Q1 Short Answer Questions and Multiple Choices

18 Points

Instructions:

- This is an open book exam -- you can use course materials posted at CCLE as your reference. Please do not access any other material during the exam. Discussion is strictly prohibited.
- This exam booklet contains **seven** problems with a total of 100 points.
- We highly recommend taking the exam in the official exam period 2/18 11:30AM -- 2/18 2:30PM. If you are facing a technical issue, contact us immediately. You must finish the exam by Saturday 2/19 11:30AM. No late submissions will be accepted; therefore, make sure to submit well before the deadline!
- You can ask **private** questions in Piazza, but only clarification questions will be answered. If there is an issue with the exam, we will make announcements during the official exam period (i.e., before Friday 12/18 2:30PM).
- If you think the question is ambiguous or if you have additional concerns, feel free to write down your explanations and comments at the end of the exam (i.e., Question 8).
- Note that we are not able to provide customized rubrics and many questions do not have partial credits.
- You can type your answer using markdown. For example, type `$$\gamma$$` will generate γ.
- **Remember to frequently save your answer**

Best of luck!

Q1.1 Spam Filter Experiment

3 Points

Z is a summer intern working on spam classification in your company. The dataset consists of 10 million non-spam emails (class 0) and 10 thousand spam emails (class 1). Z considers the following steps of conducting experiments:

- Step 1: Shuffle the dataset and split it into the train, validation, and test sets.
- Step 2: Train logistic regression models on the train set with different hyper-parameters.
- Step 3: Identify the best hyper-parameter using the validation set and report the results on the test set in accuracy.

Do you agree with the above experimental setup?
If No, what is the major issue? Provide your suggestions in one or two sentences.

No. There are 10 million emails in case 0 and 10 thousand emails in case 1. This skew goes against the model assumptions for logistic regression.

Q1.2 Code Review

3 Points

Continue Q1.1. Z decides to use the scikit-learn library for data standardization before training the model. Z sends you the following code snippet for code review

```
....
\# X_train, X_val and X_test contain train, val and test data
scaler = preprocessing.StandardScaler()
X_train_std = scaler.fit_transform(X_train)
X_val_std = scaler.fit_transform(X_val)
X_test_std = scaler.fit_transform(X_test)
\# Z uses X_train_std, X_val_std and X_test_std for train, validatio
...
```

Would you approve Z’s code? If No, please show how to correct the code.

Hint: the following scikit-learn documentation might be useful:

- fit()- Compute the mean and std to be used for later scaling.
- transform()- Perform standardization by centering and scaling. Should be called afterfit().
- fit_transform()- Fit the data, then transform it.

No. I would not approve Z's code because they only need to use fit_transform() on the training data. You need to transform the validation and testing data to standardize them.

Q1.3 ID3

2 Points

Which of the following statement(s) about ID3 algorithm are correct? Select all of them.

☐ The ID3 algorithm always finds the optimal decision tree, i.e., the decision treewith the minimal depth that can classify all training instances.

☐ The ID3 algorithm can be only used in binary classification problems.

☑ ID3 algorithm can be used to find a non-linear classifier.

☑ Decision trees can be implemented as a set of if-then-else statements.

Q1.4 Multi-class

2 Points

Which of the following statement(s) related to multi-class classification are correct? Select all of them.

☑ One-vs-One strategy decomposes a multi-class classification problem into several binary classification problems.

☑ For the same multi-class classification problem, it is possible that some of the corresponding binary classification problems are not linearly separable when using the One-against-All strategy, while the binary classification problems are all linearly separable when using the One-vs-One strategy.

□ Imbalanced training set size is a common issue with One-vs-One strategy.

☑ One-vs-One strategy requires to train more binary classifiers than the One-against-All strategy when the number of classes is 5.

Q1.5 Regression

3 Points

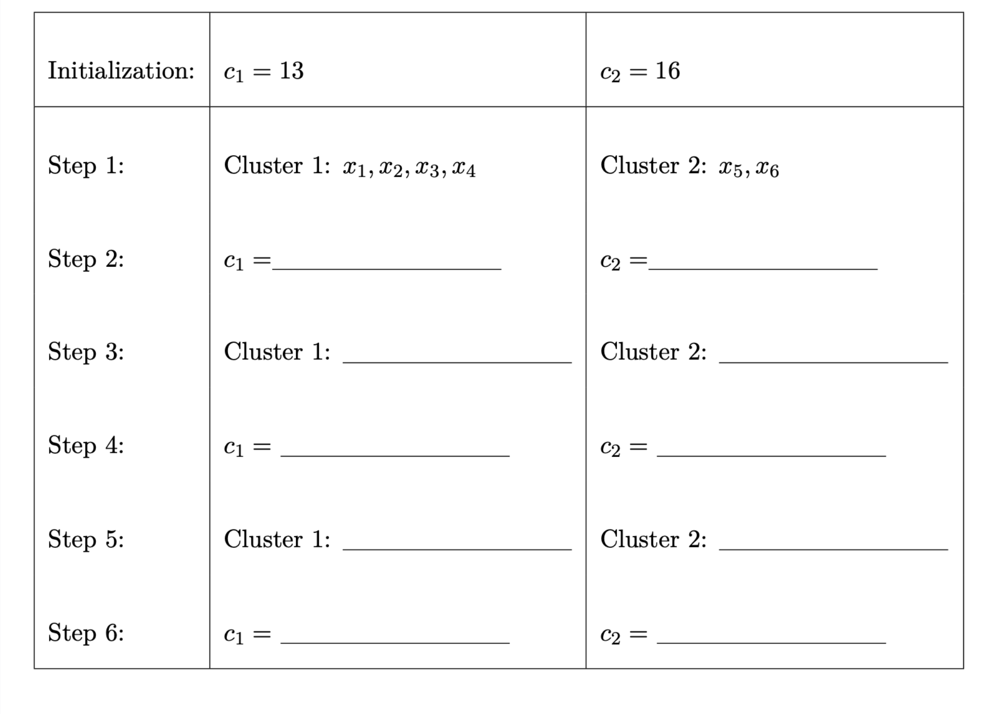
Consider a linear regression model $f : y = 0.5x + 1$.
Given a set of data points $D = \{(x_i, y_i)\}_{i=1}^{3} = \{(1.0, 1.6), (1.5, 1.5), (3.0, 2.4)\}$.
What is the mean squared error of the model f on D?
Write down your final answer.

