CSCI 567 Summer 2018 Real Final Exam
DO NOT OPEN EXAM UNTIL INSTRUCTED TO DO SO

Name	
ID #	

Read the following instructions carefully; you will be expected to conform to them during the exam.

- Multiple choice questions must be answered by filling in the appropriate segments on the Scantron form provided with a #2 pencil. These questions might not be weighted equally and partial credit may be available on some of them.

- Free response questions should be answered **concisely**. Excessive writings will incur mark reduction. Derivations should be short — no need to prove/derive extensively. No derivations will require more than 10 lines.

- Write **legibly** so the grader can read your answers without misunderstandings. Avoid cursive writings.

- You must answer each free response question on the page provided. We will not provide blank new copies of pages.

- The duration of the quiz is **90 minutes**. Please budget your time on each question accordingly.

- You **may not** leave your seat during the exam **for any reason**. If you leave your seat, you may be required to submit your exam at that moment.

- The quiz has a total of **7 physical pages** (including this cover), **13** multiple choice questions and **five(5) free response questions**. Each question may have sub-questions. Once you are permitted to open your exam (and not before), you should count the pages to ensure that you have the right number.

- This is a **closed-book** exam. Consulting classmates, the Internet, or other resources is NOT permitted. You may not use any electronics *for any reason* or say anything to a classmate *for any reason*.

- There are some spaces in the packet that can be used for scratch work. No additional scratch paper will be provided.

- When the proctor tells you the test is over, you must *immediately* cease writing and close the packet. Continuing to write after the exam is over will be reported for academic discipline, including at minimum an F **in the class**.

Question	Points Possible	Points Earned
M/C	14	
Q1: Kernel	2	
Q2: Neural Networks	4	
Q3: Mixture Models	2	
Q4: Hidden Markov Models	3	
Q5: k-Means	5	

This is the inside cover of the exam. You may use this as scratch paper if needed. Your packet must remain attached throughout the exam – you may not remove this page for any reason.

Gradient Descent Methods

1. Consider running the Perceptron algorithm (**not** batch gradient descent) on some set of training data S. Consider also some set S' that is the same elements of S, but in a different order.

   True or False: The algorithm makes the same mistakes on S as it would on S'.

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

2. Consider running **batch gradient descent** on some set of training data S. Consider also some set S' that is the same elements of S, but in a different order.

   True or False: Given the same starting value and threshold value(s), batch gradient descent will complete with the same $\boldsymbol{w}$ values.

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

Neural Networks

Two common activation functions in a neural network are the linear activation and the threshold function.

As a reminder, the linear activation function is $y = w_0 + \sum_i w_i x_i$. The threshold function is 1 if $w_0 + \sum_i w_i x_i \geq 0$ and is 0 otherwise.

3. Suppose we are building a neural network and can only use those two types of activation functions. True or false: this neural network can learn a polynomial function of degree one.

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

4. Suppose we are building a neural network and can only use those two types of activation functions. True or false: this neural network can learn the hinge loss function. As a reminder, the hinge loss function is $h(x) = \max(0, 1 - x)$

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

CNN

5. Overfitting is a major problem for neural networks. There are many techniques for mitigating overfitting. One of the following techniques will serve this purpose; select that answer choice.

   (a) Retraining on the same data many times.

   (b) Using less data.

   (c) Dropping random neurons for each training batch.

   (d) Training until you get the best accuracy on your training data.

You given a 5x5x1 black and white image shown below. You will use this image for questions 6 and 7. For both questions, assume that padding is allowed. That is, the image may be padded if needed.

0	0	0	0	0
0	1	0	1	0
0	0	0	0	0
0	1	0	0	0
0	0	1	1	0

6. Given a 3x3x1 filter, what will be the shape of the resulting image after convolving this filter with a stride of 2 and the provided 5x5x1 image.

