HW3 – Gal Kaptsenel 209404409

Q1

1.1

Let $C = \{v_1, v_2, v_3, v_4, v_5\}$ points in $\mathbb{R}^2$.
There is a subset $S \subset C$of at <u>most</u> 4 points, each of the points in S has an extreme value in at least one of the x/y axes (if there are two points with the same max/min value, take one of them). The size of S is at most 4 because there are two axes (x and y), and for each of them we will take a point with maximum value and (maybe another) point with minimum value.
Any rectangle that contains S must contain all the points in C because given a point $v_i \in C$,

$$\exists v_j, v_k \in S, v_{j_x} \leq v_{i_x} \leq v_{k_x}$$
$$\exists v_j, v_k \in S, v_{j_y} \leq v_{i_y} \leq v_{k_y}$$

Because S contains the points from C with the maximum and minimum coordinates in both axes.
S is at most with size 4, therefore there is $v_i \in C \backslash S$, given the labeling that labels v_i with **false** and the rest of the points with **true**.
Suppose there exists a rectangle that contains all the points with the **true** label and does not contain the single point (v_i) with the **false** label. This rectangle contains all points in the set $C \backslash \{v_i\}$, and because $v_i \notin S, S \subset C$, the rectangle contains all the points in S. Therefore, this rectangle contains all points in C, and explicitly also $v_i \in C$, which contradicts the assumption that the rectangle does not contain v_i. Thus, the assumption is incorrect and there is no such rectangle.
Therefore, H_{rec} cannot shatter C, and because C is subset of any 5 points in $\mathbb{R}^2$, we can conclude that $VCdim(H_{rect}) < 5$.

