Short HW3 – SVM, Optimization, and PAC learning

1. Define $\mathcal{H} = \{x \mapsto sign(w^t x): w \in \mathbb{R}^d\}$, the hypothesis class of homogeneous linear classifiers.

   1.1. In Tutorial 05, we said that the VC-dimension of homogeneous linear classifiers is $\geq d$. Provide a rigorous proof for this statement.

Answer 1.1:

We want to show that **there exists a set of d points** that is shattered by H.

Formally:

Given $C = \{x_1, \dots, x_d\} \subset X$, $H\ shatters\ C\ iff\ \forall y_1, \dots, y_d \in Y : \exists h \in H: \forall x_i \in C: h(x_i) = y_i$

Denote $C = \{e_1, \dots, e_d\}\ and\ let\ \{y_1, \dots, y_d\}\ some\ labeling.$

Denote $h : x \to (sign(y_1, \dots, y_d)x) \in H.$

Given e_j: $h(e_j) = \left((y_1, \dots, y_d)e_j\right) = sign(y_j)$, $\forall\ 1 \leq j \leq d$

Thus $H\ shatters\ C.$

1.2. Prove that $VCdim(\mathcal{H})$ is exactly d by proving that $VCdim(\mathcal{H}) < d + 1$.

Hint: Any set of $\{x_1, \dots, x_{d+1}\}$ vectors in $\mathbb{R}^d$ is linearly dependent, and at least one vector in the set (w.l.o.g x_{d+1}) satisfies $x_{d+1} = \sum_{i=1}^{d} z_i x_i$ for some scalars $z_1, \dots, z_d \in \mathbb{R}$ with at least some scalar that is not equal to 0.

Answer 1.2:

We want to show that **any set of d+1 points** cannot be shattered by H.

Given any set $C = \{x_1, \dots, x_{d+1}\} \subset \mathbb{R}^{d+1}$ is linearly dependent. Hence there is a non-trivial linear combination that satisfies $x_{d+1} = \sum_{i=1}^{d} \alpha_i x_i$.

For the set C denote the following labels:

$$y_i = \begin{cases} sign(\alpha_i), & a_i \neq 0\ \wedge i < d+1 \\ 1\,, & a_i = 0\ \wedge i < d+1 \\ -1\,, & i = d+1 \end{cases}$$

Assume that where exists $w \in \mathbb{R}^{d+1}$ that defines $h \in H$ that shatters C.

Then the prediction $\hat{y}_{d+1}$ for x_{d+1} would be:

$\hat{y}_{d+1} = h(x_{d+1}) = h\left(\sum_{i=1}^{d} \alpha_i x_i\right) = sign\left(w^T \sum_{i=1}^{d} \alpha_i x_i\right) = sign\left(\sum_{i=1}^{d} w^T \alpha_i x_i\right) =$
$sign\left(\sum_{i=1}^{d} \alpha_i y_i\right) = 1\ \neq y_{d+1}$ in contrediction.

Thus, H do not shatters C of size d+1 so $VCdim(H) < d + 1$

Together with 1.1 we have that $VCdim(H) = d.$

2. Let $\phi: \mathcal{X} \to \mathbb{R}^{n_1}, \phi': \mathcal{X} \to \mathbb{R}^{n_2}$ be two feature mappings where $n_1, n_2 \in \mathbb{N}$.

   Let $K, K': (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ be two valid kernels defined as:

   $$K(u,v) = \langle \phi(u), \phi(v) \rangle = \sum_{i=1}^{n_1} \phi_i(u)\phi_i(v), \quad K'(u,v) = \langle \phi'(u), \phi'(v) \rangle = \sum_{j=1}^{n_2} \phi'_j(u)\phi'_j(v).$$

   Prove that $G(u,v) \triangleq K(u,v) \cdot K'(u,v)$ is a valid kernel. That is, propose a feature mapping $\psi: \mathcal{X} \to \mathbb{R}^{n_3}$ for some $n_3 \in \mathbb{N}$, such that $G(u,v) = \langle \psi(u), \psi(v) \rangle$.

   Hint: You should use $n_3 = n_1 \cdot n_2$.

Answer 2:

We want to show that $G(u,v) = K(u,v) \cdot K'(u,v) = <\psi(u), \psi(v)> \forall (u,v) \in X \times X \; and \; a \; given \; \psi: X \to \mathbb{R}^{n_3}$.

