Machine Learning

Answer Sheet for Homework 5

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

December 9, 2015

Problem 1

The hard-margin support vector machine is with $d+1$ variables. For soft-margin support vector machine, there are N more variables ξ_n, $1 \leq n \leq N$.

So soft-margin support vector machine is a quadratic programming problem with $N+d+1$ variables.

$\square$

Problem 2

I wrote a `Q02.py` to help me get the answer. By using Python package `cvxopt`[2], with

$$\mathbf{z} = \begin{bmatrix} 1 & -2 \\ 4 & -5 \\ 4 & -1 \\ 5 & -2 \\ 7 & -7 \\ 7 & 1 \\ 7 & 1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ +1 \\ +1 \\ +1 \\ +1 \end{bmatrix} \tag{1}$$

and

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{p} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2}$$

$$\mathbf{A}^T = \begin{bmatrix} -1 & -1 & 2 \\ -1 & -4 & 5 \\ -1 & -4 & 1 \\ 1 & 5 & -2 \\ 1 & 7 & -7 \\ 1 & 7 & 1 \\ 1 & 7 & 1 \end{bmatrix}, \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{3}$$

To use this package, I gave `solvers.qp(`$\mathbf{Q}, \mathbf{p}, -\mathbf{A}^T, -\mathbf{c}$`)` and got

$$b = -9, \mathbf{w} = [2, 0] \tag{4}$$

So the hyperplane is

$$2z_1 - 9 = 0 \Rightarrow z_1 = 4.5 \tag{5}$$

$\square$

Problem 3

I wrote a `Q03.py` to help me get the answer. By using Python package `cvxopt`, with

$$\mathbf{Q} = \begin{bmatrix} 4 & 1 & 1 & 0 & -1 & -1 & -1 \\ 1 & 4 & 0 & -1 & -9 & -1 & -1 \\ 1 & 0 & 4 & -1 & -1 & -9 & -1 \\ 0 & -1 & -1 & 4 & 1 & 1 & 9 \\ -1 & -9 & -1 & 1 & 25 & 9 & 1 \\ -1 & -1 & -9 & 1 & 9 & 25 & 1 \\ -1 & -1 & -1 & 9 & 1 & 1 & 25 \end{bmatrix}, \mathbf{p} = \begin{bmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \tag{6}$$

$$-\mathbf{A}^T = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

with

$$\mathbf{G} = \mathbf{y}^T = \begin{bmatrix} -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix} \text{ and } h = 0 \tag{8}$$

and To use this package, I gave `solvers.qp(`$\mathbf{Q}, \mathbf{p}, -\mathbf{A}^T, \mathbf{c}, \mathbf{G}, h$`)` and got

$$\alpha = \left[4.32 \times 10^{-9} \approx 0, 0.704, 0.704, 0.889, 0.259, 0.259, 5.27 \times 10^{-10} \approx 0\right] \tag{9}$$

where `cvxopt` needs conditions

$$-\mathbf{A}^T \boldsymbol{\alpha} \preceq \mathbf{c} \text{ and } \mathbf{G}\boldsymbol{\alpha} = h \tag{10}$$

Support vectors are the corresponding $\alpha_i \neq 0$, so $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$ and $\mathbf{x}_6$ are support vectors.

□

Problem 4

I wrote a `Q04.py` to help me get the answer. By using python package `sympy` and

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n K\left(\mathbf{x}_n, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) + b \tag{11}$$

$$b = y_s - \sum_{n=1}^{N} \alpha_n y_n K\left(\mathbf{x}_n, \mathbf{x}_s\right) \tag{12}$$

we have

$$\mathbf{w} = \frac{1}{9}\left(8x_1^2 - 16x_1 + 6x_2^2 - 15\right) \tag{13}$$

□

Problem 5

Since kernel function $K\left(\mathbf{x}, \mathbf{x}'\right) = \left(1 + \mathbf{x}^T\mathbf{x}'\right)^2$ is different from $\mathbf{z} = \left(\phi\left(\mathbf{x}\right), \phi\left(\mathbf{x}\right)\right)$, the curves should be different in the $\mathcal{X}$ space.