0.0275

Q1.6 K-Means

5 Points

Consider $x \in \mathbb{R}$ and assume we have six data points $x_1 = 2$, $x_2 = 3$, $x_3 = 7$, $x_4 = 12$, $x_5 = 15$, $x_6 = 18$. In the following, we are going to apply K-means algorithm with $K = 2$ on the following six points. The initial centers $c_1 = 13$, $c_2 = 16$. Complete the following table.



Initialization:	$c_1 = 13$	$c_2 = 16$
Step 1:	Cluster 1: x_1, x_2, x_3, x_4	Cluster 2: x_5, x_6
Step 2:	$c_1 =$________________	$c_2 =$________________
Step 3:	Cluster 1: ________________	Cluster 2: ________________
Step 4:	$c_1 =$ ________________	$c_2 =$ ________________
Step 5:	Cluster 1: ________________	Cluster 2: ________________
Step 6:	$c_1 =$ ________________	$c_2 =$ ________________

Step 2 c_1:

6

Step 2 c_2:

16.5

Step 3 Cluster 1:

x_1, x_2, x_3

Step 3 Cluster 2:

x_4, x_5, x_6

Step 4 c_1:

4

Step 4 c_2:

15

Step 5 Cluster 1:

x_1, x_2, x_3

Step 5 Cluster 2:

x_4, x_5, x_6

Step 6 c_1:

4

Step 6 c_2:

15

Q2 Maximum Likelihood Estimation

9 Points

In this question, we will explore how can we predict the number of passengers waiting in LAX at a specific time. Flights could be delayed for various reasons including weather, humidity, and heavy traffic on the airport runways. Suppose that we have collected all those related features and converted them to numerical variables: x_n.

We also managed to accurately count the number (non-negative integer) of passengers y_n waiting in LAX on day n from Jan 1, 2017 to Dec 31, 2017 to construct a training data set $\{(x_n, y_n)\}_{n=1}^{N}$, where $x_n \in \mathbb{R}^M, y_n \in \{0, 1, 2, \ldots\}$.

We assume the mapping between target variable y and the input feature vector x can be modeled by a Poisson distribution with parameter θ:

$P(Y = y|X = x; \theta) = \frac{\lambda^y}{y!} \cdot e^{-\lambda}$ where $\lambda = e^{\theta^T x}$

Use the knowledge you learned about MLE to answer the following questions.

Q2.1 Likelihood

3 Points

What is the likelihood function for one specific training example (x_n, y_n)?

○ $P(y_n|x_n, \theta) = y_n(\theta^T x_n) - \theta^T x_n + constant$

○ $P(y_n|x_n, \theta) = \frac{1}{y_n!} e^{y_n \theta^T x_n - \theta^T x_n}$

● $P(y_n|x_n, \theta) = \frac{1}{y_n!} e^{y_n \theta^T x_n} e^{-e^{\theta^T x_n}}$

○ $P(y_n|x_n, \theta) = \frac{1}{y_n!} (e^{\theta^T x_n})^{y_n} e^{-\theta}$

○ $P(y_n|x_n, \theta) = (\theta^T x_n)^{y_n} - \theta^T x_n + constant$

Q2.2 Log Likelihood

3 Points

If we assume the training examples are drawn i.i.d. from the underlying data distribution, what is the log-likelihood of the training set $\{(x_n, y_n)\}_{n=1}^N$?

○ $\mathcal{LL}(\theta) = constant + \sum_{n=1}^N y_n e^{\theta^T x_n} - \theta$

○ $\mathcal{LL}(\theta) = constant \cdot \prod_{n=1}^N y_n \theta^T x_n - \theta^T x_n$

○ $\mathcal{LL}(\theta) = constant + \sum_{n=1}^N (\theta^T x_n)^{y_n} - \theta^T x_n$

○ $\mathcal{LL}(\theta) = constant + \sum_{n=1}^N y_n \theta^T x_n - \theta^T x_n$

● $\mathcal{LL}(\theta) = constant + \sum_{n=1}^N y_n \theta^T x_n - e^{\theta^T x_n}$

Q2.3 Convexity

3 Points

Considering the above Poisson regression model, which of the following statement is correct?
Hint: $e^{\theta^T x}$ is a convex function w.r.t θ.

○ $\mathcal{LL}(\theta)$ is a convex function.

● $\mathcal{LL}(\theta)$ is a concave function.

○ $\mathcal{LL}(\theta)$ is not a convex nor a concave function.

○ $\mathcal{LL}(\theta)$ is both a convex nor a concave function.

Q3 Hard Gaussian Mixture Model

14 Points

In the following, we consider a **hard-assignment** GMM. The **hard-assignment** GMM is similar to the soft-assignment GMM we learned in the class.
However, instead of having $\gamma_{nk} \in [0, 1]$, in the hard-assignment GMM, γ_{nk} is a Boolean variable that is set to 1 if and only if the data point n belongs to cluster k.
Formally, we consider a data set consists of N i.i.d. data points $\{x_n \in \mathbb{R}\}_{n=1}^N$ and our goal is to cluster them into K groups using K Gaussians $\mathcal{N}(x_n; \mu_k, \sigma_k^2), k = 1, 2, \ldots, K$. The prior probability of the cluster k is ω_k. We use $\theta = \{\omega_k, \mu_k, \sigma_k\}_{k=1}^K$ to represent all model parameters and z_n is a random variable that represents the cluster assignment of the n-th data point.

