This exam has 5 questions (for 95 points). You are to write legibly, show all work. I cannot give partial credit if no work is shown. References are made to definitions and methods as used in class and/or book. If you use additional assumptions, you have to state your definitions and assumptions clearly, otherwise they may be misinterpreted. This exam is closed book, closed notes, no calculators, and to be taken without help or assistance. Suspected academic dishonesty will be reported.

You may leave the answers in terms of factorials, permutations, combinations, binomial coefficients, logarithms, exponentials, summations, products, ..., whichever is appropriate. GOOD LUCK

1. (20 points total) SVM Find the support vector where $C = 1$ for the points (1,3) and (2,1) and sketch the separating hyperplane along with the maximal margin
   The equations you need are below

$$\begin{aligned}
\frac{\partial}{\partial\lambda_1}L(\lambda,\gamma) &= 1+\lambda_2 < \bar{x}_1, \bar{x}_2 > -\lambda_1 < \bar{x}_1, \bar{x}_1 > -\gamma = 0 \\
\frac{\partial}{\partial\lambda_2}L(\lambda,\gamma) &= 1+\lambda_1 < \bar{x}_1, \bar{x}_2 > -\lambda_2 < \bar{x}_2, \bar{x}_2 > +\gamma = 0 \\
\frac{\partial}{\partial\gamma}L(\lambda,\gamma) &= -\lambda_1+\lambda_2 = 0 \\
\bar{w} &= \lambda_1\bar{x}_1 - \lambda_2\bar{x}_2 \\
b &= 1- < \bar{w}, \bar{x}_1 > \\
&= -1- < \bar{w}, \bar{x}_2 >
\end{aligned}$$

$$\begin{aligned}
1+5\lambda-10\lambda-\gamma &= 0 \\
1+5\lambda-5\lambda+\gamma &= 0
\end{aligned} \tag{1}$$

$$\lambda = 2/5,\ \gamma = b = -1$$

$$\bar{w} = \lambda\bar{x}_1 - \lambda\bar{x}_2 = (-2/5, 4/5)$$

$$1-\left(-\frac{2}{5}\cdot 1+\frac{4}{5}\cdot 3\right) = -1 = b = \gamma$$

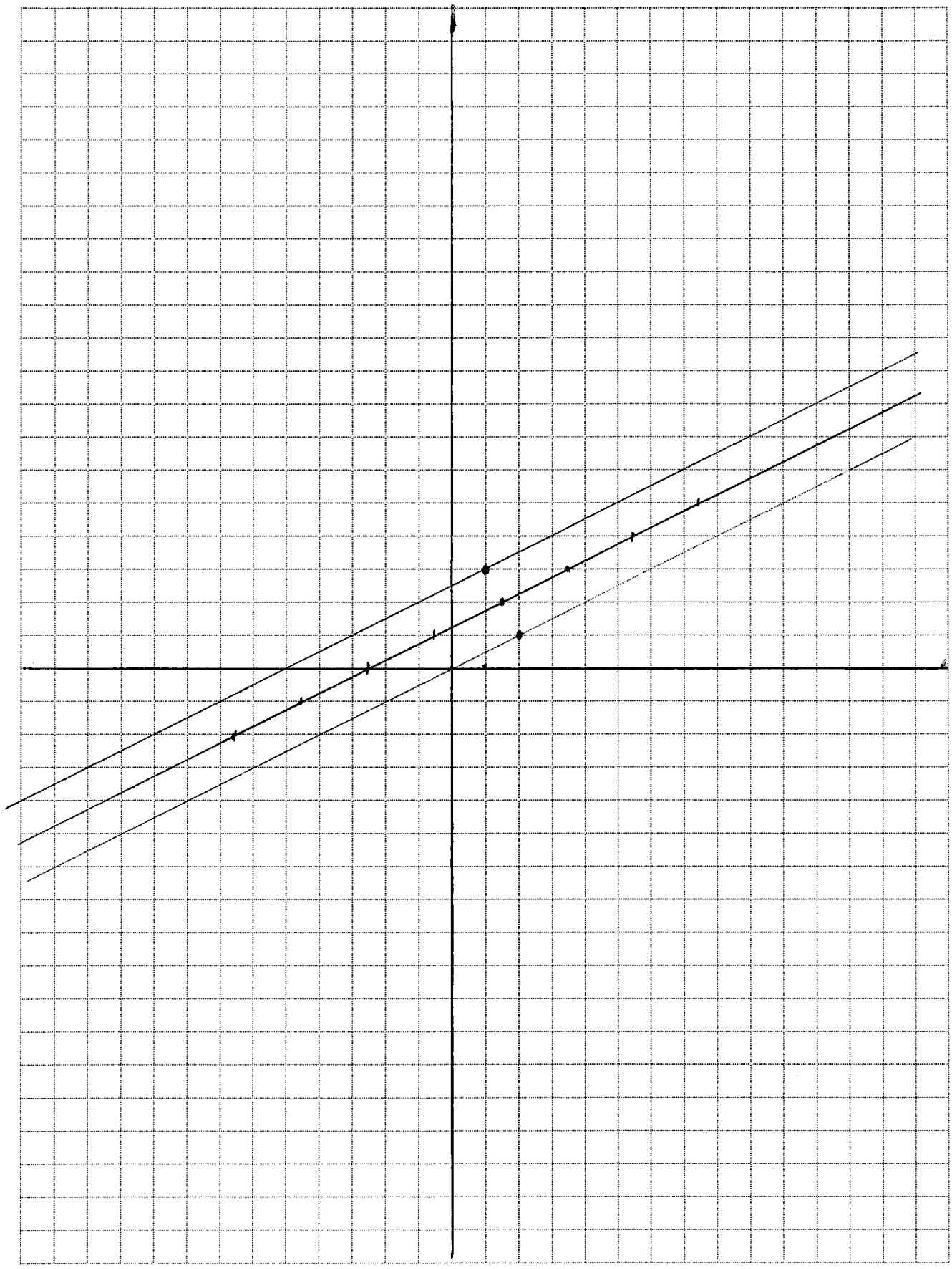
$$-1-\left(-\frac{2}{5}\cdot 1+\frac{4}{5}\cdot 1\right) = \mathbf{-1} = b = \gamma$$