□

Problem 6

Since $\left\|\mathbf{x}_n - \mathbf{c}\right\|^2 \leq R^2$, $\forall n$, the constraint to maximize is

$$\left\|\mathbf{x}_n - \mathbf{c}\right\|^2 - R^2 \leq 0 \tag{14}$$

so $L(R, \mathbf{c}, \boldsymbol{\lambda})$ is

$$L(R, \mathbf{c}, \boldsymbol{\lambda}) = R^2 + \sum_{n=1}^{N} \lambda_n \left(\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 \right) \tag{15}$$

□

Problem 7

At the optimal $(R, \mathbf{c}, \boldsymbol{\lambda})$,

$$\frac{\partial L}{\partial R} = 2R - 2R \sum_{n=1}^{N} \lambda_n = 0 \Rightarrow \sum_{n=1}^{N} \lambda_n = 1 \text{ or } R = 0 \tag{16}$$

$$\frac{\partial L}{\partial \mathbf{c}} = 2 \sum_{n=1}^{N} \lambda_n (\mathbf{c} - \mathbf{x}_n) = \mathbf{0} \Rightarrow \mathbf{c} = \left(\sum_{n=1}^{N} \lambda_n \mathbf{x}_n \right) \Big/ \left(\sum_{n=1}^{N} \lambda_n \right) \text{ if } \sum_{n=1}^{N} \lambda_n \neq 0 \tag{17}$$

So the KKT conditions are

1. primal feasible: $\|\mathbf{x}_n - \mathbf{c}\|^2 \leq R^2$.
2. dual feasible: $\lambda_n \geq 0$.
3. dual-inner optimal: if $R \neq 0$, $\sum_{n=1}^{N} \lambda_n = 1$ and $\mathbf{c} = \sum_{n=1}^{N} \lambda_n \mathbf{x}_n$.
4. primal-inner optimal: $\lambda_n \left(\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 \right) = 0$.

□

Problem 8

From Problem 6, we have

$$L(R, \mathbf{c}, \boldsymbol{\lambda}) = R^2 + \sum_{n=1}^{N} \lambda_n \left(\|\mathbf{x}_n - \mathbf{c}\|^2 - R^2 \right) = R^2 + \sum_{n=1}^{N} \lambda_n \|\mathbf{x}_n - \mathbf{c}\|^2 - \sum_{n=1}^{N} \lambda_n R^2 \tag{18}$$

$$= R^2 - R^2 + \sum_{n=1}^{N} \lambda_n \|\mathbf{x}_n - \mathbf{c}\|^2 = \sum_{n=1}^{N} \lambda_n \|\mathbf{x}_n - \mathbf{c}\|^2 \tag{19}$$

where $\sum_{n=1}^{N} \lambda_n = 1$ since $R \neq 0$.
Also, from (17), we have $\mathbf{c} = \sum_{n=1}^{N} \lambda_n \mathbf{x}_n$. Hence

$$\text{Objective}(\lambda) = \sum_{n=1}^{N} \lambda_n \left\| \mathbf{x}_n - \sum_{m=1}^{N} \lambda_m \mathbf{x}_m \right\|^2 \tag{20}$$

□

Problem 9

We have

$$\sum_{n=1}^{N} \lambda_n \left\|\mathbf{x}_n - \mathbf{c}\right\|^2 = \sum_{n=1}^{N} \lambda_n \left(\mathbf{x}_n^T \mathbf{x}_n - \mathbf{x}_n^T \mathbf{c} - \mathbf{c}^T \mathbf{x}_n + \mathbf{c}^T \mathbf{c}\right) \tag{21}$$

$$= \sum_{n=1}^{N} \lambda_n \left(\mathbf{x}_n^T \mathbf{x}_n - \mathbf{x}_n^T \sum_{m=1}^{N} \lambda_m \mathbf{x}_m - \left(\sum_{m=1}^{N} \lambda_m \mathbf{x}_m\right)^T \mathbf{x}_n + \left\|\sum_{m=1}^{N} \lambda_m \mathbf{x}_m\right\|^2\right) \tag{22}$$

So

$$\sum_{n=1}^{N} \lambda_n \left\|\phi\left(\mathbf{x}_n\right) - \phi\left(\mathbf{c}\right)\right\|^2 \tag{23}$$

$$= \sum_{n=1}^{N} \lambda_n K\left(\mathbf{x}_n, \mathbf{x}_n\right) - 2 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K\left(\mathbf{x}_n, \mathbf{x}_m\right) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K\left(\mathbf{x}_n, \mathbf{x}_m\right) \tag{24}$$

$$= \sum_{n=1}^{N} \lambda_n K\left(\mathbf{x}_n, \mathbf{x}_n\right) - \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K\left(\mathbf{x}_n, \mathbf{x}_m\right) \tag{25}$$

