I have neither given nor received unauthorized assistance on this exam.

 __________________________________________________________
                  Name Signed                      Date

 __________________________________________________________
                  Name Printed
1. Below is a small \( p = 2 \) classification training set (for 2 classes) displayed in graphical and tabular forms (circles are class \(-1\) and squares are class \(1\)).

\[
\begin{array}{cccc}
  y & x_1 & x_2 \\
-1 & 1 & 2 \\
-1 & 1 & 10 \\
-1 & 3 & 5 \\
-1 & 7 & 0 \\
-1 & 7 & 10 \\
-1 & 8 & 6 \\
-1 & 11 & 4 \\
1 & 7 & 15 \\
1 & 7 & 17 \\
1 & 10 & 19 \\
1 & 11 & 14 \\
1 & 13 & 12 \\
1 & 13 & 20 \\
1 & 14 & 16 \\
1 & 17 & 12 \\
1 & 17 & 19 \\
\end{array}
\]

\(a)\) Using the geometry above (and not trying to solve an optimization problem analytically) find the support vector machine for this classification problem. You may find it helpful to know that if \( u, v, \) and \( w \) points in \( \mathbb{R}^2 \) and \( u \neq v \), then the distance from the point \( w \) to the line through \( u \) and \( v \) is

\[
\frac{|w_1 (v_2 - u_2) - w_2 (v_1 - u_1) + v_1 u_2 - u_1 v_2|}{\sqrt{(v_1 - u_1)^2 + (v_2 - u_2)^2}}
\]

\(b)\) List the set of support vectors and the "margin" \( (M) \) for your SVM.
2. Below is a different $p = 2$ classification training set for 2 classes (specified as in problem 1).

\[ \begin{array}{ccc}
  y & x_1 & x_2 \\
  -1 & 1 & 2 \\
  -1 & 1 & 12 \\
  -1 & 3 & 5 \\
  -1 & 7 & 0 \\
  -1 & 7 & 13 \\
  -1 & 8 & 6 \\
  -1 & 15 & 4 \\
  1 & 7 & 15 \\
  1 & 7 & 17 \\
  1 & 11 & 15 \\
  1 & 12 & 16 \\
  1 & 13 & 7 \\
  1 & 13 & 20 \\
  1 & 15 & 11 \\
  1 & 17 & 16 \\
  1 & 17 & 19 \\
\end{array} \]

\begin{itemize}
\item a) Using empirical misclassification rate as your splitting criterion and standard forward selection, find a reasonably simple binary tree classifier that has empirical error rate 0. Carefully describe it below, using as many nodes as you need.
\end{itemize}

At the root node: split on $x_1$ / $x_2$ (circle the correct one of these) at the value ____________

Classify to Class $-1$ if ________________ (creating Node #1)

Classify to Class 1 otherwise (creating Node #2)

At Node # ______________: split on $x_1$ / $x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #3)

Classify to Class 1 otherwise (creating Node #4)

At Node # ______________: split on $x_1$ / $x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #5)

Classify to Class 1 otherwise (creating Node #6)
At Node # ______________: split on $x_1 / x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #7)

Classify to Class 1 otherwise (creating Node #8)

At Node # ______________: split on $x_1 / x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #7)

Classify to Class 1 otherwise (creating Node #8)

At Node # ______________: split on $x_1 / x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #9)

Classify to Class 1 otherwise (creating Node #10)

At Node # ______________: split on $x_1 / x_2$ at the value ____________

Classify to Class $-1$ if ________________ (creating Node #11)

Classify to Class 1 otherwise (creating Node #12)

(Add more of these on another page if you need them.)

b) Draw in the final set of rectangles corresponding to your binary tree on the graph of the previous page.
c) For every sub-tree, $T$, of your full binary tree above, list in the table below the size (number of final nodes) of the sub-tree, $|T|$, and the empirical error rate of its associated classifier.

| Full tree pruned at Nodes # | $|T|$ (pruned tree size) | $\overline{\text{err}}$ |
|-----------------------------|--------------------------|--------------------------|
| None (full tree)            |                          |                          |

(If you need more room in the table, add rows on another page.)

d) Using the values in your table from c), find for every $\alpha \geq 0$ a sub-tree of your full tree minimizing

$$C_\alpha = |T| + \alpha \cdot (\overline{\text{err}})$$
3. Here again is the $p = 2$ classification training set specified in problem 2.

