CS 691 Final Exam Name ___________________________

## Short Problems (20 points)

1. Let's say we have two polynomial feature maps: $\phi_1(x) = \{x, x^2\}$, and $\phi_2(x) = \{2x, 2x^2\}$. In general, is the margin we would attain using $\phi_2(x)$ greater, equal, or smaller, in comparison to the margin resulting from $\phi_1(x)$? **Solution:** Greater

2. Give one similarity and one difference between feature selection and PCA. **Solution:** Similarity=reduce the dimension of data, difference=feature selection finds a subset of features, while PCA produces a smaller, new set of features.

3. True/False: We would expect the support vectors to remain the same in general as we move from a linear kernel to higher order polynomial kernels. Explain. **Solution:** False

4. Which of the following statements are true? No need to explain.

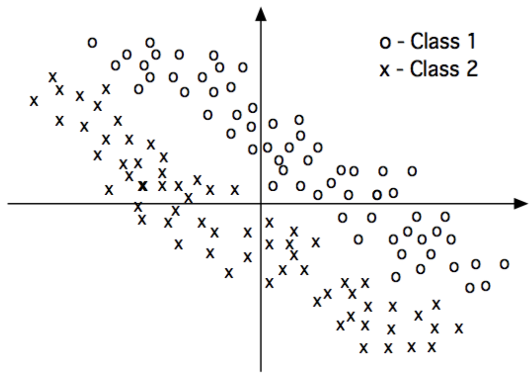
   (a) Training a k-nearest neighbors classifier takes more computational time than applying it. **Solution:** False

   (b) The more training examples, the more accurate the prediction of a k-nearest neighbors classifier. **Solution:** True

   (c) k-nearest neighbors cannot be used for regression (to predict a real-value). **Solution:** False

   (d) A k-nearest neighbors is sensitive to outliers. **Solution:** True

5. Explain how you might use the result of PCA on the below data to perform classification.




Solution on 491 Exam

6. Consider two types of Neural Network activations:

   - linear: $h = w \cdot x + b$
   - hard threshold: $h = 1$ if $w \cdot x + b \geq 0$, and $h = 0$ otherwise.

   Which of the following functions can be exactly represented by a neural network with one hidden layer which uses linear and/or hard threshold activation functions? For each case, justify your answer.

   (a) Polynomials of degree one
   (b) Hinge loss: $h(x) = max(1 - x, 0)$
   (c) Polynomials of degree two
   (d) Piece-wise constant functions

Solution on 491 Exam

7. For each of the following situations, indicate whether the classifier will produce a linear decision boundary always, sometimes or never. Give a short explanation of your answer. In every case, assume that the input has two features $x_1$ and $x_2$, which have continuous values ranging from $-1$ to 1. For example, the value of $x_1$ is any real number between $-1$ and 1. And suppose the label, $y$ has two possible values, 1 or $-1$. In all cases, assume there is plenty of training data (at least 100 examples for each class).

   (a) We build a decision tree with two levels.

   (b) We build a decision tree with one level.

   (c) We train a Perceptron.

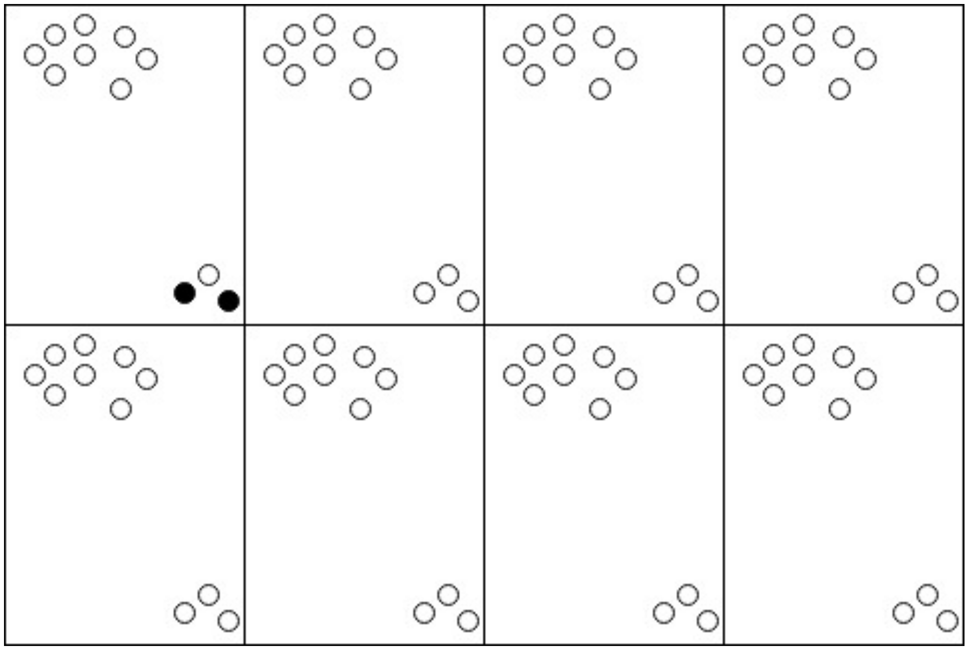
   (d) We use K-nearest neighbor, with K = 1.