□

Problem 10

From primal-inner optimal condition, pick some $\lambda_i > 0$, we have

$$\left\|\mathbf{x}_i - \mathbf{c}\right\|^2 = R^2 \tag{26}$$

so

$$R^2 = \mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \sum_{m=1}^{N} \lambda_m \mathbf{x}_m - \left(\sum_{m=1}^{N} \lambda_m \mathbf{x}_m\right)^T \mathbf{x}_i + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m \mathbf{x}_n^T \mathbf{x}_m \tag{27}$$

$$= K\left(\mathbf{x}_i, \mathbf{x}_i\right) - 2 \sum_{m=1}^{N} \lambda_m K\left(\mathbf{x}_i, \mathbf{x}_m\right) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K\left(\mathbf{x}_n, \mathbf{x}_m\right) \tag{28}$$

$$\Rightarrow R = \sqrt{K\left(\mathbf{x}_i, \mathbf{x}_i\right) - 2 \sum_{m=1}^{N} \lambda_m K\left(\mathbf{x}_i, \mathbf{x}_m\right) + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K\left(\mathbf{x}_n, \mathbf{x}_m\right)} \tag{29}$$

where $R > 0$.

□

Problem 11

Claim: Let $\tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ \sqrt{2C} \cdot y_1\xi_1 \\ \sqrt{2C} \cdot y_2\xi_2 \\ \vdots \\ \sqrt{2C} \cdot y_N\xi_N \end{bmatrix}$ and $\tilde{\mathbf{x}}_n = \begin{bmatrix} \mathbf{x}_n \\ v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$, where $v_i = \frac{1}{\sqrt{2C}} [\![i = n]\!]$.

Proof of Claim:

First, we have

$$\frac{1}{2}\tilde{\mathbf{w}}^T\tilde{\mathbf{w}} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} y_n^2\xi_n^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} \xi_n^2 \tag{30}$$

where $y_n^2 = 1$ due to $y_n \in \{+1, -1\}$.

And (P_2) can be rewritten as

$$\min_{\tilde{\mathbf{w}}, b, \boldsymbol{\xi}} \left(\frac{1}{2}\tilde{\mathbf{w}}^T\tilde{\mathbf{w}} \right) \tag{31}$$

Then we have

$$\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_n = \tilde{\mathbf{w}}^T \begin{bmatrix} \mathbf{x}_n \\ v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = \tilde{\mathbf{w}}^T \begin{bmatrix} \mathbf{x}_n \\ 0 \\ \vdots \\ \underbrace{\frac{1}{\sqrt{2C}}}_{i=n} \\ \vdots \\ 0 \end{bmatrix} \tag{32}$$

$$= \mathbf{w}^T\mathbf{x}_n + 0 + \cdots + y_n\xi_n + \cdots + 0 = \mathbf{w}^T\mathbf{x}_n + y_n\xi_n \tag{33}$$

So

$$y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) \geq 1 - \xi_n = 1 - y_n^2\xi_n \tag{34}$$

$$\Rightarrow y_n\left(\mathbf{w}^T\mathbf{x}_n + b\right) + y_n^2\xi_n = y_n\left(\mathbf{w}^T\mathbf{x}_n + y_n\xi_n + b\right) = y_n\left(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}_n + b\right) \geq 1 \tag{35}$$

Hence, (P_2) is equivalent to a linear hard-margin support vector machine primal problem.

$\square$

Problem 12

Claim: $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$ and $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$ are always valid kernels.

Proof of Calim:

1. $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') + K_2(\mathbf{x}, \mathbf{x}')$

   Consider Mercer's conditions,

   (a) Symmetric

   Since K_1 and K_2 are valid kernel, both of them are symmetric. So $K_1 + K_2$ must be symmetry.