$$y = \frac{1}{2}x - 5/4$$

If you were to plot the line, you would notice that the line does not pass between the two points. If the value of one equation of λ is equal to 0 we may need to switch the points. Now we get $\gamma = b = 1$, $\lambda = 2/5$, and $\bar{w} = (2/5, -4/5)$

$$y = \frac{1}{2}x + 5/4$$

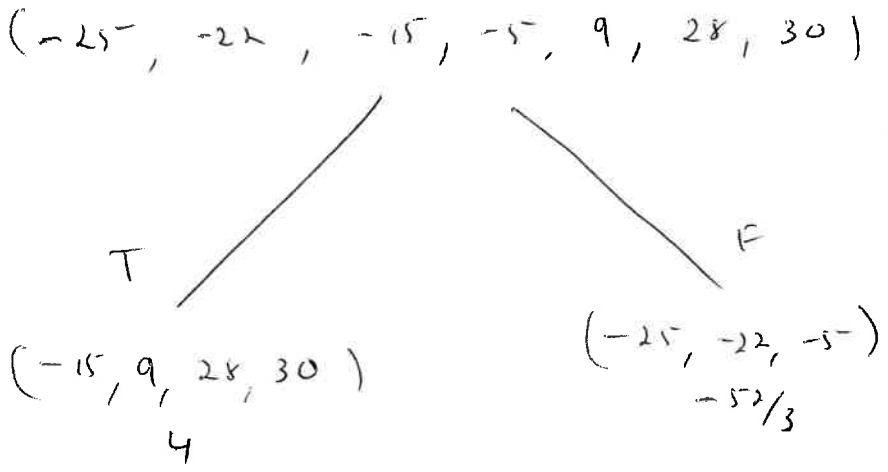
please turn the page



2. (20 points total) Decision Trees
Suppose we want to predict a person's age based on whether they prefer product A, or product B. We will do this by creating a decision tree and minimizing the squared error. The data collected is given in the table below:

ID	Age	A	B
1	22	F	T
2	25	F	T
3	32	T	T
4	42	F	T
5	56	T	F
6	75	T	T
7	77	T	F

(a) (5 points) Draw a decision tree for determining the person's age for those who like product A.