   (e) We train an SVM using a quadratic kernel.

   (f) We use a neural network. The network has two input units, two hidden units and one output unit. We use ReLU in the hidden units (remember ReLU(x) = max(0,x)) and a sigmoid after the output unit. The input is classified as belonging to class 1 if the output is greater than $\frac{1}{2}$, and as class $-1$ if the output is less than $\frac{1}{2}$.

Solution on 491 Exam

8. Run K-Means manually for the following dataset. Circles are data points and the filled-in circles in the first panel are the initial cluster centers. Draw the cluster centers and the decision boundaries that define each cluster. Use a different panel for each iteration. Use as many panels as you need until convergence.



Solution on 491 Exam

9. Suppose we have a hyperplane given by the equation $w \cdot x + b = 0$.

   (a) We want to write a loss function for this. We are given a feature vector, $x$, and a label, $y$, with $y = 1$ or $y = -1$. If $x$ is on the positive side of the hyperplane and $y = 1$, the loss should be 0. If $x$ is on the negative side of the hyperplane and $y = -1$, the loss is 0. Otherwise, the loss is the distance from $x$ to the hyperplane. Write an expression for that loss.

   (b) Suppose we want to find the hyperplane that minimizes the loss in the previous question using gradient descent. Write an update equation that shows how we should update $b$ based on an example $(x, y)$.

a) wx+b is the distance from the hyperplane. So if y(wx+b)>0, loss is 0. Else, loss is -1(wx+b)

b) dL/db = -1
So, b = b+n
n=eta=learning rate

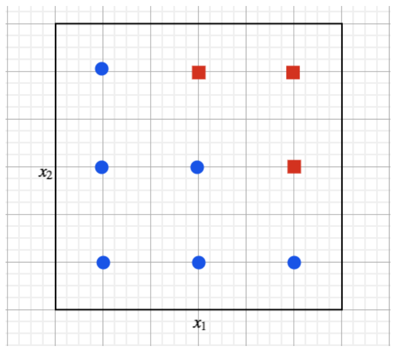
10. Recall that Adaboost learns a classifier $H$ using a weighted sum of weak learners $h_t$ as follows:

Solution on 491 Exam

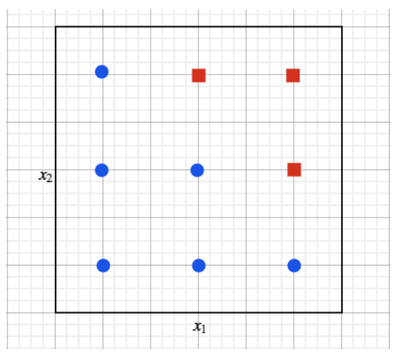
$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

In this question, we will use decision trees as our weak learners, which classify a point as $\{1, -1\}$ based on a sequence of threshold splits on its features ($x_1$ and $x_2$). Red squares are negative points and blue circles are positive points. **In the questions below, be sure to mark which regions are marked positive/negative, and assume that ties are broken arbitrarily.**

(a) Now assume that our weak learners are decision trees of maximum depth 2, which minimize the weighted training error. Using the dataset below, draw the decision boundary learned by $h_1$.




(b) On the dataset below, circle the point(s) with the highest weights on the second iteration, and draw the decision boundary learned by $h_2$.




(c) On the dataset below, draw the decision boundary $H = sign(\alpha_1 h_1 + \alpha_2 h_2)$. (Hint: you do not need to explicitly compute the $\alpha$'s).