   (b) K is positive semi-definite

   Consider any vector $\mathbf{v}$, we have

$$\mathbf{v}^T K_1 \mathbf{v} = \begin{bmatrix} \sum_{i=1}^N v_i \phi_1(\mathbf{x}_i) \phi_1(\mathbf{x}'_1) & \cdots & \sum_{i=1}^N v_i \phi_1(\mathbf{x}_i) \phi_1(\mathbf{x}'_N) \end{bmatrix} \mathbf{v} \tag{36}$$

$$= \sum_{j=1}^N \sum_{i=1}^N v_i \phi_1(\mathbf{x}_i) \phi_1\left(\mathbf{x}'_j\right) v_j \geq 0 \tag{37}$$

$$\mathbf{v}^T K_2 \mathbf{v} = \begin{bmatrix} \sum_{i=1}^N v_i \phi_2(\mathbf{x}_i) \phi_2(\mathbf{x}'_1) & \cdots & \sum_{i=1}^N v_i \phi_2(\mathbf{x}_i) \phi_2(\mathbf{x}'_N) \end{bmatrix} \mathbf{v} \tag{38}$$

$$= \sum_{j=1}^N \sum_{i=1}^N v_i \phi_2(\mathbf{x}_i) \phi_2\left(\mathbf{x}'_j\right) v_j \geq 0 \tag{39}$$

$$\mathbf{v}^T K \mathbf{v} = \mathbf{v}^T (K_1 + K_2) \mathbf{v} = \sum_{j=1}^N \sum_{i=1}^N v_i \left(\phi_1(\mathbf{x}_i) \phi_1\left(\mathbf{x}'_j\right) + \phi_2(\mathbf{x}_i) \phi_2\left(\mathbf{x}'_j\right)\right) v_j \tag{40}$$

$$= \mathbf{v}^T K_1 \mathbf{v} + \mathbf{v}^T K_2 \mathbf{v} \geq 0 \tag{41}$$

   Hence, K is positive semi-definite.

   By satisfying the Mercer's conditions, K is a valid kernel.

2. $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$

   Similarly, K is symmetry since $K_m\left(\mathbf{x}_i, \mathbf{x}'_j\right) = K_m(\mathbf{x}_j, \mathbf{x}'_i)$, for $m = 1$ or 2. Then

$$K\left(\mathbf{x}_i, \mathbf{x}'_j\right) = K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right) K_2\left(\mathbf{x}_i, \mathbf{x}'_j\right) = K_1(\mathbf{x}_j, \mathbf{x}'_i) K_2(\mathbf{x}_j, \mathbf{x}'_i) = K(\mathbf{x}_j, \mathbf{x}'_i) \tag{42}$$

   Applying Cholesky decomposition, rewrite

$$K_1 = A^T A = \sum_{k=1}^N a_{ik} a_{jk} \text{ and } K_2 = B^T B = \sum_{k=1}^N b_{ik} b_{jk} \tag{43}$$

then for any $\mathbf{v}$, we have

$$\mathbf{v}^T K \mathbf{v} = \sum_{j=1}^{N} \sum_{i=1}^{N} v_i \left(\sum_{k=1}^{N} a_{ik} a_{jk} \right) \left(\sum_{\ell=1}^{N} b_{i\ell} b_{j\ell} \right) v_j \tag{44}$$

$$= \sum_{k,\ell=1}^{N} \sum_{i,j=1}^{N} v_i v_j a_{ik} a_{jk} b_{i\ell} b_{j\ell} \tag{45}$$

$$= \sum_{k,\ell=1}^{N} \left(\sum_{i=1}^{N} v_i a_{ik} b_{i\ell} \right) \left(\sum_{j=1}^{N} v_j a_{jk} b_{j\ell} \right) \tag{46}$$

$$= \sum_{k,\ell=1}^{N} \left(\sum_{i=1}^{N} v_i a_{ik} b_{i\ell} \right)^2 \geq 0 \tag{47}$$

Hence, K is positive semi-definite.

By satisfying the Mercer's conditions, K is a valid kernel.

□

Problem 13

Claim: $K(\mathbf{x}, \mathbf{x}') = 1126 \cdot K_1(\mathbf{x}, \mathbf{x}')$ and $K(\mathbf{x}, \mathbf{x}') = (1 - K_1(\mathbf{x}, \mathbf{x}'))^{-1}$ are always valid kernels.

Proof of Claim:

1. $K(\mathbf{x}, \mathbf{x}') = 1126 \cdot K_1(\mathbf{x}, \mathbf{x}')$

   Consider Mercer's conditions,

   (a) Symmetric

   Since K_1 is valid kernel, K_1 is symmetric. So K must be symmetric.