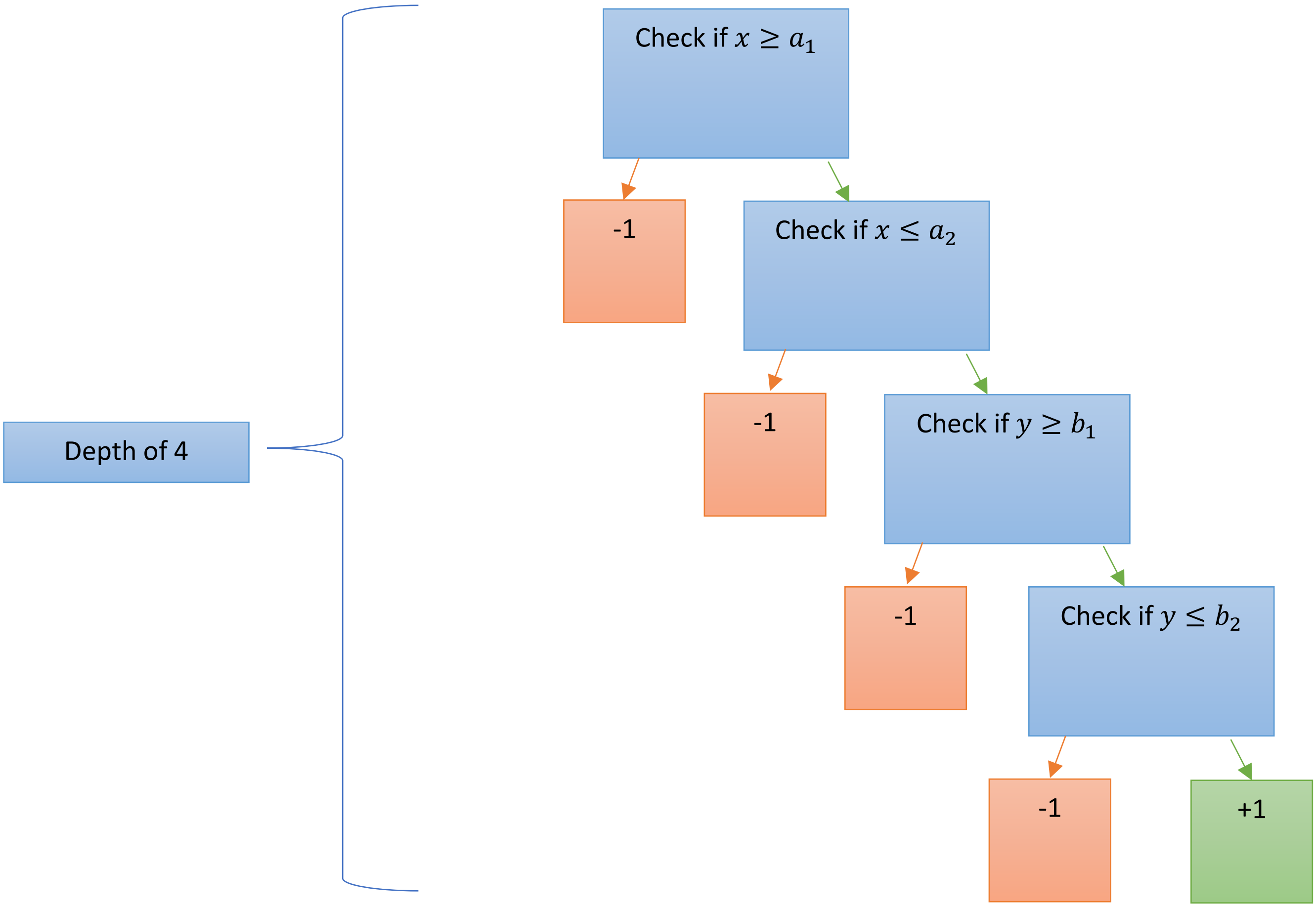
1.2

Given two hypothesis classes such that $H_1 \subseteq H_2$. Let C be a set with size of $VCdim(H_1)$ that is shattered by H_1, exists such set by the definition of $VCdim(H_1)$.
Given labeling for set C, there exists a hypothesis $h \in H_1 \subseteq H_2$which completely agrees with the labeling, and therefore H_2 shatters C.
Therefore, $VCdim(H_2) \geq |C| = VCdim(H_1)$, because there exists a set with size $VCdim(H_1)$ which shattered by H_2

1.3

It could be said that $VCdim(H_{DT}) \geq 4$.
Given $h = h_{(a_1,a_2,b_1,b_2)} \in H_{rect}$, we could construct a decision tree h_t that implements h(i.e. makes the exact same prediction as h).
The tree, which has a depth of 4, will make the following decision for a given (x, y) ,
Check whether the x coordinate is between a_1 and a_2, if no, return -1 (2 comparisons, therefore this check requires a depth of 2), <u>afterwards</u>, check whether the y coordinate is between b_1 and b_2, if no return -1, otherwise return 1 (2 comparisons, therefore requires additional depth of 2). The total depth of this tree will be 4, $h_t \in H_{DT}$.
For any given point (x, y) for prediction, the described tree will return a prediction of 1, iff $a_1 \leq x \leq a_2$ and $b_1 \leq y \leq b_2$, otherwise it will return -1, and therefore it makes the exact same predictions as $h = h_{(a_1,a_2,b_1,b_2)} \rightarrow h_{(a_1,a_2,b_1,b_2)} = h_t \in H_{DT}$.
$H_{rect} \subseteq H_{DT}$
Therefore, according to 1.1 and 1.2 above, $4 = VCdim(H_{rect}) \leq VCdim(H_{DT})$.

Visualization of h_t:



Q2

Let $n_3 = n_1 + n_2$,$\phi_3 : \chi \to \mathbb{R}^{n_3}$

$\phi_3(w) = \begin{pmatrix} 2\phi_1(w) \\ 3\phi_2(w) \end{pmatrix}$

Therefore,

$$K_3(u,v) = 4K_1(u,v) + 9K_2(u,v) = 2 \cdot \phi_1^T(u) \cdot 2\phi_1(v) + 3\phi_2^T(u) \cdot 3\phi_2(v)$$
$$= (2\phi_1^T(u) \quad 3\phi_2^T(u)) \begin{pmatrix} 2\phi_1(v) \\ 3\phi_2(v) \end{pmatrix} = \phi_3^T(u) \cdot \phi_3(v) = \langle \phi_3(u), \phi_3(v) \rangle$$

Moreover, it could be seen that K_3 is valid, according to the kernel algebra (lecture SVM slide 46):

it is given that K_1, K_2 are valid kernels, therefore, according to rule 3 form kernel algebra, $4K_1$ and $9K_2$ are valid kernels as well, and according to rule 4 from kernel algebra, $4K_1 + 9K_2 = K_3$ is a valid kernel.