Using "stumps" (binary trees with only 2 nodes) as your base classifiers, find the $M = 2$ term AdaBoost classifier for this problem. (Show your work!)
4. The machine learning/data mining folklore is full of statements like "combining uncorrelated classifiers through majority voting produces a committee classifier better than every individual in the committee." Consider the scenario outlined in the table below as regards classifiers $f_1$, $f_2$, and $f_3$ and a target (class variable) taking values in $\{0,1\}$.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>$f_3$</th>
<th>Committee Classification</th>
<th>$y$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.008</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.008</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.008</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.08</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.08</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.08</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>$(.1)^3$</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>.9 $(.1)^2 - .008$</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.9 $(.1)^2 - .008$</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.9 $(.1)^2 - .008$</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.9 $(.1) - .08$</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>.9 $(.1) - .08$</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.9 $(.1) - .08$</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.9 $(.1)^3$</td>
<td></td>
</tr>
</tbody>
</table>

Fill in the (majority vote) Committee Classification column and say carefully (show appropriate calculations to support your statement(s)) what the example shows about the folklore.
5. Suppose that in a $p = 2$ linear discriminant analysis problem, 4 transformed means

$$
\mu_k^* = \Sigma^{-\frac{1}{2}}(\mu_k - \bar{\mu})
$$

are $\mu_1^* = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mu_2^* = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \mu_3^* = \begin{pmatrix} 3.5 \\ 1.5 \end{pmatrix},$ and $\mu_4^* = \begin{pmatrix} .5 \\ 2.5 \end{pmatrix}$. These have sample covariance matrix

$$
\begin{pmatrix}
3.125 & 1.625 \\
1.625 & 3.125
\end{pmatrix} = \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{pmatrix} \text{diag}(4.75, 1.5) \begin{pmatrix}
\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}
\end{pmatrix}
$$

\[\text{a)}\] Suppose that one wants to do reduced rank ($\text{rank} = 1$) linear discrimination based on a single real variable

$$
w = (u_1, u_2) \Sigma^{-\frac{1}{2}} (x - \bar{\mu})
$$

Identify an appropriate vector $(u_1, u_2)$ and with your choice of vector, give the function $f(w)$ mapping $\mathbb{R} \to \{1, 2, 3, 4\}$ that defines this 4-class classifier.

\[\text{b)}\] What set of "prototypes" $(w)$ yields the classifier in a)?
6. In classification problems, searching for a small number of effective basis functions $h_1, h_2, \ldots, h_k$ to use in transforming a complicated situation involving input $x \in \mathbb{R}^p$, into a simpler situation involving "feature vector" $(h_1(x), h_2(x), \ldots, h_k(x))^\prime \in \mathbb{R}^k$ is a major concern.

a) The $p = 1$ toy classification problem with $K = 2$ classes and training data below is not simple.

<table>
<thead>
<tr>
<th>$y$</th>
<th>1</th>
<th>1</th>
<th>-1</th>
<th>-1</th>
<th>-1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

But it's easy to find a single function $h(x)$ that makes the problem into a linearly separable problem with $k = p = 1$. Name one such function.

b) In what specific way(s) does the use of kernels and SVM methodology typically lead to identification of a *small* number of important features (basis functions) that are effective in 2-class classification problems?

c) If you didn't have available SVM software but DID have available Lasso or LAR regression software, say how you might use the kernel idea and the Lasso/LAR software in the search for a few effective features in a 2-class classification problem.
7. Consider a Bayesian model averaging problem where \( x \) takes values in \( \{0,1\} \) and \( y \) takes values in \( \{0,1\} \). The quantity of interest is \( P[y=1|x=1]/P[y=0|x=1] \) and there are \( M = 2 \) models under consideration. We'll suppose that joint probabilities for \((x,y)\) are as given in the tables below for the two models for some \( p \in (0,1) \) and \( r \in (0,1) \)

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y ) ( x )</td>
<td>(0) (1) (0) (1)</td>
<td>(0) (1) (0) (1)</td>
</tr>
<tr>
<td>0</td>
<td>((1-p)/2) (p/2)</td>
<td>.25 .25</td>
</tr>
<tr>
<td>1</td>
<td>.25 .25</td>
<td>((1-r)/2) (r/2)</td>
</tr>
</tbody>
</table>

so that under Model 1, the quantity of interest is \( .5/p \) and under Model 2, it is \( r/.5 \). Suppose that under both models, training data \((x_i,y_i)\) \(i=1,2,\ldots,N\) are iid. For priors, in Model 1 suppose that \( a \text{ prior } p \sim \text{Beta}(2,2) \) and in Model 2 suppose that \( a \text{ prior } r \sim \text{Beta}(2,2) \). Further, suppose that the prior probabilities of the two models are \( \pi(1) = \pi(2) = .5 \).

Find the posterior probabilities of the 2 models, \( \pi(1|T) \) and \( \pi(2|T) \) and the Bayes model average squared error loss predictor of \( P[y=1|x=1]/P[y=0|x=1] \). (You may think of the training data as summarized in the 4 counts \( N_{(x,y)} = \) the number of training vectors with value \((x,y)\).)