   (b) K is positive semi-definite

   Consider any vector $\mathbf{v}$, we have

$$\mathbf{v}^T K \mathbf{v} = \sum_{j=1}^{N} \sum_{i=1}^{N} v_i \left(1126 K_1 \left(\mathbf{x}_i, \mathbf{x}_j' \right) \right) v_j = 1126 \sum_{j=1}^{N} \sum_{i=1}^{N} v_i K_1 \left(\mathbf{x}_i, \mathbf{x}_j' \right) v_j \tag{48}$$

   Since $\sum_{j=1}^{N} \sum_{i=1}^{N} v_i K_1 \left(\mathbf{x}_i, \mathbf{x}_j' \right) v_j \geq 0$, we have

$$1126 \sum_{j=1}^{N} \sum_{i=1}^{N} v_i K_1 \left(\mathbf{x}_i, \mathbf{x}_j' \right) v_j \geq 0 \tag{49}$$

where $\sum_{j=1}^{N}\sum_{i=1}^{N} v_i K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right) v_j \geq 0$ due to K_1 is positive semi-definite. Hence, K is positive semi-definite. By satisfying the Mercer's conditions, K is a valid kernel.

2. $K\left(\mathbf{x}, \mathbf{x}'\right) = \left(1 - K_1\left(\mathbf{x}, \mathbf{x}'\right)\right)^{-1}$

   Consider Mercer's conditions,

   (a) Symmetric

   Since K_1 is valid kernel, K_1 is symmetric. So K must be symmetric.

   (b) K is positive semi-definite

   Consider any vector $\mathbf{v}$, we have

$$\mathbf{v}^T K \mathbf{v} = \sum_{j=1}^{N}\sum_{i=1}^{N} v_i \left(\frac{1}{1 - K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right)}\right) v_j \tag{50}$$

   Since $\sum_{j=1}^{N}\sum_{i=1}^{N} v_i K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right) v_j \geq 0$ and $0 < K_1\left(\mathbf{x}, \mathbf{x}'\right) < 1$, we have

$$\frac{1}{1 - K_1\left(\mathbf{x}, \mathbf{x}'\right)} > K_1\left(\mathbf{x}, \mathbf{x}'\right) \tag{51}$$

   where $\sum_{j=1}^{N}\sum_{i=1}^{N} v_i K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right) v_j \geq 0$ due to K_1 is positive semi-definite. So we have

$$\sum_{j=1}^{N}\sum_{i=1}^{N} v_i \left(\frac{1}{1 - K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right)}\right) v_j \geq \sum_{j=1}^{N}\sum_{i=1}^{N} v_i K_1\left(\mathbf{x}_i, \mathbf{x}'_j\right) v_j \geq 0 \tag{52}$$

   Hence, K is positive semi-definite. By satisfying the Mercer's conditions, K is a valid kernel.

$\square$

Problem 14

Claim: $\tilde{C} = \frac{C}{p}$, $\tilde{\beta}_n = \frac{\beta_n}{p} = \frac{C}{p} - \frac{\alpha_n}{p} = \tilde{C} - \tilde{\alpha}_n$, $\forall n$ for optimal solution.

Proof of Claim:

$$\tilde{g}_{\mathrm{SVM}}(\mathbf{x}) = \operatorname{sign}\left(\sum_{n=1}^{N} \tilde{\alpha}_n y_n \tilde{K}(\mathbf{x}_n, \mathbf{x}) + b\right) \tag{53}$$

$$= \operatorname{sign}\left(\sum_{n=1}^{N} \tilde{\alpha}_n y_n \left(pK(\mathbf{x}_n, \mathbf{x}) + q\right) + b\right) \tag{54}$$

$$= \operatorname{sign}\left(p\sum_{n=1}^{N} \tilde{\alpha}_n y_n K(\mathbf{x}_n, \mathbf{x}) + q\sum_{n=1}^{N} \tilde{\alpha}_n y_n + b\right) \tag{55}$$

$$= \operatorname{sign}\left(p\sum_{n=1}^{N} \left(\tilde{C} - \tilde{\beta}_n\right) y_n K(\mathbf{x}_n, \mathbf{x}) + q \cdot 0 + b\right) \tag{56}$$

$$= \operatorname{sign}\left(\sum_{n=1}^{N} \left(p \cdot \frac{C}{p} - p\tilde{\beta}_n\right) y_n K(\mathbf{x}_n, \mathbf{x}) + b\right) \tag{57}$$

$$= \operatorname{sign}\left(\sum_{n=1}^{N} \left(C - p\tilde{\beta}_n\right) y_n K(\mathbf{x}_n, \mathbf{x}) + b\right) \tag{58}$$