Let $\phi: X \to \mathbb{R}^{n_1}, \phi': X \to \mathbb{R}^{n_2} \; s.t. \; K(u,v) = <\phi(u), \phi(v)>, \; K'(u,v) = <\phi'(u), \phi'(v)>,$

Define $n_1 \cdot n_2 = n_3, \; \psi: X \to \mathbb{R}^{n_3} \; by \; \psi_{i,j \in n_1 \times n_2}(x) = \phi_i(x) \cdot \phi'_j(x).$

Consider the following:

$$G(u,v) = K(u,v) \cdot K'(u,v) = <\phi(u), \phi(v)> \cdot <\phi'(u), \phi'(v)>$$

$$= \sum_{\substack{i=1 \\ |n_1 \times n_2| = n_3}}^{n_1} \phi_i(u)\phi_i(v) \cdot \sum_{j=1}^{n_2} \phi'_j(u)\phi'_j(v) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} (\phi_i(u)\phi'_j(u))(\phi_i(v)\phi'_j(v))$$

$$= \sum_{i,j=1} \psi_{i,j}(u)\psi_{i,j}(v) = <\psi(u), \psi(v)>.$$

3. For a given parameter $\gamma > 0$, define the Gaussian Kernel for 1-D input in the following manner:

$$K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad K(a,b) = \exp(-\gamma(a-b)^2)$$

3.1. Provide a feature mapping $\phi: \mathbb{R} \to \mathbb{R}^p$ with $p \in \mathbb{N} \cup \{\infty\}$, and prove that K is indeed a valid kernel

Hint: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

Answer 3.1:

We want to find a feature mapping $\phi: \mathbb{R} \to \mathbb{R}^p \; s.t. K(a,b) = <\phi(a), \phi(b)> =$
$\sum_{n=0}^{\infty} \frac{(-\gamma(a-b)^2)^n}{n!} = e^{-\gamma(a-b)^2}$.

Define $\phi_n(x) = e^{-\gamma x^2} \cdot \sqrt{\frac{(2\gamma)^n}{n!}} \cdot x^n$, $\forall\ 0 \leq n \leq p$

Given $(a,b) \in \mathbb{R} \times \mathbb{R}$, all together we have:

$$<\phi(a), \phi(b)> = \sum_{n=0}^{p} \phi_n(a) \cdot \phi_n(b) = \sum_{n=0}^{p} e^{-\gamma a^2} \cdot \sqrt{\frac{(2\gamma)^n}{n!}} \cdot a^n \cdot e^{-\gamma b^2} \cdot \sqrt{\frac{(2\gamma)^n}{n!}} \cdot b^n$$

$$= \sum_{n=0}^{p} e^{-\gamma(a^2+b^2)} \cdot \frac{2\gamma^n}{n!} \cdot (ab)^n = e^{-\gamma(a^2+b^2)} \sum_{n=0}^{p} \frac{(2\gamma ab)^n}{n!} = e^{-\gamma(a^2+b^2)} \cdot e^{2\gamma ab}$$

$$= e^{-\gamma(a^2-2ab+b^2)} = e^{-\gamma(a-b)^2} = K(a,b)$$

3.2. Assume that you are given a very large dataset with 1-D samples. We would like to apply the Gaussian Kernel to train a classifier on the dataset. Would it be better to optimize the **primal problem** with the feature mapping you found, or is it better to optimize the **dual problem** with the kernel that we defined? Is it even possible? Explain.

Answer 3.2:

For a very large database with 1-D samples, I would prefer to use the dual problem with the Gaussian kernel to train the classifier, mostly for its computational performance.
In the primal problem, after choosing $\phi(x)$ we will apply $x_i' = \phi(x_i)$ for all x_i in the dataset and only then use SVM algorithm to learn on the modified data (x_i', y_i). For any new sample x we will classifie it by applying $w^T\phi(x)$. Hence, we will need to calculate $\phi_n(x)$ for any coordinate $0 \leq n \leq p$ and store the result for later predictions. If p is large the computation will be long, multiplied by the size of the database will be even longer and storing the massive data could be challenging. If p is infinite this solution might be not possible, we cannot store infinite vector and the runtime will be infinite as well.
On the other hand, in the dual problem, on training time we will go over all pairs in the dataset and store the K(a,b) result in the kernel matrix. In this case each pair calculation is O(1) since we

have a close formula for K(a,b) and the memory cost is the size of the dataset squared. Then we could apply Dual SVM and get w which contains a_i that indicated the support vectors. Then for prediction we don't need to implicitly do inner product with w and sample x rather then we could calculate $f_\alpha(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x)$. But since dual svm optimizes directly over α , we expect to have many $a_i = 0$ thus the prediction is computationally faster, each element is O(1) and the sum is at most O(|database|) but we expect it so be faster.

4. **Refute** (with a simple example)**:** Let $f, g: \mathbb{R} \to \mathbb{R}$ be two convex functions.

   The composition $h \triangleq f \circ g$ (that is, $h(x) = f\big(g(x)\big)$) is also a convex function.

Answer 4:

Denote $f(x) = e^{-x}$, $g(x) = -\ln\left(\sqrt{x}\right)$, $h(x) = f\big(g(x)\big) = \sqrt{x}$

First, we will show that f(x) and g(x) are convex. Notice that

1. $f'(x) = -e^{-x}, f''(x) = e^{-x} > 0 \; \forall x \in \mathbb{R}$ => f(x) is convex
2. $g'(x) = -\frac{1}{2x}$, $g''(x) = \frac{1}{2x^2} > 0 \; \forall \; 0 \neq x \in \mathbb{R}$ => g(x) is convex

Now we will show that h(x) is not convex!

