



Indian Institute of Technology Kharagpur

QUESTION-CUM-ANSWERS SCRIPT

Stamp/Signature of the Invigilator

MID-SEMESTER EXAMINATION

SEMESTER (Autumn)

Roll Number										Section		Name	
Subject Number	C	S	4	0	0	0	3			Subject Name	DATA ANALYTICS		
Department/Centre/School											Additional Sheets		

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession; even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, the exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer-script and do not tear off any page. **Use last page(s) of the answer-script for rough work.** Report to the invigilator if the answer-script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the Examination Hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer-script to the invigilator. **In any case, you are not allowed to take away the answer-script with you.** After the completion of the examination, do not leave your seat until the invigilators collect all the answer-scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as 'unfair means'. Don't adopt unfair means and also don't indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be Filled by the Examiner

Question Number	1	2	3	4	5	6	7	8	9	10	Total
Marks Obtained											
Marks Obtained (in words)				Signature of the Examiner				Signature of the Scrutineer			

Computer Science & Engineering Department

Data Analytics CS40003

Autumn Semester, 2015

Mid Semester Examination

Date: 18/09/15

Time Limit: 2 Hours

Full Marks: 60

Name:

Roll No:

All answers must be in this question paper. You may use the back of the pages for rough work.

1. (a) Answer whether True or False

[3]

- i. _____ When a decision tree is grown to full depth, it is more likely to fit the noise in the data
- ii. _____ When the hypothesis space is more flexible, overfitting is more likely.
- iii. _____ When the feature space is larger, over fitting is more likely.

(b) Answer whether True or False AND explain in at most 2 sentences.

[6]

- i. _____ Using a model with less bias is always better than using a model with more bias.

- ii. _____ Variance of a model typically decreases as the number of features increases.

- iii. _____ As the amount of data increases, the true error of 1-Nearest Neighbour approaches 0, assuming noise-free data.

- (c) An econometrician fits a model with 500 parameters to a set of 800 house prices and the regression explains 99 % of the variation. Please comment on how successful the model may be to predict the price of houses next year accurately.

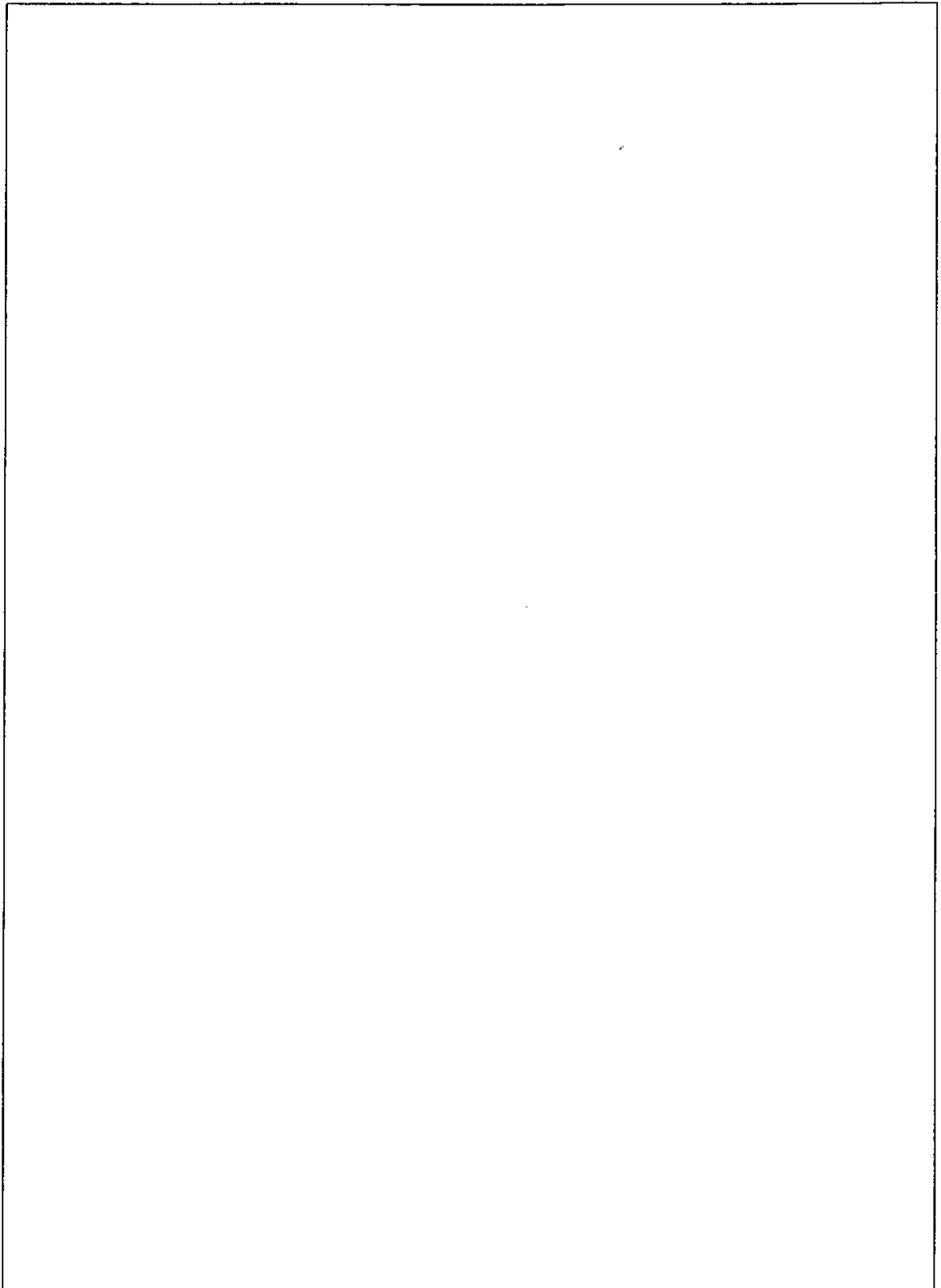
[3]