Probability density function of Gaussian distribution: $\mathbf{P}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

Q3.1 Interpretation of gamma

2 Points

What is the value of $\sum_{k=1}^K \gamma_{nk}$ for the n-th data point?

$\sum_{k=1}^K \gamma_{nk} =$

1

Q3.2 Derivation

4 Points

GMM is an iterative algorithm. We alternatively update γ_{nk} and θ. Given a fixed γ_{nk}, we derive the optimal θ in the following.

Please complete the following derivation by filling the blanks:

$$
\begin{aligned}
\mathcal{LL}(\theta) &= \sum_{n=1}^{N} \log P(x_n, z_n|\theta) \\
&= \sum_{n=1}^{N} \log \prod_{k=1}^{K} P(x_n, z_n|\theta)^{\gamma_{nk}} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \underline{\hspace{3cm}} \log P(x_n, z_n|\theta) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\{\underline{\hspace{8cm}} + \log P(z_n|\theta)\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \mathcal{N}(x_n; \mu_k, \sigma_k^2) + \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log w_k
\end{aligned}
$$

blank #1:

γ_{nk}

blank #2:

$logP(x_n|\theta)$

Q3.3 Hard GMM M-step

4 Points

Let $\mu_l^* = \arg\max_{\mu_l} \mathcal{LL}(\theta)$ be the optimal solution of maximizing $\mathcal{LL}(\theta)$ with respect to μ_l. Which of the following equations are true? Select all of them.

□ $\mu_l^* = \arg\max_{\mu_l} \ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \ \mathcal{N}(x_n; \mu_k, \sigma_k^2)$

☑ $\mu_l^* = \arg\min_{\mu_l} \ \sum_{n=1}^{N} \gamma_{nl} \log \sigma_l^2 + \frac{1}{2}\gamma_{nl}(x_n - \mu_l)^2 + \gamma_{nl} \log \ w_l$

☑ $\mu_l^* = \arg\max_{\mu_l} \ \sum_{n=1}^{N} \gamma_{nl} \log \ \mathcal{N}(x_n; \mu_l, \sigma_l^2)$

☑ $\mu_l^* = \arg\min_{\mu_l} \ \sum_{n=1}^{N} \gamma_{nl}(x_n - \mu_l)^2$

Q3.4 Hard GMM E-step

4 Points

To update γ_{nk}, we assign each data point to the cluster with the largest $P(z_n \mid x_n; \theta)$
(i.e., $z_n^* = \arg\max_k P(z_n = k \mid x_n; \theta)$).
Which of the following statements are true? Select all of them.

□ $z_n^* = \arg\max_k \mathcal{N}(x_n; \mu_k, \sigma_k^2)$

☑ $z_n^* = \arg\max_k \omega_k \mathcal{N}(x_n; \mu_k, \sigma_k^2)$

☑ $z_n^* = \arg\max_k \log \omega_k + \log \mathcal{N}(x_n; \mu_k, \sigma_k^2)$

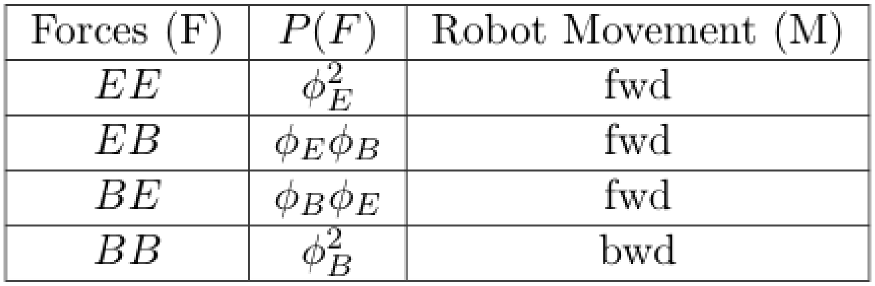
□ $z_n^* = \arg\max_k \log \omega_k$

Q4 Expectation Maximization (EM)

13 Points

Suppose we have a simulator that simultaneously exhibits 2 field forces f_1, f_2 on a robot. Both field force generators **independently** generate field forces E or B with probabilities ϕ_E, ϕ_B, respectively ($\phi_E + \phi_B = 1$).
In this simulator, the robot is exposed to two combination forces at each step. We do not observe the field forces but we observe the robot's movement moving either forward (**fwd**) or backward (**bwd**).

Specifically, the robot moves forward when the combination of the two field forces is either EE, EB, or BE. On the other hand the robot moves backwards if the field force combination is BB.



Forces (F)	$P(F)$	Robot Movement (M)
EE	ϕ_E^2	fwd
EB	$\phi_E\phi_B$	fwd
BE	$\phi_B\phi_E$	fwd
BB	ϕ_B^2	bwd

We conduct N experiment trials, and we observe that the robot moves forward n_{fwd} times and moves backward n_{bwd} times ($n_{fwd} + n_{bwd} = N$). Our objective is to determine ϕ_E and ϕ_B using EM algorithm based on the robot's movements in these N experiment trials.

As a reminder, EM is an iterative process. In the E-Step, we estimate the numbers of expected force combinations, $\mathrm{n}_{EE}, \mathrm{n}_{BB}, \mathrm{n}_{BE}$ n_{EB} in the N trials ($\mathrm{n}_{EE} + \mathrm{n}_{BE} + \mathrm{n}_{EB} + \mathrm{n}_{BB} = N)$ based on ϕ_E and ϕ_B . In the M-Step, we estimate ϕ_E and ϕ_B based on $\mathrm{n}_{EE}, \mathrm{n}_{BB}, \mathrm{n}_{BE}$ n_{EB}.

Q4.1 E-Step

4 Points

We first derive the E-Step. Suppose we are given initial estimates ϕ_E and ϕ_B, calculate the expectations of counts n_{EE}, n_{EB}, n_{BE} and n_{BB} from the observed data.