where $\sum_{n=1}^{N} \tilde{\alpha}_n y_n = 0$ due to optimal constraint.
Since $\tilde{\beta}_n = \dfrac{\beta_n}{p}$, then we have

$$\tilde{g}_{\mathrm{SVM}}(\mathbf{x}) = \operatorname{sign}\left(\sum_{n=1}^{N} \left(C - \beta_n\right) y_n K(\mathbf{x}_n, \mathbf{x}) + b\right) \tag{59}$$

$$= \operatorname{sign}\left(\sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b\right) \tag{60}$$

$$= g_{\mathrm{SVM}}(\mathbf{x}) \tag{61}$$

□

Problem 15

By using `scikit-learn` package, we have

$$\|\mathbf{w}\| \text{ list} = \left[6.021 \times 10^{-5}, 6.019 \times 10^{-3}, 5.713 \times 10^{-1}, 1.133 \times 10^{1}, 1.309 \times 10^{1}\right] \tag{62}$$

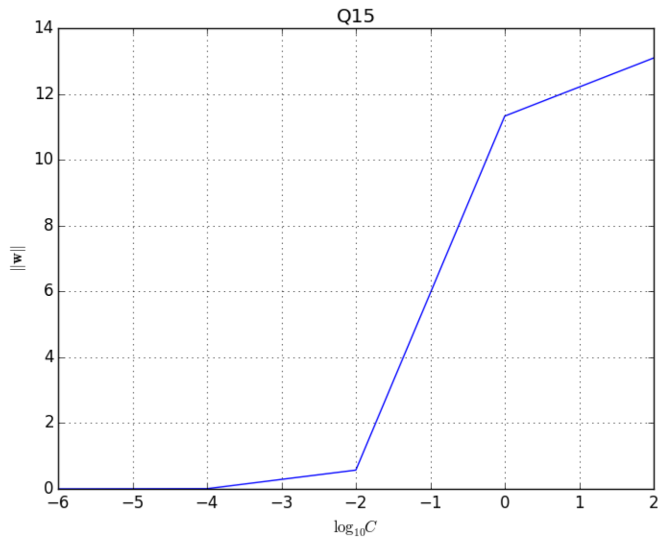

Figure 1: Q15

As C becomes larger, $\|\mathbf{w}\|$ becomes larger, too.

$\square$

Problem 16

All E_{in} versus $\log_{10}(C)$ is 7.434×10^{-2}.

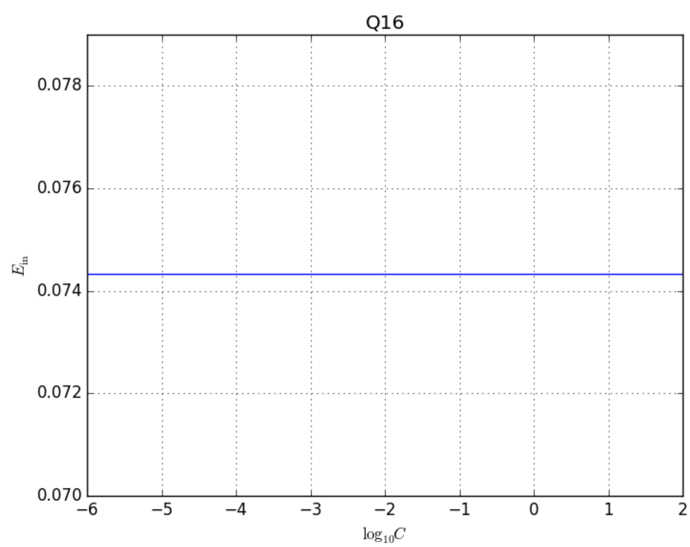

Figure 2: Q16

E_{in} is independent of C.

□

Problem 17

$$\sum_{n=1}^{N} \alpha_n \text{ list} = \left[1.084 \times 10^{-3}, 1.084 \times 10^{-1}, 1.084 \times 10^{1}, 1.084 \times 10^{3}, 1.084 \times 10^{5}\right] \tag{63}$$