2. Use the following dataset to learn a decision tree for predicting if people will get hired at a great company (Y) or not (N), based on their Data Analytics grade (High or Low), their CGPA (High or Low) and on whether or not they did an internship.

[8]

DA Grade	CGPA	Internship	(output) Hired
L	H	Y	Y
L	L	N	N
L	L	Y	N
L	L	N	N
H	H	Y	Y
H	L	Y	Y
H	H	N	Y
H	L	N	Y

- (a) What is the entropy $H(\text{Hired} \mid \text{Internship} = \text{N})$?
- (b) What is the entropy $H(\text{Hired} \mid \text{GPA} = \text{H})$?
- (c) Draw the full decision tree that would be learned for this data (assuming no pruning). You do not need to show the calculation of information gain.



3. Consider a regression problem in which we want to predict variable y from a single feature x . We have $n \geq 3$ data points, $(x_i, y_i)_{i=1}^n$. Consider two possible models to be estimated by ordinary linear regression,

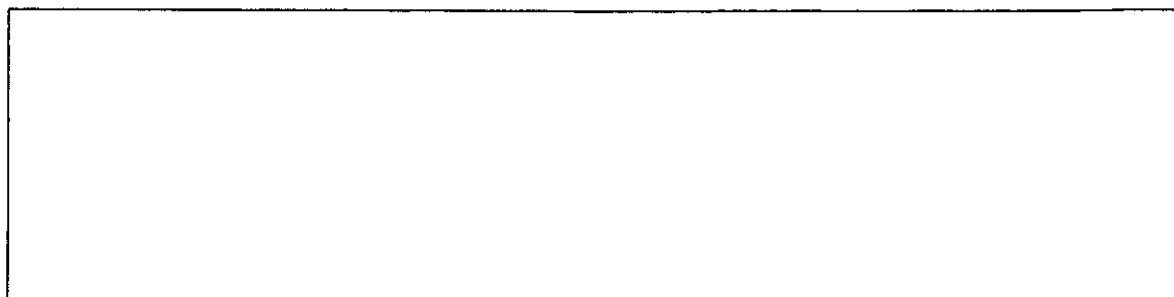
[7]

$$\text{Model1 : } y_i = w_0 + w_1 x_i + \epsilon_i \quad (1)$$

$$\text{Model2 : } y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i \quad (2)$$

