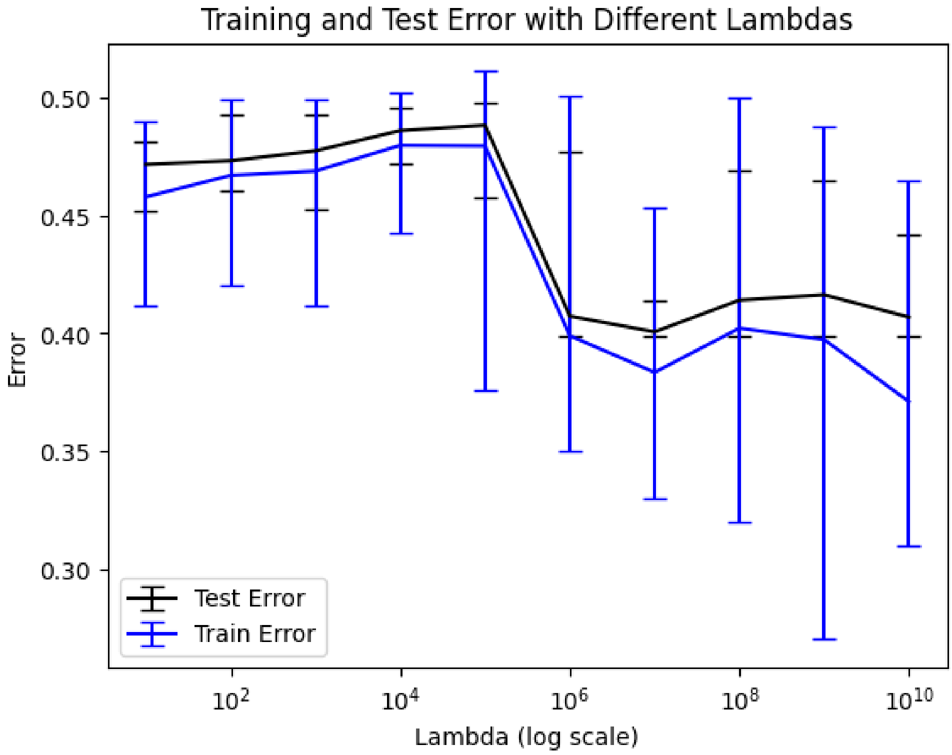
Introduction to Machine Learning

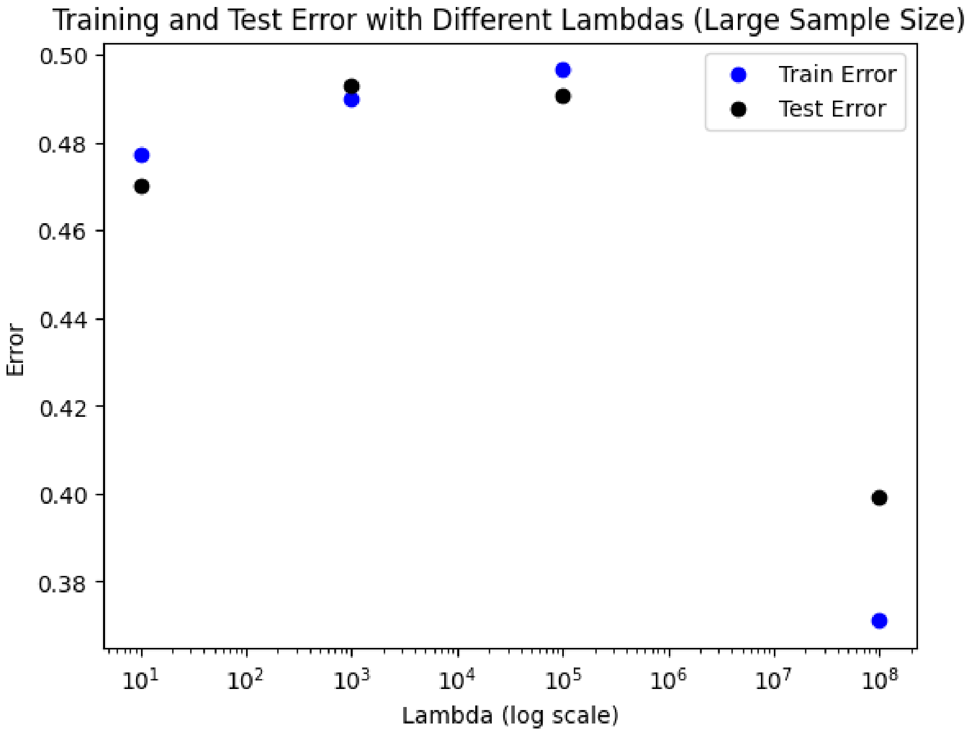
Exercise 2

<u>Question 2:</u>

a.



b.



c. The expected errors, regarding the sample size, we thought the large sample size would have better results, especially on the training data, due to more data that would help determine the best predictor, with less probability for overfitting and a small error rate. We discovered it is not entirely true, the error rates in different sample sizes and with same λ values were quite similar, probably because the random sampling captures a trend. As we saw in class, when increasing the λ value we will penalize more for a soft margin, meaning we will have a harder margin. So, the training data error will decrease when we increase the λ value, because on large λ we have a large hinge loss and hence we make fewer errors for the training data, as our graph shows. In the test data, we expect that when increasing the λ too much we might be overfitting because we don't allow mistakes on the training samples, even though we might need to make them for better results. Our graph shows the same trend in the test data as it showed in the training data, with a slight increase in the error rate on the large λ, maybe for even larger λs we will see this increase more clearly.

Question 3:

a. Running the ERM algorithm with hypothesis class $\mathcal{H}_n$ for $\boldsymbol{n} < \mathbf{10}$, may not be a good idea. We know that the doctor's decision rule has 10 characters at most. If the decision rule has 10 characters exactly, this $\mathcal{H}_{n<10}$ does not have the hypothesis that labels using the same condition as the doctor. We do not have a promise on other conditions and their hypothesis may cause a large error, or just larger than the hypothesis that uses the doctor's condition. In addition, running the ERM algorithm with the hypothesis class $\mathcal{H}_n$ for $\boldsymbol{n} > \mathbf{10}$ may not be a good idea as well, because we might have overfitting. The hypothesis that the ERM will choose will perform very well on the sample, but not that well on the distribution, because it contains conditions that match the specific samples. Therefore, the model will not be able to generalize or learn the patterns to use it on unseen data.

b. Since we know the doctor's decision rule has 10 characters at most, the hypothesis that decides according to this rule is in $\mathcal{H}_{10}$. Therefore, there exist such $h_c * \in \mathcal{H}_{10}$ that $err(h_c *, D) = 0$, what makes it the realizable case. The PAC bounds for the realizable case: $m \geq \frac{log(|\mathcal{H}_{10}|)+log(1/\delta)}{\epsilon}$. To compute the minimum size of m, we will place $\delta = 0.01$ (which gives probability of 0.99), and $\epsilon = 0.1$. In addition, we need to compute and place in the formula $|\mathcal{H}_{10}|$. The definition for $\mathcal{H}_n$ is:

$$\mathcal{H}_n := \{h_c \mid c\ is\ a\ condition\ which\ can\ be\ described\ using\ at\ most\ n\ characters\}$$

We know there are 128 ASCII characters, where each character in a condition has 128 options.