$h'(x) = \frac{1}{2\sqrt{x}}$, $h''(x) = -\frac{1}{4x^{\frac{3}{2}}} = -\frac{1}{4x \cdot \sqrt{x}} < 0 \; \forall \; 0 < x \in \mathbb{R}$ => h(x) is not convex!

5. We will now prove that the following Soft-SVM problem is convex:

$$\underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \max\{0,1-y_i\cdot w^\top x_i\} + \lambda\|w\|_2^2$$

Let $f, g: C \to \mathbb{R}$ be two convex functions defined over a convex set C.

Lemma (no need to prove)**:** $q(z) \triangleq \max\{f(z),\ g(z)\}$ is convex w.r.t z.

Lemma (no need to prove)**:** the sum of <u>any</u> number of convex functions is convex.

5.1. Prove (by definition): Given a constant $\alpha \in \mathbb{R}_{\geq 0}$, the function $\alpha f(z)$ is convex w.r.t z.

<u>Answer 5.1:</u>

Given f(x) is convex by definition:

$\forall x_1, x_2 \in C\ , t \in [0,1]: f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$

Then simply by multiplying the inequality by a non-negative scaler we have:

$\alpha f(tx_1 + (1-t)x_2) \leq \alpha t f(x_1) + \alpha(1-t)f(x_2)$ hence, we showed $\alpha f(z)$ is convex by def.

5.2. Using a rule from Tutorial 07, conclude that $\max\{0, 1-y_i\boldsymbol{w}^\top \boldsymbol{x}_i\}$ is convex w.r.t $\boldsymbol{w}$.

<u>Answer 5.2:</u>

In tutorial 07 we proved that any linear function is convex over a convex set.

First of all, notice that $\mathbb{R}^d$ is a convex set (using vectors linearity).

Now, denote $f(w) = 0\ , \forall i \in [m]:\ g_i(w) = 1 - y_i w^T x_i\ , q_i(w) = \max\{f(w), g_i(w)\}$.

$f(w)$ is trivially convex over a convex set.

Let $w_1, w_2 \in \mathbb{R}^d,\ fix\ i \in [m], t \in [0,1]$, we will show $g_i(w) = 1 - y_i w^T x_i$ is convex with respect to w.

$$\begin{aligned} g_i(tw_1 + (1-t)w_2) &= 1 - y_i(tw_1 + (1-t)w_2)^T x_i = 1 - y_i(tw_1^T + (1-t)w_2^T)x_i \\ &= 1 - ty_i w_1^T x_i - (1-t)y_i w_2^T x_i \leq 1 + t - ty_i w_1^T x_i - (1-t)y_i w_2^T x_i \\ &= t(1 - y_i w_1^T x_i) + (1-t)(1 - y_i w_2^T x_i) = tg_i(w_1) + (1-t)g_i(w_2) \end{aligned}$$

Thus, for all $i \in [m]$, $g_i(w)$ is convex over a convex set.

Denote

Using the given first lemma we have that $q_i(w)$ is the maximum of two convex functions, hence $q_i(w)$ is convex itself, over the same convex set.

<u>MORE ANSWERS ON THE NEXT PAGE</u>

5.3. Using the above (and properties from Tutorial 07), conclude that the Soft-SVM optimization problem is convex w.r.t $\boldsymbol{w}$.

Answer 5.3:

$$\underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \max\{0,1-y_i\cdot w^\top x_i\} + \lambda\|w\|_2^2$$

In Tutorial 07, using the 2nd derivate test we showed that $p(w) = \left|\left|w\right|\right|_2^2$ is convex w.r.t.w.
Since $\lambda \geq 0$ using answer 5.1 we have $\lambda p(w) = \lambda\left|\left|w\right|\right|_2^2$ is also convex w.r.t.w.
Using the 2nd lemma provided, $r(w) = \sum_{i=1}^{m} \max\{0,1-y_i w^T x_i\} = \sum_{i=1}^{m} q_i(w)$ is also convex as a sum of convex functions, also w.r.t.w.
Since $\frac{1}{m} \geq 0$ using answer 5.1 we have $z(w) = \frac{1}{m} r(w)$ is also convex w.r.t.w.
Lastly, in Tutorial 07 we showed that argmin of a convex function $z(w)$ over a convex set $\mathbb{R}^d$, is convex, w.r.t.w.
It is given from the definition of argmin:

$$\arg\min_{x\in S} f(x) := \{x \in S \ : \ f(s) \geq f(x) \text{ for all } s \in S\}$$

Since $z(w)$ is convex over a convex set, it achieves its global minimum. Meaning the set $\{w \in \mathbb{R}^d : z(x) \geq z(w)\ \forall x \in \mathbb{R}^d\} \neq \emptyset$ and of course, if w is in the set, z(w) is smaller\equal to any other z(x) since w is the minimum. We also saw in the tutorial that if a function is convex over a convex set, if it has more then one minimum, then they're function values are all equal. Any local minimum is also global in that case. Thus, the convexity inequality still holds.
All together we have that Soft-SVM problem is convex w.r.t.w.