 + \xi_k] = 0, \quad w = \sum_{k=1}^{N} \alpha_k y_k x_k, \quad k(x_k, x_l) = x_k^T x_l \\ y_l \left[\left(\sum_{k-1}^{N} \alpha_k y_k k(x_k, x_l) + b \right) - 1 + \xi_l \right] = 1, \quad y \in \{-1, 1\} \\ sign \left[\left(\sum_{k=1}^{N} \alpha_k y_k k(x_k x_l) + b \right) + \xi_k \right]\end{gathered}$$

Apendix A - Extra features of the SVM Approach

We present the following content as interesting thoughts and questions about some fundamental assumptions, characteristics and features of SVM general formulations.

Recall that the fundamental idea in SVM is that the method maps the input vectors x into a high dimensional feature space Z, by some nonelinear mapping that is choosen *a priori*, and in this space Z an optimal separating hyperplane is constructed. With that fundamental idea, one can consequently think that the dimensionality of the feature space will be huge, and, a hyperplane that separates the training data will not necessarily generalize well. So two different and equally interesting questions can be formulated:

1. Why a high dimensional hyperplane will generalize sufficiently well?
   See *Generalization of the hyperplane*
2. How such a high dimensional hyperplane could be found ?
   See *The role of the convolution of the inner product*

5.8 Generalization of the hyperplane

We will start by proposing a conceptual definition of generalization: *Generalization is the act whereby a learned response is made to a stimulus similar to but not identical wit the conditioned (learned) stimulus.* In other words, for any similar input data a similar output value will be produced, similar to the learned response (since is supervised learning).

For the first question: *Why a high dimensional hyperplane will generalize sufficiently well?*, some useful concepts for this are:

- **Δ-margin separating hyperplane**

  We call hyperplane the following:

  $$(w^* \cdot x) - b = 0, \quad |w^*| = 1 \tag{62}$$

  The canonical form would is:

  $$y_i[(w \cdot x_i) - b] \geq 1, \quad i = 1, 2, ..., l \tag{63}$$

  And, a Δ-margin separating hyperplane if it classifies vectors x as follows:

  $$y = \begin{cases} +1 & \text{if} \quad (w^* \cdot x) - b \geq +\Delta \\ -1 & \text{if} \quad (w^* \cdot x) - b \leq -\Delta \end{cases} \tag{64}$$

  Its easy to see check that the optimal hyperplane as defined above is the Δ-margin separating hyperplane with $\Delta = {}^1/_{|w^*|}$

- **Soft margin separating hyperplane**

  Also called the generalized optimal hyperplane, this is determined by the vector of w that minimizes the functional:

  $$\Phi(w, \xi) = \frac{1}{2}(w \cdot w) + c\left(\sum_{i=1}^{l} \xi_i\right) \tag{65}$$

  here c is a fiven value that is subject to the constraint $y_i((w \cdot x_i) - b) \geq 1 - \xi_i \quad i = 1, 2, ..., l$. As part of a non separable case, the solution of this quadratic optimization problem is almost equivalent to the technique used in the separable case.

- **VC dimension**

  We can define it by: The maximum number of vectors that can be *shattered* by the set of functions.

  Therefore, as stated with help of the above concepts and definitions, the VC dimension of the set of Δ-margin separating hyperplanes with large Δ is small.

So, the generalization ability of a constructed hyperplane is high, if, the VC Dimension of the set of Delta-margin separating hyperplanes with large Delta is small. With the condition that the training set contain l examples that are separated by the maximal margin hyperplane, we have that the expectation of the probability of test error (in training set) is bounded by the expectation of the minimum of the following three values:

$$EP_{error} \leq Emin\left(\frac{m}{l}, \frac{R^2|w|^2}{l}, \frac{n}{l}\right) \tag{66}$$

where:
m = The number of support vectors.
R = The radius of the sphere containing the data.
$|w|^2$ = The value of the margin.
n = The dimensionality of the input space.

Above interpretations are the following: $^m/_l$ is the proportion of the support vectors and the observations that are separated by the maximal margin hyperplane. For $^{R^2|w|^2}/_l$ we can infer that is the proportion of the area that contains the data within the margin and the separated observations, and finally, we see that the factor $^n/_l$ expresses the proportion between the dimensionality of the input space and the separated observations.

With the result stated with (1), the answer to the generalization question, We rely in that the expectation of the of the data compression is large and that the expectation of the margin is large. With this we have indeed a high generalization capacity of the SVM method.

5.9 The role of the convolution of the inner product

Now we have addressed the previous question of the ability of generalization of the SVM method, even if the optimal hyperplane generalizes well and can theoretically be found, there remains another situation, a technical one, and that is how to treat the high-dimensional feature space in a practical-computational sense.

In order to construct the optimal separating hyperplane in the feature space Z, one has only to be able to calculate the inner products between support vectors and the vectors of the feature space, so the following particular calculations are important:

We consider the following functional, obtaining it by putting the expression $\alpha_i^0[(x_i \cdot w_0) - b_0]y_i - 1$ for w_0 into Lagrangian form and taking into account the Kuhn-Tucker conditions:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{67}$$

and a particularlly important constraint:

$$y_i((w \cdot x_i) - b) \geq 1 - \xi_i, \quad i = 1, 2, ..., l, \tag{68}$$

Therefore, one does not need to consider the feature space in its *explicit form.* And more over, if we consider a general expression for the inner product in *Hilbert space*:

$$(z_i \cdot z) = K(x, x_i) \tag{69}$$

where:
z: Is the image in feature space of the vector x in input space. x: Vector in the input space.
According to Hilbert-Schmidt theory, $k(x, x_i)$ can be any symmetric function satistying the follow general conditions:

$$K(u, v) = \sum_{k=1}^{\infty} \alpha_k \Psi_k(u) \Psi_k(u) \tag{70}$$

With positive coefficient $\alpha_k > 0$, and with $K(u, v)$ as an inner product in some feature space. It is necessary and sufficient that:

$$\int \int K(u, v)g(u)g(v)dudv > 0 \tag{71}$$

be valid for all $g \neq 0$ for which:

$$\int g^2(u)du < \infty \tag{72}$$

Another interesting feature is that the convolution of inner products allows the construction of decision functions that are nonlinear in the input space, for example:

$$f(x) = sign\left[\sum_{supportvectors} y_i\alpha_i k(x_i, x) - b\right] \tag{73}$$

And those decision functions are in fact equivalent to linear decision functions in the high-dimensional feature space $\Psi_1(x), \Psi_2(x), ..., \Psi_N(x)$. So, $k(x_i, x)$ is a convolution of the inner product for this feature space.