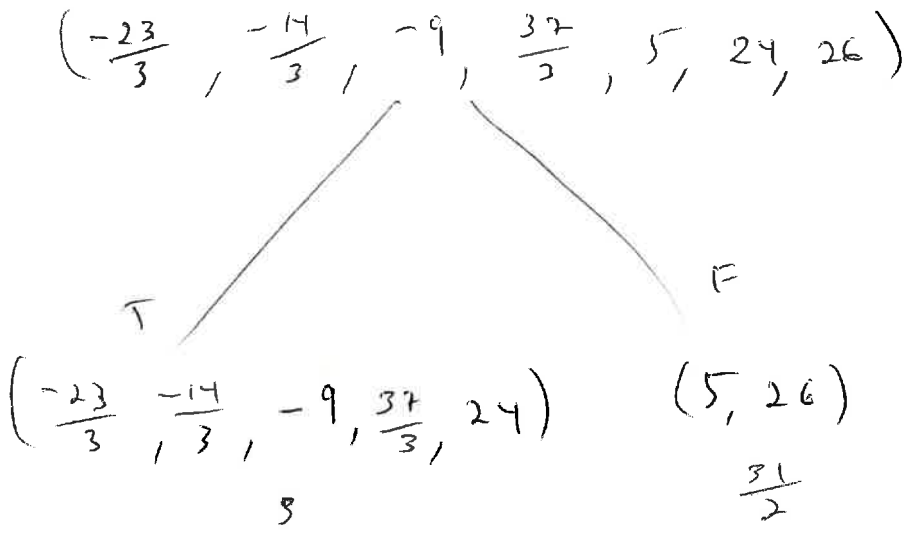


(b) (5 points) Calculate the residuals (Age - Prediction) for your tree and fill in the table below including the total SSE.

ID	Age	A	B	F_0	Ps_0	h_0	F_1
1	22	F	T	47	-25	-52/3	89/3
2	25	F	T	47	-22	-52/3	89/3
3	32	T	T	47	-15	4	51
4	42	F	T	47	-5	-52/3	89/3
5	56	T	F	47	9	4	51
6	75	T	T	47	28	4	51
7	77	T	F	47	30	4	51

please turn the page

(c) (5 points) Draw a second decision tree for determining the person's age for those who like product B based on the residuals of the tree you just constructed.



(d) (5 points) Calculate the residuals (Age - Prediction) for your second tree and fill in the table below including the total SSE.

ID	Age	A	B	F_0	Ps_0	h_0	F_1	Ps_1	h_1	F_2
1	22	F	T	47	-25	-52/3	89/3	-23/3	3	98/3 = 32.6
2	25	F	T	47	-22	-52/3	89/3	-14/3	3	32.6
3	32	T	T	47	-15	4	51	-9	3	54
4	42	F	T	47	-5	-52/3	89/3	37/3	3	32.6
5	56	T	F	47	9	4	51	5	31/2	66.5
6	75	T	T	47	28	4	51	24	3	54
7	77	T	F	47	30	4	51	26	31/2	66.5

Combined SSE = 1403.98

please turn the page

3. (15 points total) Consider two smoothing splines, $\hat{g}_1$ and $\hat{g}_2$ defined by

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(3)}(x) \right]^2 dx \right),$$

$$\hat{g}_2 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(4)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the mth derivative of g.

(a) (5 points) As $\lambda \longrightarrow \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training RSS?
The smoothing spline $\hat{g}_2$ will probably have the smaller training RSS because it will be a higher order polynomial due to the order of the penalty term (it will be more flexible).

(b) (5 points) As $\lambda \longrightarrow \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller test RSS?
It is hard to say. As mentioned above we expect $\hat{g}_2$ to be more flexible, so it may overfit the data, but $\hat{g}_1$ may underfit. It will probably be more likely $\hat{g}_1$ that have the smaller test RSS.

(c) (5 points) For $\lambda = 0$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training and test RSS?
If $\lambda = 0$, we have $\hat{g}_1 = \hat{g}_2$, so they will have the same training and test RSS.

please turn the page

4. (15 points total) We are given a set of $n = 8$ observations in $p = 2$ dimensions.