   (a) 3x3x1

   (b) 2x2x1

   (c) 2x3x1

   (d) 3x2x1

7. Given this 3x3x1 filter what is the output of convolving it with a stride of 2 and the given 5x5x1 image and then passing that through a 5x5x1 maxpool filter.

0	0	0
1	0	1
0	0	0

   (a) 0

   (b) 1

   (c) 2

   (d) 3

Various

8. Q-learning can only be used when the learner has prior knowledge of how its actions will affect the environment.

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

9. True or false: Density estimation (using say, the kernel density estimator) can be used to perform classification.

   Fill in bubble (A) to indicate true and bubble (B) to indicate false.

10. Which of the following datasets would **not** be appropriate to use HMM for learning?

    (a) Gene sequences (e.g., human genome)

    (b) Movie reviews (e.g., IMDB)

    (c) Stock prices for a particular company or set of companies

    (d) Daily rainfall data for Seattle (Seattle a city that experiences much rain).

SVM

Recall the primal form of linear SVMs:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_n \xi_n \\ \text{s.t.} \quad & y_n[\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n + b] \geq 1 - \xi_n, \quad \forall \;\; n \\ & \xi_n \geq 0, \quad \forall \; n \end{aligned}$$

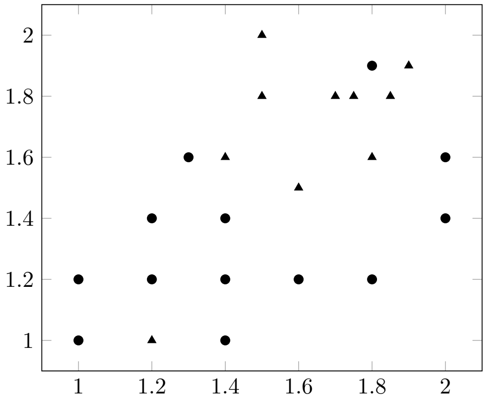
11. True or false: Support vector machines, like logistic regression models, give a probability distribution over the possible labels given an input example.

    Fill in bubble (A) to indicate true and bubble (B) to indicate false.

12. True or false: Suppose I train a SVM classifier, either with the points directly or with a linear kernel. I then train a separate SVM classifier using a higher order polynomial kernel. I expect the support vectors to remain the same.

    Fill in bubble (A) to indicate true and bubble (B) to indicate false.

13. (2 points) Consider the following dataset. It consists of two classes.



Suppose I solve this SVM using a *quadratic kernel.* Note that this will involve using also the dual which is not written here (you do not need it to solve this problem) and I use a reasonable value for the primal C. I then go through the process to get the values of the primal solution from this. How many points have $\xi_i > 1$?

Note: You do not need to know the dual for SVM to solve this exam question. You do, however, need to know how duals work in general.

(a) 1

(b) 2

(c) 3

(d) 4

(e) 5

This page intentionally left blank. You may use it as scratch paper if you wish.

1 Kernels (2 points)

The Mercer Theorem (the rules that tell us how to form a valid kernel from other valid kernels) tells us that $k(\boldsymbol{x}_m, \boldsymbol{x}_n) = c \cdot k'(\boldsymbol{x}_m, \boldsymbol{x}_n)$ is a valid kernel if c is a constant and $k'(\cdot, \cdot)$ is a valid kernel. For example, we can see that $k(\boldsymbol{x}_m, \boldsymbol{x}_n) = 5(\boldsymbol{x}_m^{\mathrm{T}} \boldsymbol{x}_n)$ is a valid kernel because the identity is a valid kernel.

Show that $k(\boldsymbol{x}_m, \boldsymbol{x}_n) = 5(\boldsymbol{x}_{\boldsymbol{m}}^{\mathrm{T}} \boldsymbol{x}_n)$ is a valid kernel *without appealing to the Mercer Theorem*, but instead by giving an explicit mapping $\phi(\boldsymbol{x})$ such that $k(\boldsymbol{x}_m, \boldsymbol{x}_n) = \phi(\boldsymbol{x}_m)^{\mathrm{T}} \phi(\boldsymbol{x}_n)$

Despite a large amount of blank space, there is a very short answer to this. If you plan to use the rest of the page as scratch paper, please draw a line to separate your answer to this question from scratch work for others.