Q3

3.1

$\forall z_1, z_2 \in C, \forall t \in [0,1],$

$t \cdot q(z_1) + (1-t) \cdot q(z_2) = t \cdot \max\{f(z_1), g(z_1)\} + (1-t) \cdot \max\{f(z_2), g(z_2)\} \overset{t,1-t\geq 0}{=}$

$\max\{t \cdot f(z_1), t \cdot g(z_1)\} + \max\{(1-t) \cdot f(z_2), (1-t) \cdot g(z_2)\} \overset{1}{\geq}$

$\max\{t \cdot f(z_1) + (1-t) \cdot f(z_2), t \cdot g(z_1) + (1-t) \cdot g(z_2)\} \overset{2}{\geq}$

$\max\{f(t \cdot z_1 + (1-t) \cdot z_2), g(t \cdot z_1 + (1-t) \cdot z_2)\} = q(t \cdot z_1 + (1-t) \cdot z_2)$

1. $\max\{t \cdot f(z_1), t \cdot g(z_1)\} \geq t \cdot f(z_1), t \cdot g(z_1)$

   $\max\{(1-t) \cdot f(z_2), (1-t) \cdot g(z_2)\} \geq (1-t) \cdot f(z_2), (1-t) \cdot g(z_2)$

   And therefore,

   $\max\{t \cdot f(z_1), t \cdot g(z_1)\} + \max\{(1-t) \cdot f(z_2), (1-t) \cdot g(z_2)\}$

   $\geq t \cdot f(z_1) + (1-t) \cdot f(z_2), t \cdot g(z_1) + (1-t) \cdot g(z_2)$

   and therefore,

   $\max\{t \cdot f(z_1), t \cdot g(z_1)\} + \max\{(1-t) \cdot f(z_2), (1-t) \cdot g(z_2)\} \geq$
   $\max\{t \cdot f(z_1) + (1-t) \cdot f(z_2), t \cdot g(z_1) + (1-t) \cdot g(z_2)\}$
2. f, g are convex functions.

In conclusion, by definition, $q(z)$ is convex $w.r.t. z$, because
$\forall z_1, z_2 \in C, \forall t \in [0,1],$
$t \cdot q(z_1) + (1-t) \cdot q(z_2) \geq q(t \cdot z_1 + (1-t) \cdot z_2)$

3.2

According to the rule from Tutorial 01, that states that any linear function $g(x) = a^T x + b$ is convex (slide 13, lemma 1), we can conclude that both the functions

- $0 = \mathbf{0}^T w + 0$
- $1 - y_i w^T x_i \overset{w \text{ and } x_i \text{ are row and column vectors}}{=} 1 - y_i x_i^T w \overset{define\ a_i^T = -y_i x_i^T}{=} a_i^T w + 1$

Are convex $w.r.t\ w$.

According to 3.1 above, $\max\{0, 1 - y_i w^T x_i\}$ is convex $w.r.t\ w$

3.3

According to slide 12 from tutorial 7, $\|w\|^2$ is convex, and according to the given two lemma 1 in this question, and because $\lambda \in \mathbb{R}_{\geq 0}$, $\lambda\|w\|^2$ is convex.

According to given lemma 2 in this question and 3.2 above, $\sum_{i=1}^{m} \max\{0, 1 - y_i w^T x_i\}$ is convex.

According to lemma 1 in this question, $\frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i w^T x_i\}$ is convex.

According to lemma 2 in this question, $\frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i w^T x_i\} + \lambda\|w\|^2$ is convex.

In addition, $\mathbb{R}^d$ is convex set because,
$\forall a_1, a_2 \in \mathbb{R}^d, \forall t \in [0,1], t \cdot a_1 + (1-t) \cdot a_2 \in \mathbb{R}^d$, because $\mathbb{R}^d$ is a vector space.
Therefore, according to the property from slide 15, tutorial 7, which states that if we restrict a convex function to a convex subset, then it is a convex function,

the above $f(w) = \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i w^T x_i\} + \lambda\|w\|^2$, $f|_{\mathbb{R}^d}$ is convex.

We can conclude that $f|_{\mathbb{R}^d}$ is convex and therefore the problem $\underset{w \in \mathbb{R}^d}{\operatorname{argmin}}\ f(w)$,

i.e. Soft-SVM, is a convex optimization problem.