○ $n_{EE} = (n_{fwd} + n_{bwd})\phi_E; n_{BB} = (n_{fwd} + n_{bwd})\phi_B; n_{EB} = n_{BE} = (n_{fwd} + n_{bwd})\phi_B\phi_E$

○ $n_{EE} = n_{fwd}\frac{\phi_E}{2\phi_E\phi_B+\phi_B^2}; n_{BB} = n_{bwd}; n_{EB} = n_{BE} = (n_{fwd} + n_{bwd})\phi_B\phi_E$

● $n_{EE} = n_{fwd}\frac{\phi_E^2}{2\phi_E\phi_B+\phi_E^2}; n_{BB} = n_{bwd}; n_{EB} = n_{BE} = n_{fwd}\frac{\phi_E\phi_B}{2\phi_E\phi_B+\phi_E^2}$

○ $n_{EE} = n_{fwd}\frac{\phi_E^2}{2\phi_E\phi_B+\phi_B^2}; n_{BB} = n_{bwd}\frac{\phi_B^2}{\phi_B^2+2\phi_E\phi_B}; n_{EB} = n_{BE} = (n_{fwd} + n_{bwd})\frac{\phi_E\phi_B}{2\phi_E\phi_B+\phi_E^2}$

Q4.2 Likelihood

3 Points

Assume if we know the expected numbers of force combinations are $\mathrm{n}_{EE}, \mathrm{n}_{BB}, \mathrm{n}_{BE}\ \mathrm{n}_{EB}$. Which of the following correspond to the likelihood function?

● $\phi_E^{2\mathrm{n}_{EE}+\mathrm{n}_{EB}+\mathrm{n}_{BE}}\ \phi_B^{2\mathrm{n}_{BB}+\mathrm{n}_{EB}+\mathrm{n}_{BE}}$

○ $\phi_E^{\mathrm{n}_{EE}+\mathrm{n}_{EB}}\ \phi_B^{\mathrm{n}_{BB}+\mathrm{n}_{BE}}$

○ $\phi_E^{\mathrm{n}_{EE}+\mathrm{n}_{EB}+\mathrm{n}_{BE}}\ \phi_B^{\mathrm{n}_{BB}+\mathrm{n}_{EB}+\mathrm{n}_{BE}}$

○ $\phi_E^{\mathrm{n}_{EE}+\mathrm{n}_{EB}+\mathrm{n}_{BE}+\mathrm{n}_{BB}}\ \phi_B^{\mathrm{n}_{EE}+\mathrm{n}_{EB}+\mathrm{n}_{BE}+\mathrm{n}_{BB}}$

Q4.3 M-Step

3 Points

Following the previous question, write down the M-Step for estimating ϕ_E and ϕ_B by maximizing the likelihood function.

● $\phi_E = \frac{2n_{EE}+n_{EB}+n_{BE}}{2N}, \phi_B = \frac{2n_{BB}+n_{EB}+n_{BE}}{2N}$

○ $\phi_E = \frac{n_{EE}+2n_{EB}}{N}, \phi_B = \frac{n_{BB}+2n_{EB}}{N}$

○ $\phi_E = \frac{n_{EE}+2n_{EB}}{N}, \phi_B = \frac{n_{BB}}{N}$

○ $\phi_E = \frac{n_{EE}}{2N}, \phi_B = \frac{n_{BB}}{2N}$

Q4.4 EM Algorithm

3 Points

What are the properties of this EM algorithm? Select all of the true statements

□ The EM algorithm is guaranteed to converge to the global optimum regardless of initialization.

☑ The EM algorithm is only guaranteed to converge to a local optimum.

☑ If we select a good initialization, it is possible that EM converges to the global optimum.

□ There is no way the EM algorithm can converge to the global optimum.

Q5 Learning Theory

22 Points

Ring Classifier: In this problem, we consider data points in 2-dimensional space $x = (x_1, x_2) \in \mathbb{R}^2$. A ring classifier assigns the label 1 to a data point x if and only if x is inside a ring. Formally, given $t \leq r$, where $t, r \in \mathbb{R}$,
a ring classifier $h_{(t,r)}$ labels data x by

$$h_{(t,r)}(x) = \begin{cases} 1, & \text{if } t \leq \|x\|_2 \leq r \\ 0, & \text{otherwise} \end{cases} .$$

In the following, we consider the ring hypothesis set

$$\mathcal{H}_{ring} = \{h_{(t,r)} : t \leq r\}.$$

We assume that the training dataset S_{train} consists of N examples x_i drawn i.i.d. from a distribution $\mathcal{D}$. The labels are provided by a target function $h^*_{(t^*,r^*)} \in \mathcal{H}_{ring}$. In addition, we use S_p to denote the set of positive examples in S_{train} (i.e. they are inside the ring of $h^*_{(t^*,r^*)}$) and use S_n to denote the set of negative examples in S_{train}.

Q5.1

4 Points

Let $\mathcal{A}$ be an algorithm that learns to select a hypothesis $\mathcal{A}(S_{train}) \in \mathcal{H}_{ring}$ from the training dataset S_{train}, where $\mathcal{A}(S_{train})$ is the tightest ring enclosing all positive examples in S_{train}. Specifically, $\mathcal{A}(S_{train}) = h_{(t_a,r_a)}$, where
$t_a = \min_{x \in S_p} \|x\|_2, \quad r_a = \max_{x \in S_p} \|x\|_2$.

Show that $\mathcal{A}(S_{train})$ achieves zero traning error.