Obs.	X_1	X_2	Y	Class
1	0	0	-2	Blue
2	2	2	$-\frac{46}{25}$	Blue
3	-5	-2	$-\frac{29}{25}$	Blue
4	5	5	1	Red
5	-5	5	5	Red
6	7	7	$\frac{97}{25}$	Red
7	-7	-7	$\frac{97}{25}$	Red
8	5	-10	19	Red
9	-10	5	10	Red
10	-5	-5	1	Red

A Support Vector Machine classifier using the following kernel $< \overline{x}_1, \overline{x}_2 >^2$ was used to classify the data. The equation for this hyperplane is given as

$$y =< \overline{w}, \overline{x} >^2 + b$$

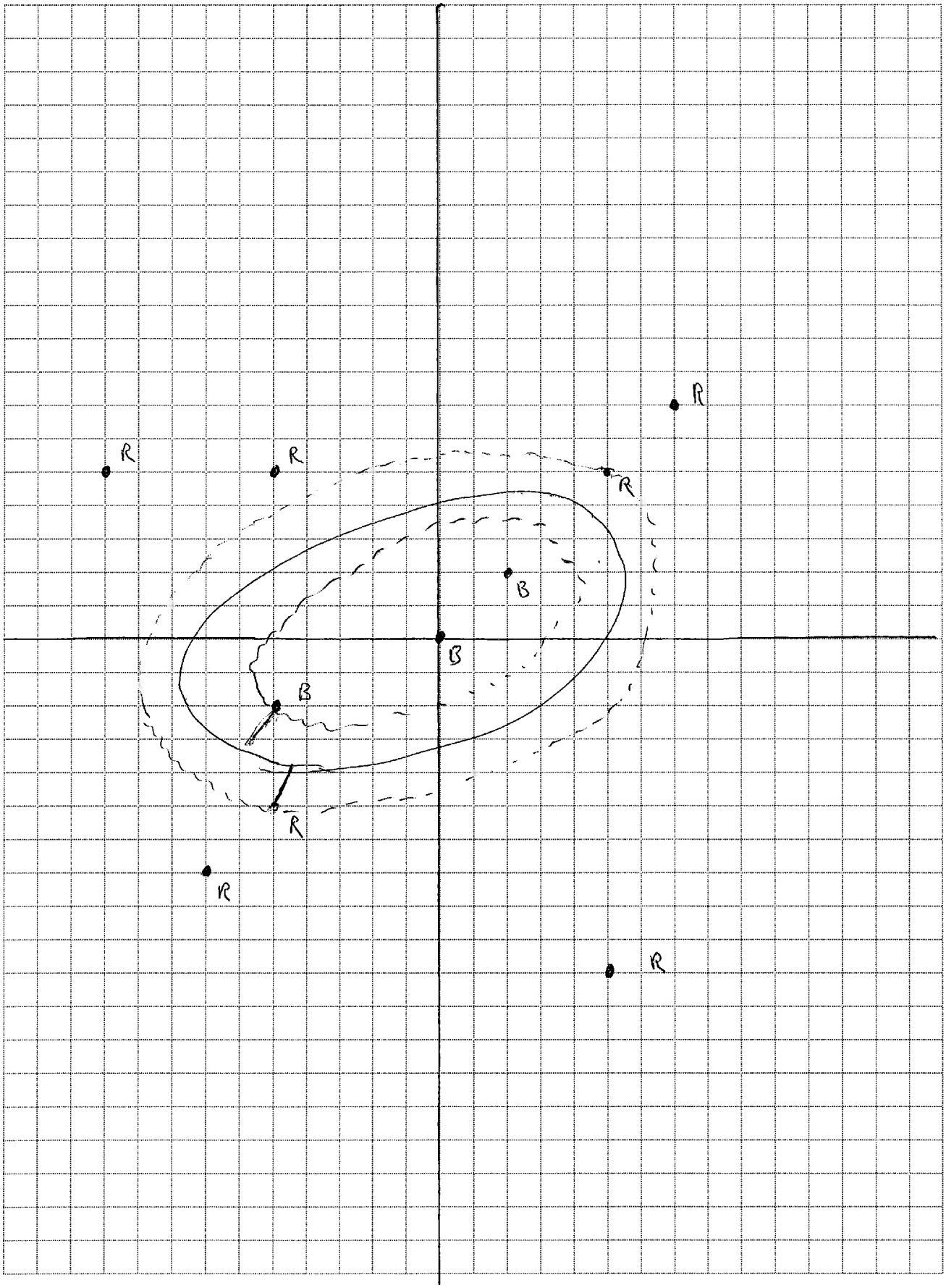
where $\overline{w} = (\frac{-1}{5}, \frac{2}{5})$ and $b = -2$. For $y > 0$ classify red and for $y < 0$ classify blue. The expression is:

$$\left(-\frac{1}{5} \cdot x_1\right)^2 + 2\left(-\frac{1}{5} \cdot x_1 \cdot \frac{2}{5} \cdot x_2\right) + \left(\frac{2}{5} \cdot x_2\right)^2 - 2$$

(a) (5 points) Sketch and label the observations.

(b) (5 points) Sketch a separating hyperplane (note: it may not be linear, it need not be exact)

(c) (5 points) Indicate the support vectors for the maximal margin classifier.


R
R
R
R
R
B
B
B
R
R
R

5. (20 points)Parameter Selection and Model Selection
Circle the correct answer and give justification in the space provided: (justification must be given and must advocate circled answer to earn any credit)

(a) (2 points) When selecting parameters, the best subset of size $p+1$ always has an equal or lower RSS in the training data than the best subset of size p: **T F** Explain:

False: For training, adding predictors always reduces the RSS

(b) (2 points) When selecting parameters using Lasso, the best subset of size $p+1$ always has an equal or lower RSS in the training data than the best subset of size p: **T F** Explain:

False: Lasso introduces a shrinkage penalty to prevent the size of the model (p) from getting too large.

(c) (2 points) Once a set of candidate models have been chosen, cross validation using a test set is the most accurate method for choosing the best model (best set of parameters.): **T F** Explain:

True: This is the most accurate method, but also requires the most computation and a large data set.

(d) (2 points) Lasso will choose the best subset for any value of p out of all the 2^p possible predictors: **T F** Explain:

False: Like backward selection, Lasso selects a model starting with all predictors. However, it usually performs much better, sometimes as good as best subset.

(e) (2 points) Ridge regression is frequently used because it's results are more interpretable than Lasso: **T F** Explain:

False: Since the predictors are never reduced to 0 in ridge regression, it is difficult to interpret. In Lasso it is clear which set of predictors are important for a model of size k

(f) (2 points) Which measure for choosing the proper model is the most parsimonious? **AIC**, **BIC**, $\boldsymbol{C_P}$, $\boldsymbol{R^2}$ Explain:

BIC is usually the most parsimonious, choosing the smallest model

(g) (2 points) Which measure for choosing the proper model results in a model closest to the *real* distribution **AIC**, **BIC**, $\boldsymbol{C_P}$, $\boldsymbol{R^2}$ Explain:

This was meant to say "which distribution assumes that a real model exists". That would be BIC. AIC assumes that all models are pseudo models and will tend to choose a larger model than BIC. Cp is similar to AIC.

(h) (6 points) Explain and compare the methods of Cross Validation, **AIC**, and R^2 for choosing a model. Strengths and weaknesses. Explain:

R^2 itself as was shown is not used as it will always choose the largest model. Adjusted R^2 is better since it adds a penalty. Cross validation is the most accurate, but can be computationally complex.
AIC has a benefit over cross validation by being able to narrow down the set of candidate models during training. etc...

please turn the page

- Boosting algorithm version 1.
  1. F_0 Start with the mean of the data set
  2. Root of the first tree is the residual

$$Ps_0 = x_i - F_0$$

  3. fit the tree to the training data.
  4. h_0 = mean of all residuals in each leaf .
  5. $F_1 = F_0 + h_0$
  6. $Ps_1 = Ps_0 - h_0$
  7. h_1 = mean of all values in each leaf.
  8. $F_2 = F_1 + h_1$

- Dot product squared

$$<\bar{x}, \bar{y}>^2 = <\bar{x}, \bar{y}><\bar{x}, \bar{y}> = (x_1y_1 + x_2y_2 + \ldots + x_ny_n), (x_1y_1 + x_2y_2 + \ldots + x_ny_n)$$

  For two dimensions this is equivalent to

$$<\bar{x}, \bar{y}>^2 = (x_1y_1)^2 + 2x_1y_1x_2y_2 + (x_2y_2)^2$$