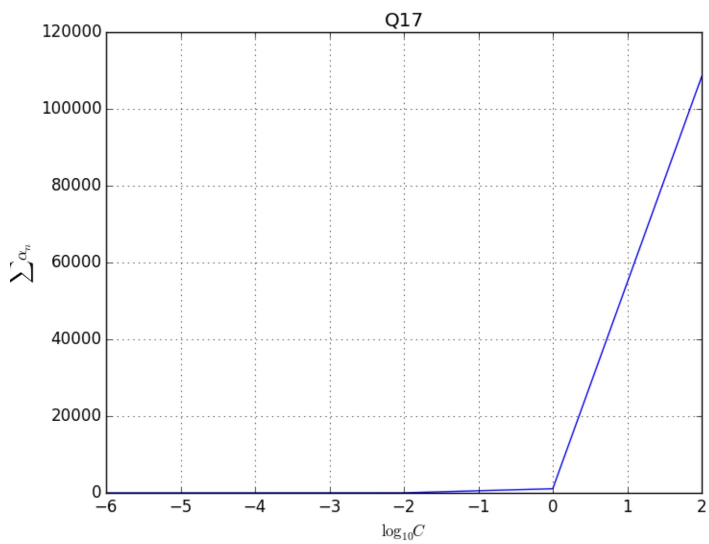

Figure 3: Q17

From this figure, we have

$$\sum_{n=1}^{N} \alpha_n \propto C \tag{64}$$

□

Problem 18

$$\text{distance list} = \left[8.237, 8.248 \times 10^{-1}, 1.288 \times 10^{-1}, 8.416 \times 10^{-2}, 4.109 \times 10^{-2}\right] \tag{65}$$

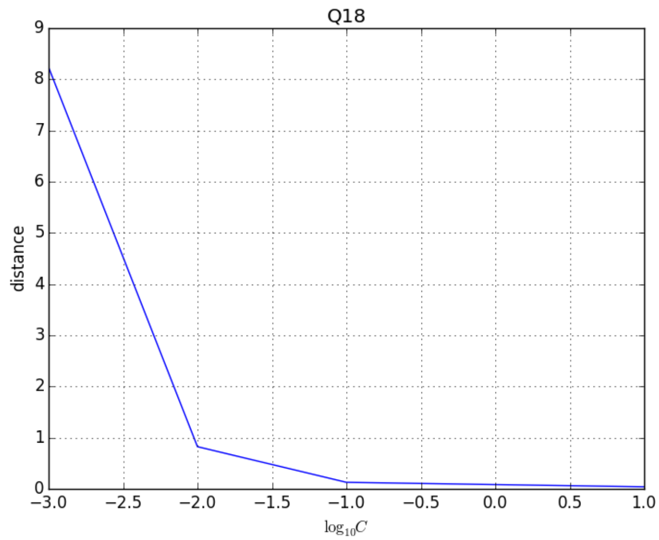

Figure 4: Q18

The distance is strictly decreasing as C becomes larger.

□

Problem 19

$$E_{\text{out}} \text{ list} = \left[1.071 \times 10^{-1}, 9.915 \times 10^{-2}, 1.051 \times 10^{-1}, 1.789 \times 10^{-1}, 1.789 \times 10^{-1}\right] \quad (66)$$

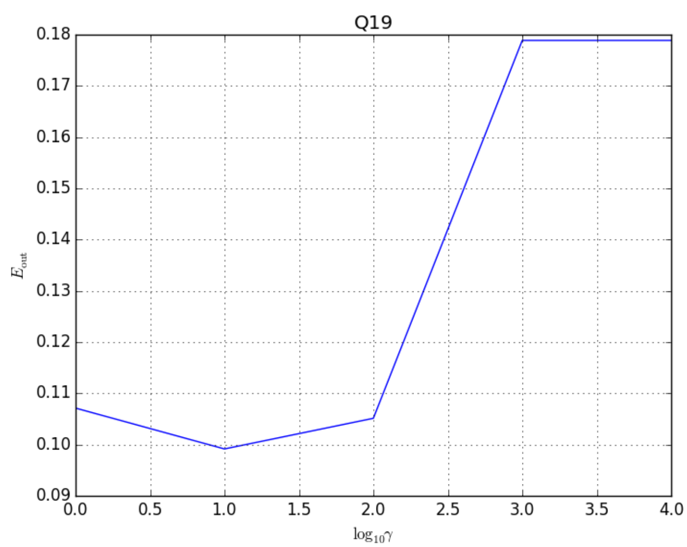

Figure 5: Q19

The minimum occurs at $\gamma = 10$. E_{out} becomes larger as γ becomes larger.

□

Problem 20

The corresponding γ of $\min E_{\text{val}}$ are 10 in 99 results, one result is 1.