Given a positive $x_p \in S_p$
$t_a \leq ||x_p||_2 \leq r_a$
$\forall x_p \in S_p$ it must have a magnitude that satisfies the above
$\therefore$ positive examples are classified as positives by $A(S_{train})$

Given a negative $x_n \in S_n$
$t > ||x_n||_2$ or $r < ||x_n||_2$ (gotten by the ring classifier)
$t_a \geq t$ and $r_a \leq r$
$||x_n||_2 < t \leq t_a$ or $||x_n||_2 > r \geq r_a$
$\forall x_n \in S_n : x_n$must be less than or equal to t_a or greater than or equal to r_a
This is because to be in S_n it must not be within the boundary between t_a and r_a
Because t_a is equal to the x with the min magnitude and because r_a is equal to the x with the greatest magnitude. $\forall x_i \in S_{train}$
$\therefore$ negative examples are classified as negatives by $A(S_{train})$

Because positive and negative points are classified correctly by $A(S_{train})$
achieves zero training error.

Q5.2

4 Points

We draw another set of samples S_{test} from D as the test set.

Prove that $\mathcal{A}(S_{train})$ will not make any mistake on **negative examples** in S_{test}.

Given a negative $x_n \in S_{test}$
Our negative data point satisfies $||x_n||_2 < t$ or $||x_n||_2 > r$
We know $t \leq t_a$ and $r \geq r_a$
$||x_n||_2 < t \leq t_a$ or $||x_n||_2 > r \geq r_a$
We can see the $||x_n||_2$ is out of the ring so it is labeled as a negative.

All negative data points in S_{test} classified as negatives by $A(S_{train})$
$\therefore$ there are no mistakes made by $A(S_{train})$.

Q5.3

4 Points

We define the error as

$$\epsilon = \mathbf{P}_{x \sim \mathcal{D}} \left(h^*_{(t^*, r^*)}(x) \neq h_{(t_a, r_a)}(x) \right).$$

If we draw i.i.d. m examples from $\mathcal{D}$, what is the probability that $h_{(t_a, r_a)}$ makes no mistakes for all m examples?

○ $\left(\frac{\epsilon}{1-\epsilon}\right)^m$

○ $\left(\frac{1-\epsilon}{\epsilon}\right)^m$

○ ϵ^m

● $(1-\epsilon)^m$

Q5.4

6 Points

Show that the VC dimension $VC(\mathcal{H}_{ring}) \geq 2$ by completing the following proof.

Proof: To show $VC(\mathcal{H}_{ring}) \geq 2$, we consider the following two points: S_1 and S_2

S_1:

$(3, 4)$ where $||S_1|| = 5$

S_2:

$(6, 8)$ where $||S_2|| = 10$

$\mathcal{H}_{ring}$ can shatter these two points because for all the following label combinations, we can find the corresponding t, r such that $h_{t,r}(X)$ classifies S_1 and S_2, correctly.

When S_1 is labeled as 1 and S_2 is labeled as 1,
we pick t=

4

r=

11

When S_1 is labeled as 1 and S_2 is labeled as 0,
we pick t=

4

r=

6

When S_1 is labeled as 0 and S_2 is labeled as 1,
we pick t=

9

r=

11

When S_1 is labeled as 0 and S_2 is labeled as 0, we pick t=

1

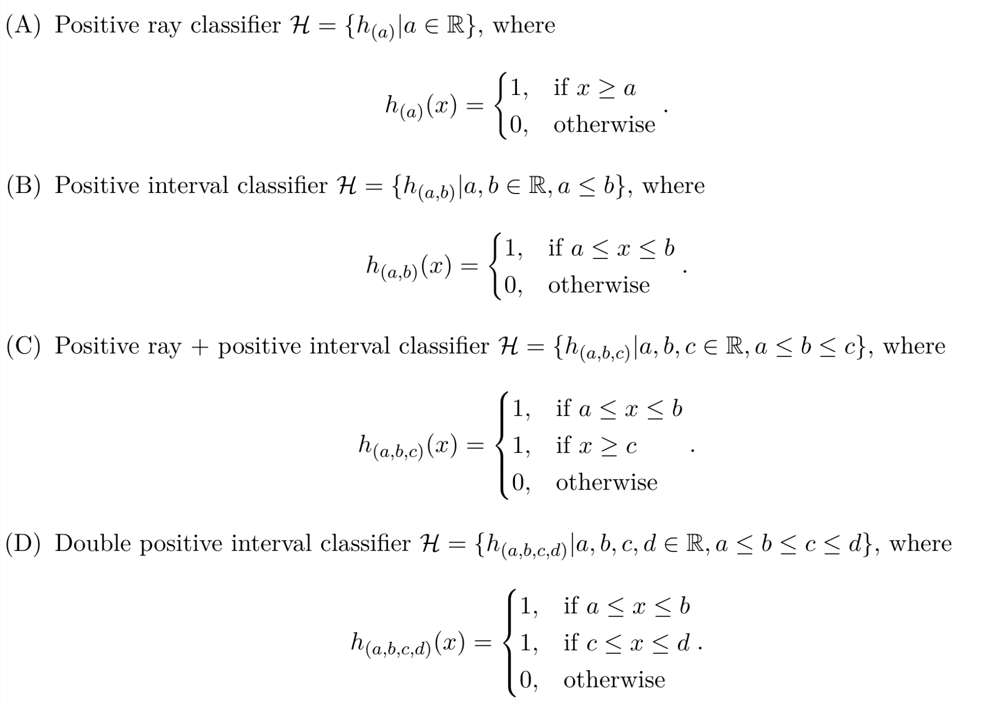
r=

1

Q5.5

4 Points

The VC dimension of $\mathcal{H}_{ring}$ in $\mathbb{R}^2$ is the same as which of the following problem(s) in $\mathbb{R}$? Select all of them.