2 Neural Networks (4 points)

In this problem, we will compare two different cost (loss) functions, 1) the quadratic cost function and 2) the cross entropy cost function. Let's assume the following network: A 3 dimension vector is fed into the input nodes and weight vectors and a bias are connecting the input nodes with the output node. The final output has a sigmoid function as an activation function. As a reminder, the sigmoid function is $s(a) = \frac{1}{1+e^{-a}}$ There are N different training samples, $(\mathbf{x}^{(n)}, y^{(n)})$, $n = 1, \cdots, N$. $\mathbf{x}^{(n)} \in \mathbb{R}^3$, $y^{(n)} \in \mathbb{R}$.

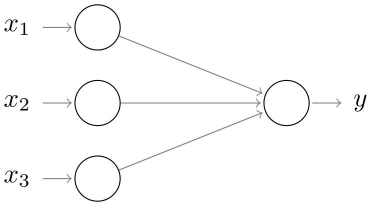


Figure 1: Neural Network

$$z = \sum_{i=1}^{3} w_i x_i^{(n)} + b, \quad \hat{y}^{(n)} = \sigma(z)$$

$$\text{Quadratic Cost} = C_1(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2}(y^{(n)} - \hat{y}^{(n)})^2$$

$$\text{Cross Entropy Cost} = C_2(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{n=1}^{N} \left[y^{(n)} \log \hat{y}^{(n)} + (1 - y^{(n)}) \log(1 - \hat{y}^{(n)}) \right]$$

- Find a derivative of C_1 function in terms of a weight w_j, i.e., $\frac{\partial C_1}{\partial w_j}$

- Find a derivative of C_2 function in terms of a weight w_j, i.e., $\frac{\partial C_2}{\partial w_j}$

- According to the gradient descent, w_j will be updated by $w_j^{new} = w_j^{old} - \gamma \frac{\partial C}{\partial w_j}$. If the learning rate γ is same for both C_1 and C_2, explain which one has larger $|\gamma \frac{\partial C}{\partial w_j}|$ and *briefly* explain why.

3 Mixture Model (2 points)

Let X be a random variable of a mixture of three univariate Bernoulli distributions,

$$X = \omega_1 Y + \omega_2 W + (1 - \omega_1 - \omega_2) Z$$

where $Y \sim$ Bernoulli(p_1) and $W \sim$ Bernoulli(p_2) and $Z \sim$ Bernoulli(p_3) and $\omega_1 + \omega_2 < 1$. If you prefer to refer to $1 - \omega_1 - \omega_2$ as ω_3, that is fine.

As a reminder, a Bernoulli(p) has value 1 with probability p and value 0 with probability $1 - p$.

If you are having trouble picturing this, we could phrase it this way: Imagine that I have given three (possibly fair, possibly unfair, possibly unfair in different ways) coins to three individuals. When I want a coin flip result, I ask one of them with probability ω_1, a different one with probability ω_2, and the third otherwise.

The only unknown parameter is the mixing parameters ω_i; we know the various p_i values.

Now we observe a single sample x.

There are two interpretations of this question; one as a mixture model and one as the sum of random variables. While I think this problem is easier as a mixture model, you should answer as whichever you prefer. I encourage you to indicate which one you are using if it isn't clear by your answer.