Let $\left|\widehat{\mathcal{H}}_n\right| := \{h_{c} \mid c\ is\ a\ condition\ which\ can\ be\ described\ using\ \boldsymbol{exactly}\ n\ characters\}$

therefore: $|\mathcal{H}_{10}| = \left|\widehat{\mathcal{H}}_1\right| + \cdots + \left|\widehat{\mathcal{H}}_{10}\right| = \sum_{i=1}^{10} 128^{i}$. Now to the computation:

$$m \geq \frac{log\left(\sum_{i=1}^{10} 128^{i}\right) + log(1/0.01)}{0.1} = 531.3$$

For a sufficient sample size the minimum of m is **532**.

c. Since we know the doctor's decision rule has 10 characters at most, we can't be sure the condition that describes this rule has an hypothesis in $\mathcal{H}_n$ for $n < 10$, as we explained in section (a). Hence, we are in the agnostic case. The PAC bound for the agnostic case is: $m \geq \frac{2log(|\mathcal{H}_n|)+2log(2/\delta)}{\epsilon^2}$. We place in this formula $\delta = 0.01$, $\epsilon = 0.1$, and receive $m \geq \frac{2log(|\mathcal{H}_n|)+2log(2/0.01)}{0.1^2} = \frac{log\left(40000\cdot|\mathcal{H}_n|^2\right)}{0.01}$. To find the upper bound for $n < 10$ that will give a minimum value for a sufficient sample size we can choose $n = 9$, and then $|\mathcal{H}_9| = 128^9$.

$$m \geq \frac{log(40000 \cdot |\mathcal{H}_n|^2)}{0.01} = \frac{log\left(40000 \cdot (\sum_{i=1}^{9} 128^{i})^2\right)}{0.01} = 9794.887$$

For a sufficient sample size the minimum of m is **9795**.

Question 4:

Given a hypothesis class of homogeneous linear predictors $\mathcal{H}_{\mathrm{L}}^{d} := \{h_w \mid w \in \mathbb{R}^d\}$ where $h_w(x) = sign(< w, x >), \forall x \in \mathbb{R}^d$, we need to prove that any linearly independent set of d vectors $u_1, \dots, u_d \in \mathbb{R}^d$ are shattered by $\mathcal{H}_{\mathrm{L}}^{d}$.

First, we will see that the set of the d basis vectors of $\mathbb{R}^d$: $\{e_1, \dots, e_d\}$ are shattered by $\mathcal{H}_{\mathrm{L}}^{d}$, meaning that they can get all the possible labelings by predictors from $\mathcal{H}_{\mathrm{L}}^{d}$:

Since each basis vector e_i $(1 \leq i \leq d)$ has one coordinate in the ith position with the value

of 1, and the rest of the coordinates with the value of 0, e_i's label using a certain h_w hypothesis, will be determined by ith coordinate of w, where $w = [y_1, \dots, y_d]$,
$y_1, \dots, y_d \in -1, +1$. More formally:

$$h_w(e_i) = sign(< w, e_i >) = sign\left(\sum_{k=1}^{d} e_i\,(k) \cdot w(k)\right) = sign\big(w(i)\big) = sign(y_i) = y_i$$

Since the label of e_i is determined by the ith coordinate of w, each coordinate of w will be responsibole for each label of a different basis vectors. $\mathcal{H}_{\mathrm{L}}^{d}$ has all h_w for any $w \in \mathbb{R}^d$, hence any possible labeling can be reached by one $w \in \mathbb{R}^d$, and therefore the basis vectors of $\mathbb{R}^d$: $\{e_1, \dots, e_d\}$ are shattered by $\mathcal{H}_{\mathrm{L}}^{d}$.

To prove that any linearly independent set of d vectors $u_1, \dots, u_d \in \mathbb{R}^d$ are shattered by $\mathcal{H}_{\mathrm{L}}^{d}$, we address to the fact that the basis vectors span the space, thus any vector in $\mathbb{R}^d$ can be represented as a linear combination of the basis vectors, including any linearly independent vectors.

Question 5

a. We will write a quadratic minimization problem with a set of constraints equivalent to the problem.