(A) Positive ray classifier $\mathcal{H} = \{h_{(a)} | a \in \mathbb{R}\}$, where

$$h_{(a)}(x) = \begin{cases} 1, & \text{if } x \geq a \\ 0, & \text{otherwise} \end{cases}.$$

(B) Positive interval classifier $\mathcal{H} = \{h_{(a,b)} | a, b \in \mathbb{R}, a \leq b\}$, where

$$h_{(a,b)}(x) = \begin{cases} 1, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}.$$

(C) Positive ray + positive interval classifier $\mathcal{H} = \{h_{(a,b,c)} | a, b, c \in \mathbb{R}, a \leq b \leq c\}$, where

$$h_{(a,b,c)}(x) = \begin{cases} 1, & \text{if } a \leq x \leq b \\ 1, & \text{if } x \geq c \\ 0, & \text{otherwise} \end{cases}.$$

(D) Double positive interval classifier $\mathcal{H} = \{h_{(a,b,c,d)} | a, b, c, d \in \mathbb{R}, a \leq b \leq c \leq d\}$, where

$$h_{(a,b,c,d)}(x) = \begin{cases} 1, & \text{if } a \leq x \leq b \\ 1, & \text{if } c \leq x \leq d \\ 0, & \text{otherwise} \end{cases}.$$

☐ (A) Positive ray classifier

☑ (B) Positive interval classifier

☐ (C) Positive ray + Positive interval classifier

☐ (D) Double-positive interval classifier

Q6 Support Vector Machines & Kernel Trick

15 Points

Q6.1 Construct Transformation I

4 Points

Let x and $y \in \mathbb{R}^n$ be the input feature vectors. Let ϕ_a: $\mathbb{R}^n \to \mathbb{R}^k$ and ϕ_b: $\mathbb{R}^n \to \mathbb{R}^k$ be two feature transform functions. Consider two kernels $K_a(x, y)$ = $\phi_a^T(x)\phi_a(y)$ and $K_b(x, y)$ = $\phi_b^T(x)\phi_b(y)$. Let us define a new kernel $K_c(x, y)$ = 3$K_b(x, y)$ + 4 and let ϕ_c: $\mathbb{R}^n$ $\to \mathbb{R}^{k+1}$ be its corresponding feature transformation. Write down the transformation ϕ_c in terms of ϕ_a and ϕ_b.

ϕ_c =

$[\sqrt{3 * \phi_b}, 2]$

Q6.2 Construct Transformation II

4 Points

Continue Q6.1. Now, consider $K_d(x, y)$ = $K_a(x, y)$($K_b(x, y)$ + 1) and let ϕ_d: $\mathbb{R}^n \to \mathbb{R}^{k^2+k}$ be its corresponding feature transformation. Write down the transformation ϕ_d in terms of ϕ_a and ϕ_b.

Hint: you can use $\phi_a \times \phi_b$ (`$$\phi_a \times \phi_b $$`) to represent the Cartesian product between ϕ_a and ϕ_b.
If $\phi_a = [1, 2, 3], \phi_b = [1, 2], \phi_a \ \times \phi_b =$

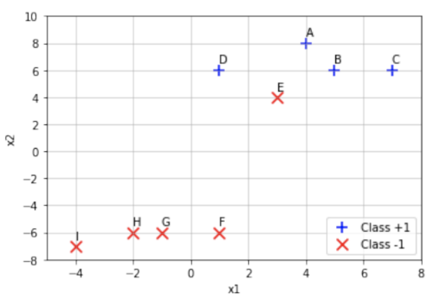
$[\phi_{a1}\phi_{b1}, \phi_{a2}\phi_{b1}, \phi_{a3}\phi_{b1}, \phi_{a1}\phi_{b2}, \phi_{a2}\phi_{b2}, \phi_{a3}\phi_{b2}] = [1, 2, 3, 2, 4, 6].$

ϕ_d =

$$[\sqrt{\phi_a \times \phi_b}, \sqrt{\phi_a}]$$

Q6.3 SVM I

4 Points



Point	(x_1, x_2)	Label
A	(4,8)	+1
B	(5,6)	+1
C	(7,6)	+1
D	(1,6)	+1
E	(3,4)	-1
F	(1,-6)	-1
G	(-1,-6)	-1
H	(-2,-6)	-1
I	(-4,-7)	-1

Figure 1: Dataset

Figure 1 provides labeled training data in a 2-D plane. We learn a $hard\ SVM$ classifier using this data. Answer the following question.

Which of the following statement(s) are True? Select all of them.

☑ (A) There are exactly 4 support vectors - B, C, D & E

□ (B) There are exactly 3 support vectors - A, B, & G

□ (C) The decision boundary of the hard SVM classifier will be $x_2 = 0$

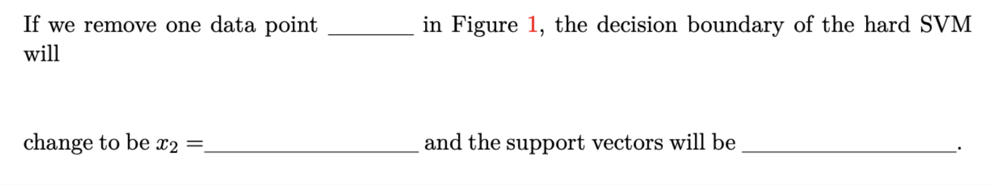
☑ (D) The decision boundary of the hard SVM classifier will be $x_2 = 5$

□ (E) The training error of the hard SVM classifier is Zero.

☑ (F) There will be no feasible solution for the hard margin SVM optimization problem if we flip the label of point B.

Q6.4 SVM II

3 Points



If we remove one data point ________ in Figure 1, the decision boundary of the hard SVM will

change to be $x_2 =$____________________ and the support vectors will be ____________________.

(blank #1) data point (use A, B, . . ., I to represent the data points):

E

(blank #2) $x_2 =$:

0

(blank #3) Support vectors are (use A, B, . . ., I to represent the data points):

B,C,D,F,G,H

Q7 Perceptron

9 Points

Q7.1

0 Points

Q7.2 Perceptron Update Rule

3 Points

Consider training a Perceptron model $y = w_1 x_1 + w_2 x_2 + b$ in the 2-dimensional feature space. If Perceptron makes a mistake on the data point (x_1, x_2) with label y where $x_1, x_2 \in \mathbb{R},\ y \in \{-1, 1\}$. Write down the update rule of w_1, w_2, and b.