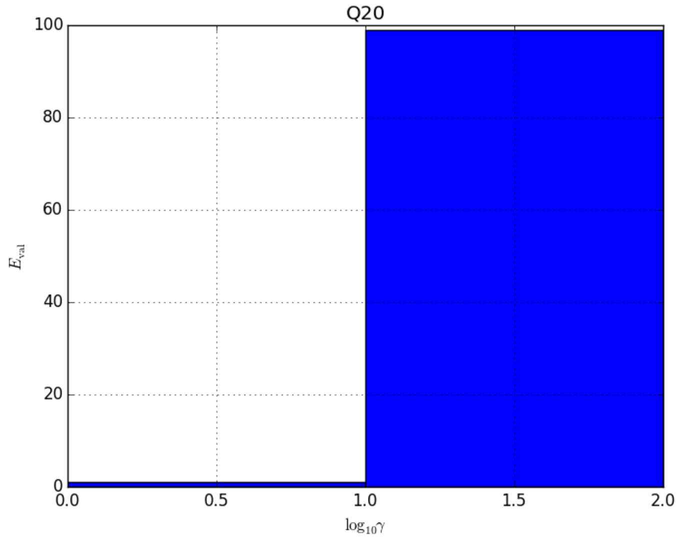

Figure 6: Q20

Over almost all random choices, $\gamma = 10$ always minimizes E_{val}.

□

Problem 21

Hard-margin SVM dual:

$$\min_{\boldsymbol{\alpha}} \left(\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n \right) \text{ subject to } \sum_{n=1}^{N} y_n \alpha_n = 0,\ \alpha_n \geq 0,\ \forall n \quad (67)$$

then we have

$$\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n \right) - \sum_{n=1}^{N} \beta_n \alpha_n \quad (68)$$

Use $\max_{\beta_n \geq 0, \forall n}$ to eliminate condition $\alpha_n \leq 0$ since if there exists some $\alpha_n < 0$ for some n, then $\max_{\beta_n \geq 0, \forall n} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \to \infty$. So the dual SVM is

$$\min_{\boldsymbol{\alpha}} \left(\max_{\beta_n \geq 0, \forall n} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \right) \quad (69)$$

Hence the Lagrangian dual problem is

$$\min_{\boldsymbol{\alpha}} \left(\max_{\beta_n \geq 0, \forall n} \mathcal{L}\left(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) \right) \geq \max_{\beta_n \geq 0, \forall n} \left(\min_{\boldsymbol{\alpha}} \mathcal{L}\left(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) \right) \tag{70}$$

At optimal, we have

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 = y_i \left(\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n^T \right) \mathbf{z}_i - 1 - \beta_i \Rightarrow \beta_i = y_i \left(\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n^T \right) \mathbf{z}_i - 1 \tag{71}$$

Substitute this into (70), we have

$$\begin{aligned} \max_{\beta_n \geq 0, \forall n} \mathcal{L}\left(b, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) &= \max_{\beta_n \geq 0, \forall n} \left(\left(\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n \right) \right. \\ &\qquad \left. - \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m + \sum_{n=1}^{N} \alpha_n \right) && (72) \\ &= \max_{\beta_n \geq 0, \forall n} \left(-\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m \right) && (73) \end{aligned}$$

which is

$$\begin{aligned} \min_{\beta_n \geq 0, \forall n} \left(\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m \right) &= \min_{\beta_n \geq 0, \forall n} \left(\frac{1}{2} \left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 \right) && (74) \\ &= \min_{\beta_n \geq 0, \forall n} \left(\frac{1}{2} \left\| \mathbf{w} \right\|^2 \right) && (75) \end{aligned}$$

This is the same as the original problem, but subject to

$$\beta_n \geq 0 \Rightarrow y_n \left(\sum_{m=1}^{N} \alpha_m y_m \mathbf{z}_m^T \right) \mathbf{z}_n = y_n \mathbf{w}^T \mathbf{z}_n \geq 1 \tag{76}$$

this is different from

$$y_n \left(\mathbf{w}^T \mathbf{z}_n + b \right) \geq 1 \tag{77}$$

The reason that b is missing is due to the optimal condition $\sum_{n=1}^{N} \alpha_n y_n = 0$.

□

Reference

[1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.

[2] Quadratic Programming with Python and CVXOPT

https://courses.csail.mit.edu/6.867/wiki/images/a/a7/Qp-cvxopt.pdf

[3] scikit-learn documentation: 1.4. Support Vector Machines

http://scikit-learn.org/stable/modules/svm.html#multi-class-classificat