Write out the likelihood function of event $X = x$

4 Hidden Markov Models (HMMs, 3 points)

Recall the algorithm for HMMs:

```
$\delta_0(\text{Start}) = 1$
$\forall s \neq \text{Start}\ \delta_0(s) = 0$
for $i = 1 \ldots n$ do
    $\delta_i(s) = \max_{s_{i-1}}\{P(s_i|s_{i-1})\delta_{i-1}(s_{i-1})P(w_i|si)\}$
    $\psi_i(s) = \arg\max s_{i-1}\{P(s_i|s_{i-1})\delta_{i-1}(s_{i-1})P(w_i|s_i)\}$
end for
Fill in the end state as well.
```

After I grade this final, I am going to write a computer program. I will use at least one of two computer programming languages to do so: either I will use Erlang or JavaScript. Right now, you can think of this as that I am in the "start state." From that state, the probability of transitioning to Erlang is 0.75 and probability of transitioning to JavaScript is 0.25. Both languages can interact with others, so I might switch between them to best achieve my goals. The program will take me three days to complete, and I will only use one language any given day. For example, I might use Erlang tomorrow and Thursday, but JavaScript on Friday.

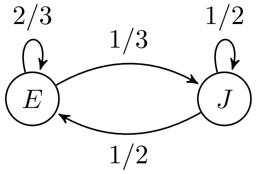
You cannot observe my computer to see what programming language I am using at any given time. You can, however, observe my mood to know if I am happy or angry. You know that I hate programming in JavaScript and I love Erlang, so you can guess what language I am using from this.

More specifically, you know the probability I am happy or angry given what language I am using and the likelihood I will switch languages from one.

E	$p(E\|X = \text{Erlang})$
happy	4/5
angry	1/5

E	$p(E\|X = \text{JavaScript})$
happy	1/4
angry	3/4

In the diagram below, E denotes Erlang and J denotes JavaScript



You notice that I am happy for days one and two and angry for day three.

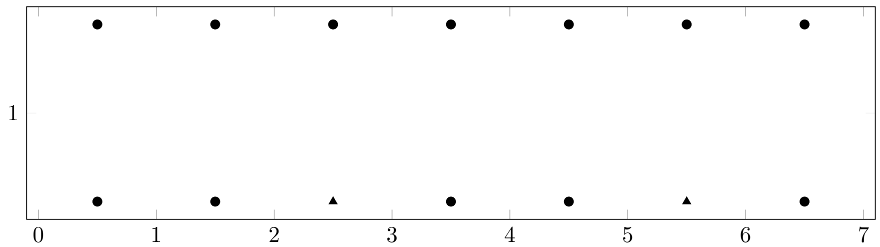
Write your answer in this space, one state per box. A blank page follows on which you may do your work. In order to get credit for this question, you must show your work.

Start				End

This page intentionally left blank. You may wish to use it for showing your work for HMMs.

5 K-Means

Consider the following dataset. All points are unlabeled and part of the same set. The triangles are used to distinguish two points later. *Please do not draw on this diagram until you have read the problems below.*



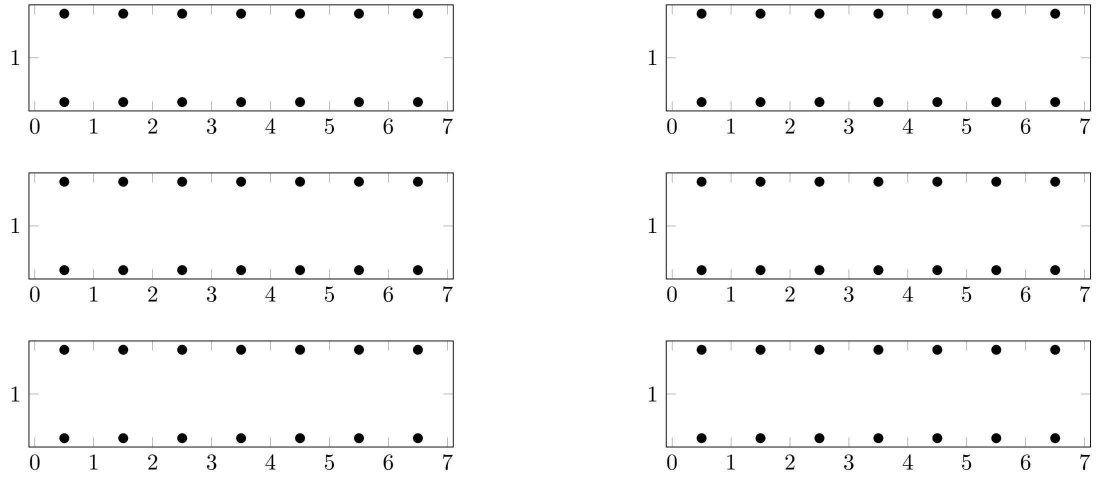
Suppose we want to cluster the data using Lloyd's k-means algorithm, which has the following pseudocode:

```
Select $k$ points as the initial values of $\boldsymbol{\mu}$
while not convered do
    Assign each point $\boldsymbol{x}_n$ to its nearest mean $\boldsymbol{\mu}_j$
    Re-assign each mean to be the center of mass of its assigned points.
end while
```