Hint: In the lecture, we show a version of the Perceptron update rule where we augment the weight vector w with the bias term b. Write down the update rule for each element of the vector.

$w_1 \leftarrow$

$w_1 + y * x_1$

$w_2 \leftarrow$

$w_2 + y * x_2$

$b \leftarrow$

b

Q7.3 Perceptron Mistake Bound

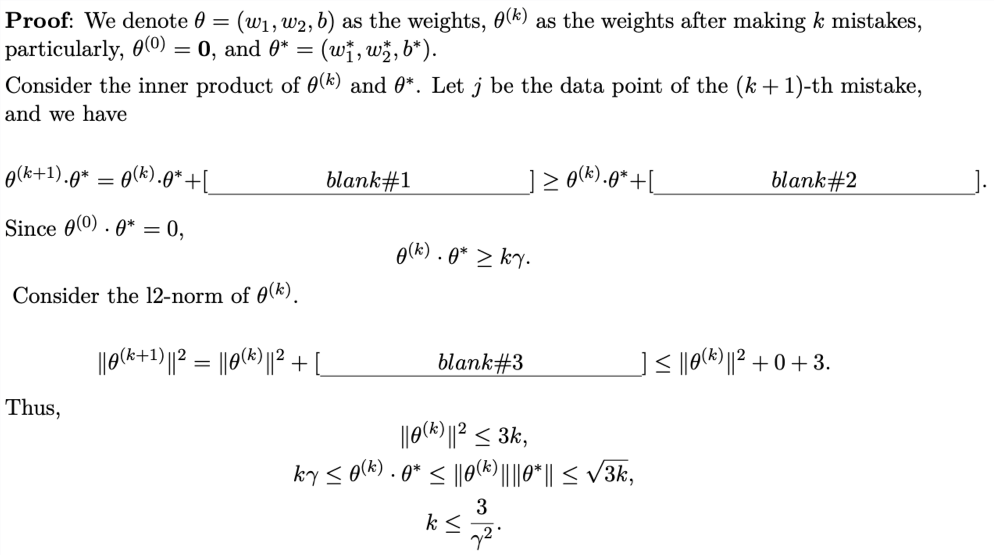
6 Points

Please fill the blanks to complete the proof of the following mistake bounds:

Given a linear separable dataset $\mathcal{D} = \{(x_1^{(i)}, x_2^{(i)}), y^{(i)}\}$ ($-1 \leq x_1^{(i)}, x_2^{(i)} \leq 1,\ y^{(i)} \in \{-1, 1\}$.) with margin γ, i.e., there exists a linear function $y = w_1^* x_1 + w_2^* x_2 + b^*$ satisfying

$$w_1^{*2} + w_2^{*2} + b^{*2} = 1, \quad \forall i,\ y^{(i)}(w_1^* x_1^{(i)} + w_2^* x_2^{(i)} + b^*) \geq \gamma.$$

If the Perceptron model is initialized as $w_1 = w_2 = b = 0$, prove that the Perceptron algorithm will make no more than $\frac{3}{\gamma^2}$ mistakes when training on D.



Proof: We denote $\theta = (w_1, w_2, b)$ as the weights, $\theta^{(k)}$ as the weights after making k mistakes, particularly, $\theta^{(0)} = \mathbf{0}$, and $\theta^* = (w_1^*, w_2^*, b^*)$.

Consider the inner product of $\theta^{(k)}$ and θ^*. Let j be the data point of the $(k+1)$-th mistake, and we have

$$\theta^{(k+1)} \cdot \theta^* = \theta^{(k)} \cdot \theta^* + [\underline{\qquad blank\#1 \qquad}] \geq \theta^{(k)} \cdot \theta^* + [\underline{\qquad blank\#2 \qquad}].$$

Since $\theta^{(0)} \cdot \theta^* = 0$,

$$\theta^{(k)} \cdot \theta^* \geq k\gamma.$$

Consider the l2-norm of $\theta^{(k)}$.

$$\|\theta^{(k+1)}\|^2 = \|\theta^{(k)}\|^2 + [\underline{\qquad blank\#3 \qquad}] \leq \|\theta^{(k)}\|^2 + 0 + 3.$$

Thus,

$$\|\theta^{(k)}\|^2 \leq 3k,$$
$$k\gamma \leq \theta^{(k)} \cdot \theta^* \leq \|\theta^{(k)}\| \|\theta^*\| \leq \sqrt{3k},$$
$$k \leq \frac{3}{\gamma^2}.$$

blank #1:

$$y^{(j)}(w_1^* x_1^j + w_2^* x_2^j + b^*)$$

blank #2:

$$\gamma$$

blank #3:

$$||x^j||^2 + 2y^j(w_1^k x_1^j + w_2^k x_2^j + b^k)$$

Q8 Additional Comment

0 Points

If you have any questions or concerns about this exam, you can type them in the following input box or upload a file. Please note that we are not able to provide customized rubrics and many questions do not have partial credits. However, if you think the question is ambiguous and you would like to provide some additional explanations to clarify your understanding of the exam questions, you can provide them here. We can only regrade the exam based on the write-up you turn in.