In addition to the hinge loss part which is replaced in the original soft-SVM optimization problem, we also have the $||w||_1$ which is basically defined as the sum of absolute values of w's products. Since absolute value is not a quadratic form, we will replace it as well. Let's define the following variables:

$$\xi_i = \ell_h(w, (x_i, y_i);\ \forall i \in [1 \dots m]|$$

$$\gamma_i = |w_i|\ ;\ \forall i\ \in [1 \dots d]$$

$$\theta = \sum_{i=1}^{d} \gamma_i$$

Now let's define the corresponding quadratic minimization problem:

$$min_{w \in \mathbb{R},\ \xi_1 \dots \xi_m,\ \gamma_1 \dots \gamma_d,\ \theta}\ \lambda\theta^2 + \sum_{i=1}^{m} \xi_i$$

With the following constraints:

a. $\xi_i \geq 0;\ \forall i \in [1 \dots m]$

b. $y_i \langle w, x_i \rangle \geq 1 - \xi_i\ ;\ \forall i \in [1 \dots m]$

c. $\gamma_i \geq w_i\ ;\ \forall i \in [1 \dots d]$

d. $\gamma_i \geq -w_i\ ;\ \forall i \in [1 \dots d]$

e. $\theta \geq \sum_{i=1}^{d} \gamma_i$

f. $\theta \geq -\sum_{i=1}^{d} \gamma_i$

b. Based on the above definitions, we will define H, u, A, v as they should be set to solve the quadratic programming problem we defined.

$$\xi_{vec} \coloneqq \begin{bmatrix} \xi_1 \\ \cdot \\ \cdot \\ \cdot \\ \xi_m \end{bmatrix}, \gamma_{vec} \coloneqq \begin{bmatrix} \gamma_1 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_d \end{bmatrix}, X_{mat} \coloneqq \begin{bmatrix} x_1^T \\ \cdot \\ \cdot \\ \cdot \\ x_m^T \end{bmatrix}, Y_{mat} \coloneqq\ diag\{y_1, \dots y_m\}$$

$$\boldsymbol{z} = (\xi_{vec}, \gamma_{vec}, \theta, w) \in \mathbb{R}^{(m+2d+1)\times 1}$$

$$\boldsymbol{H} = 2\lambda \begin{bmatrix} 0_{m\times m} & 0_{m\times d} & 1 & 0_{m\times d} \\ 0_{d\times m} & 0_{d\times d} & 0 & 0_{d\times d} \\ 0 & 0 & 0 & 0 \\ 0_{d\times d} & 0_{d\times d} & 0 & 0_{d\times d} \end{bmatrix} \in \mathbb{R}^{(m+2d+1)\times(m+2d+1)}$$

$$\boldsymbol{u} = (1_{m\times 1}, 0_{d\times 1}, 0\ , 0_{d\times 1})^T \in \mathbb{R}^{(m+2d+1)\times 1}$$

$$\boldsymbol{v} = (0_{m\times 1}, 1_{m\times 1}, 0_{d\times 1}\ , 0_{d\times 1}, 0, 0\ )^T \in \mathbb{R}^{(2m+2d+2)\times 1}$$

$$\boldsymbol{A} = \begin{bmatrix} I_m & 0_{m\times d} & 0_{m\times 1} & 0_{m\times d} \\ Y_{mat}X_{mat} & 0_{m\times d} & 0_{m\times 1} & I_m \\ 0_{d\times m} & I_d & 0_{d\times 1} & -I_d \\ 0_{d\times m} & I_d & 0_{d\times 1} & I_d \\ 0_{1\times d} & -1_{1\times d} & 1 & 0_{1\times d} \\ 0_{1\times d} & 1_{1\times d} & 1 & 0_{1\times d} \end{bmatrix} \in \mathbb{R}^{(2m+2d+2)\times(m+2d+1)}$$

Question 6

a. We will prove the claim for any $i \leq d,\ \left|w^{(t+1)}(i)\right| \leq t$ using induction on the perceptron update:

$$w^{(t+1)} \leftarrow w^{(t)} + \ y_j x_j$$

Base case: $t = 1$

$$\left|w^{(t+1)}(i)\right| = \ \left|w^{(2)}(i)\right| \underset{\substack{perceptron\ update\\ step}}{=} \left|(w^{(1)} + y_k x_k)(i)\right|$$

$$= \ |(0, \dots ,0) + y_k x_k)(i)| = \ |y_k x_k(i)|\,.$$

Based on the definition of $y\ and\ x$:

$$y_k x_k = (-1)^{i+1} * \left((-1)^i, \dots, (-1)^{i+1}, 0, \dots ,0\right)$$

So $x_k(i) = \{0, 1, -1\}$, hence $|y_k x_k(i)| \leq |\pm 1 \mid 0| \leq 1 = \boldsymbol{t}$