5.1 Clustering Data (2 points)

Use the points indicated by triangles as the initial values of $\boldsymbol{\mu}$ for $k = 2$. When the algorithm converges, there will be a clear dividing line between the two clusters. Draw the line clearly in the diagram above.

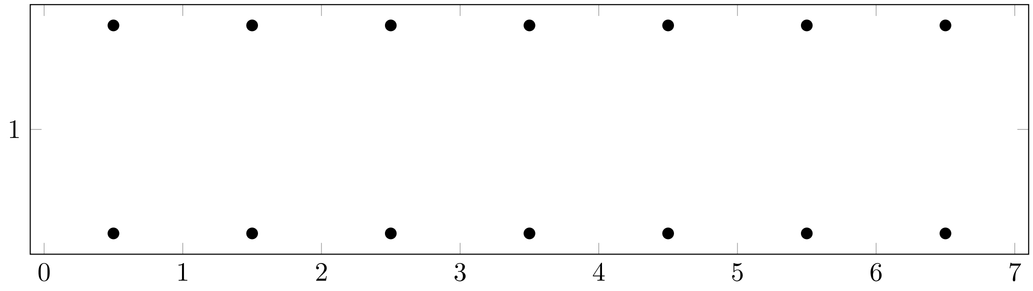
It is possible you may want copies of the diagram to use when working. They are drawn with circles for all points. **Nothing you write on these copies will be graded, but you may use them to solve the problem if you so desire. Write your answer to this part above.**



5.2 Initializing Lloyd's Algorithm (2 points)

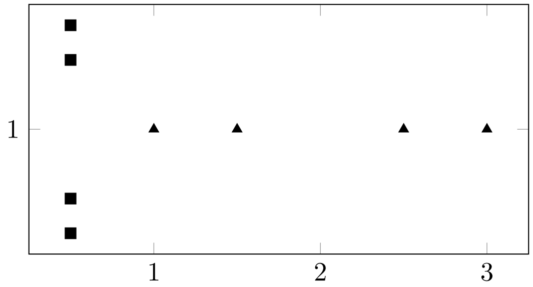
On the diagram below, draw two triangles with the following properties:

- Do not draw the same triangles as the problem statement on the previous page.
- Choose two points such that, if we run Lloyd's k-means algorithm with these two as the initial points, we will get the same clustering you found for the previous problem. Draw a triangle at these points.
- You are not obligated to select a data point as a triangle, although you may do so.



5.3 Forming Lloyd's k-means (1 point)

I might have run Lloyd's algorithm on the following data set. If I did so, I ran the algorithm until it fully converged (another iteration of the **while** loop would have the same partitioning / means). The points assigned to $\boldsymbol{\mu}_1$ are drawn as triangles and the points assigned to $\boldsymbol{\mu}_2$ are drawn as squares. I have drawn the result below:



Could this actually be the result of a run of that algorithm? **Clearly** indicate if you think the answer is "yes" or "no" and explain your answer briefly.