No files uploaded

Final Exam

● GRADED

15 MINUTES LATE

STUDENT

WILLIAM CULVER RANDALL

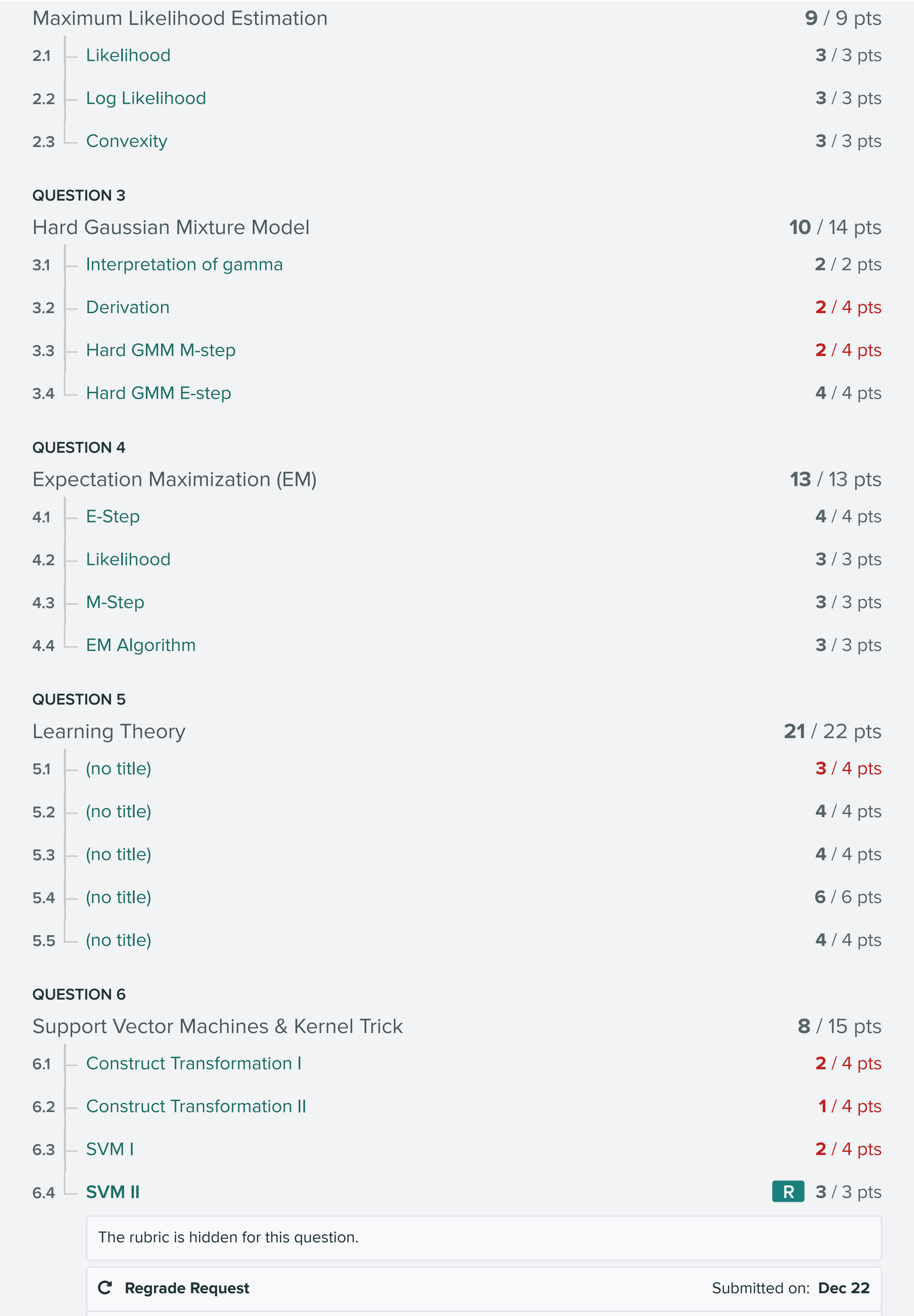
TOTAL POINTS

82 / 100 pts

QUESTION 1

	Short Answer Questions and Multiple Choices	**15** / 18 pts
1.1	Spam Filter Experiment	**0** / 3 pts
1.2	Code Review	**3** / 3 pts
1.3	ID3	**2** / 2 pts
1.4	Multi-class	**2** / 2 pts
1.5	Regression	**3** / 3 pts
1.6	K-Means	**5** / 5 pts

QUESTION 2



	Maximum Likelihood Estimation	**9** / 9 pts
2.1	Likelihood	**3** / 3 pts
2.2	Log Likelihood	**3** / 3 pts
2.3	Convexity	**3** / 3 pts

QUESTION 3

	Hard Gaussian Mixture Model	**10** / 14 pts
3.1	Interpretation of gamma	**2** / 2 pts
3.2	Derivation	**2** / 4 pts
3.3	Hard GMM M-step	**2** / 4 pts
3.4	Hard GMM E-step	**4** / 4 pts

QUESTION 4

	Expectation Maximization (EM)	**13** / 13 pts
4.1	E-Step	**4** / 4 pts
4.2	Likelihood	**3** / 3 pts
4.3	M-Step	**3** / 3 pts
4.4	EM Algorithm	**3** / 3 pts

QUESTION 5

	Learning Theory	**21** / 22 pts
5.1	(no title)	**3** / 4 pts
5.2	(no title)	**4** / 4 pts
5.3	(no title)	**4** / 4 pts
5.4	(no title)	**6** / 6 pts
5.5	(no title)	**4** / 4 pts

QUESTION 6

	Support Vector Machines & Kernel Trick	**8** / 15 pts
6.1	Construct Transformation I	**2** / 4 pts
6.2	Construct Transformation II	**1** / 4 pts
6.3	SVM I	**2** / 4 pts
6.4	**SVM II**	R **3** / 3 pts

The rubric is hidden for this question.

Regrade Request Submitted on: **Dec 22**

I’m not sure why this question is wrong because if I remove E then I can divide the points by a line at $x_1 = 0$ and then each of the points I listed would still work

Resolved

Reviewed on: **Dec 22**

QUESTION 7

Perceptron		**6** / 9 pts
7.1	(no title)	**0** / 0 pts
7.2	Perceptron Update Rule	**0** / 3 pts
7.3	Perceptron Mistake Bound	**6** / 6 pts

QUESTION 8

Additional Comment **0** / 0 pts