So $|y_k x_k(i)| \leq |\pm 1| \leq 1 = \boldsymbol{t}$

Assumption: assume the induction hypothesis holds for iteration $n = t$ so that

for any $i \leq d,\ \left|w^{(t+1)}(i)\right| \leq t$

Induction step: we need to show the following inequality hold for $n + 1 = t\ + 1$, i.e: for any $i \leq d,\ \left|w^{(t+2)}(i)\right| \leq t + 1.$

Proof:

$$\left|\boldsymbol{w}^{(t+2)}(\boldsymbol{i})\right| \underset{\substack{perceptron\ update\\ step}}{=} \left|(w^{(t+1)} + \ y_k x_k)(i)\right| \underset{triangle\ inequality}{\leq} \left|(w^{(t+1)}(i)\right| + \mid y_k x_k(i)|$$

$$\underset{\substack{Induction\\ hypothesis}}{\leq} t + \mid y_k x_k(i)| \underset{\substack{\mid y_k x_k(i)| \leq 1\\ as\ we\ showed\ above\\ in\ the\ base\ case}}{\leq} \boldsymbol{t+1}$$

b. We will prove the claim for every coordinate $j,\ \left|w^{(T)}(i)\right| \geq 2^{i-1}$ using induction on the coordinate j. We will prove the above for a stronger statement: $\forall i\ \in \{1, \dots, d\}$, $w^{(T)}(i) \geq 2^{i-1}$, since we are dealing with positive numbers ($2^{i-1} > 0$), it will naturally lead to proving $\left|w^{(T)}(i)\right| \geq 2^{i-1}$

We will use the following **fact(*)**:

$w^{(T)}$ is a separator, hence by definition it correctly labels all examples in the given sample S, so the following holds: $\forall i \ \in \{1, \dots, d\},\ \ y_i\langle w^{(T)}, x_i \ \rangle > 0$

Base case: $j = 1$

$$(*)\ y_1\langle w^{(T)}, x_1 \ \rangle \underset{\substack{defintion\ of\ x_i, y_i \\ and\ inner\ product}}{=} (-1)^2 * \left(w^{(T)}(1)x_1(1) + w^{(T)}(2) * 0 \dots w^{(T)}(d) * 0\right) = \ w^{(T)}(1)x_1(1)$$

$$= \ w^{(T)}(1) * (-1)^2 = \ \boldsymbol{w^{(T)}(1)} \underset{based\ on\ (*)\ above}{\geq} 1 = \ 2^0$$

$$y_1\langle w^{(T)}, x_1 \ \rangle \underset{based\ on\ the\ fact\ (*)\ above}{=} \boldsymbol{w^{(T)}(1)} \underset{\substack{based\ on\ the\ fact\ above \\ for\ \ y_1\langle w^{(T)}, x_1 \ \rangle}}{>} 0 \underset{\substack{coordinates\ of\ w^{(T)} \\ are\ whole\ numbers}}{\geq} \geq 1 \ = \ 2^0$$

Assumption: assume the induction hypothesis holds for every coordinate $n < j$.

Induction step: we need to show the following inequality hold for $n = j$, i.e $\left|w^{(T)}(n)\right| \geq 2^{n-1}$

$$y_i\langle w^{(T)}, x_i\rangle = \sum_{j=1}^{d} w^{(T)}(j) * y_i x_i(j)$$

$$\underset{\substack{defintion\ of\ x_i, y_i \\ and\ inner\ product}}{=} \sum_{j=1}^{d} w^{(T)}(j) * \left((-1)^{i+1} * \left((-1)^{i}, \dots, (-1)^{i}, (-1)^{i+1}, 0, \dots, 0\right)\right)(j)$$

$$= \ (-1)^{i+1} \sum_{j=1}^{d} w^{(T)}(j) * \left((-1)^i, \dots, (-1)^i, (-1)^{i+1}, 0, \dots, 0\right) \underset{\substack{based\ on\ the\ fact \\ \forall i\ \ y_i\left\langle w^{(T)}, x_i \right\rangle \ > \ 0}}{>} 0$$

Observation: based on the sample definition and the fact above that $w^{(T)}$ label correctly all the samples of S, we can say that:

$$\forall i \ \in \{1, \dots, d\}, \qquad y_i\langle w^{(T)}, x_i \ \rangle \geq 1$$

So, we will turn > 0 into ≥ 1.

Now we have split it into cases:

Case 1: i is even

In that case we will get:

$$(-1)^i \sum_{j=1}^{d} w^{(T)}(j) * \left((-1)^i, \dots, (-1)^i, (-1)^{i+1}, 0, \dots, 0\right) =$$

$$1 * \left(\sum_{j=1}^{d} w^{(T)}(j) * (-1, \dots, -1, 1, 0, \dots, 0) \right) \geq 1$$

$$1 * \left(-w^{(T)}(1) - w^{(T)}(2) \dots + w^{(T)}(j) + 0 \dots 0\right) \geq 1$$

$$\longrightarrow w^{(T)}(j) \geq 1 + \sum_{k=1}^{j-1} w^{(T)}(k) \underset{induction\ assumption}{\geq} 1 + \sum_{k=1}^{j-1} 2^{k-1}$$

$$= 1 + \frac{1}{2}\sum_{k=1}^{j-1} 2^k \underset{geometric\ sequence\ sum}{=} 1 + \frac{1}{2} * 2 * \left(2^{j-1} - 1\right) = 2^{j-1} - 1 + 1 = 2^{n-1}$$

Case 2: i is odd

In that case we will get:

$$(-1)^i \sum_{j=1}^{d} w^{(T)}(j) * \left((-1)^i, \dots, (-1)^i, (-1)^{i+1}, 0, \dots, 0\right) =$$

$$-1 * \left(\sum_{j=1}^{d} w^{(T)}(j) * (1, \dots, 1, -1, 0, \dots, 0) \right) \geq 1$$

$$-1 * \left(w^{(T)}(1) + w^{(T)}(2) \dots - w^{(T)}(j) + 0 \dots 0\right) \geq 1$$

$$\longrightarrow w^{(T)}(j) \geq 1 + \sum_{k=1}^{j-1} w^{(T)}(k) \underset{induction\ assumption}{\geq} 1 + \sum_{k=1}^{j-1} 2^{k-1}$$

$$= 1 + \frac{1}{2}\sum_{k=1}^{j-1} 2^k \underset{geometric\ sequence\ sum}{=} 1 + \frac{1}{2} * 2 * \left(2^{j-1} - 1\right) = 2^{j-1} - 1 + 1 = 2^{n-1}$$

In both cases, we got that: $w^{(T)} \geq 2^{n-1}$ as required.

c. Let T be the number of updates the perceptron algorithm performs until it stops and output $w^{(T)}$.

So, using our proof from item a above, we can say that:

$$\forall i \;\; 1 \le i \le d, \qquad \left|w^{(T)}(i)\right| \le T - 1$$

Using our proof from item b above, we can say that

$$\forall i \;\; 1 \le i \le d, \qquad \left|w^{(T)}(i)\right| \ge 2^{i-1}$$

Combining both, we will get the following:

$$\forall i \;\; 1 \le i \le d, \qquad T - 1 \ge \left|w^{(T)}(i)\right| \ge 2^{i-1}$$

This holds also for $i = d$ hence:

$$T - 1 \ge \left|w^{(T)}(d)\right| \ge 2^{d-1}$$

$$\longrightarrow T \ge \; 2^{d-1} + 1 = O(2^d)$$