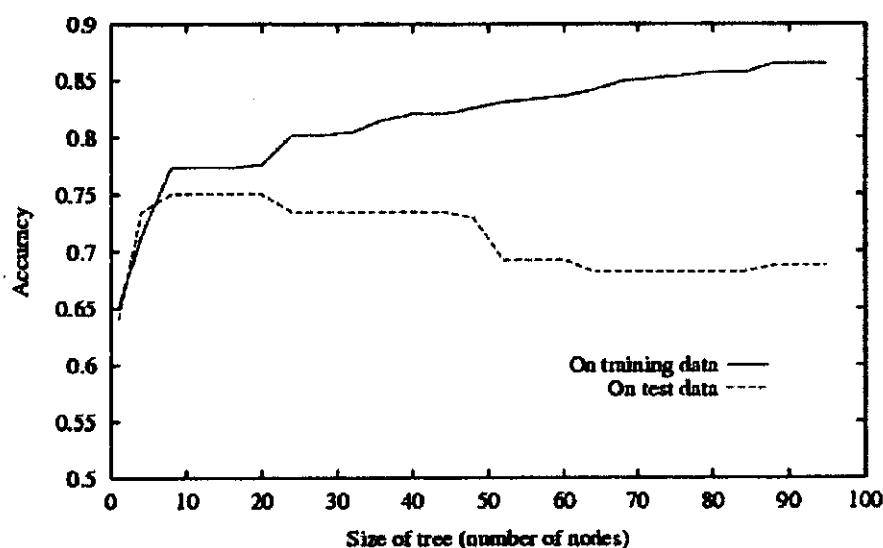
The error terms ϵ_i are independent and identically distributed from a normal distribution with zero mean.

- (a) Will one model fit the **training** data better than the other, will they fit equally well, or is it impossible to say? Explain your reasoning.
- (b) Will one model fit the **testing** data better than the other, will they fit equally well, or is it impossible to say? Explain your reasoning.
- (c) Assume the true model is either Model 1 or 2. How will you determine whether Model 1 or Model 2 is a better model. Explain any one method.



4. Consider the training set accuracy and test set accuracy curves plotted below, during decision tree learning, as the number of nodes in the decision tree grows. This decision tree is being used to learn a function $f : X \rightarrow Y$, where training and test set examples are drawn independently at random from an underlying distribution $P(X)$, after which the trainer provides a noise-free label Y . Note $error = 1 - accuracy$. Please answer each of these true/false questions, and explain/justify your answer in 1 or 2 sentences.

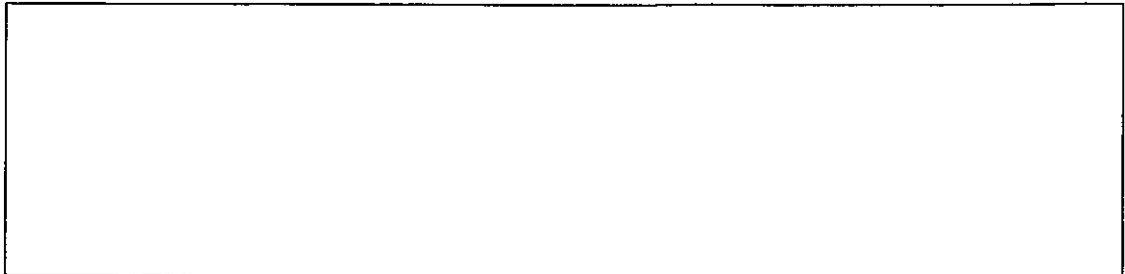
[8]



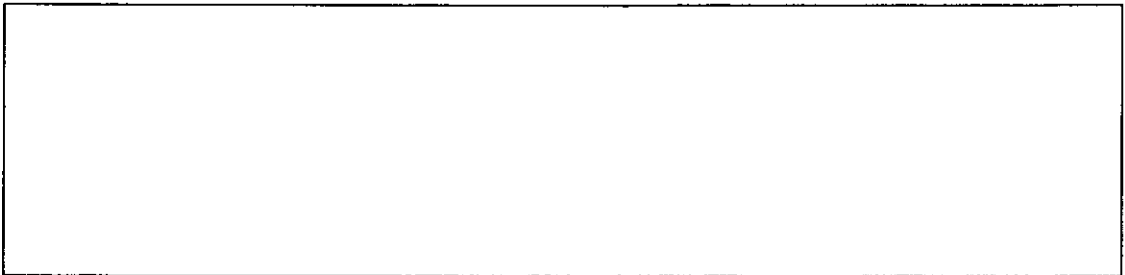
- (a) _____ Training error at each point on this curve provides an unbiased estimate of true error.

- (b) _____ Test error at each point on this curve provides an unbiased estimate of true error.

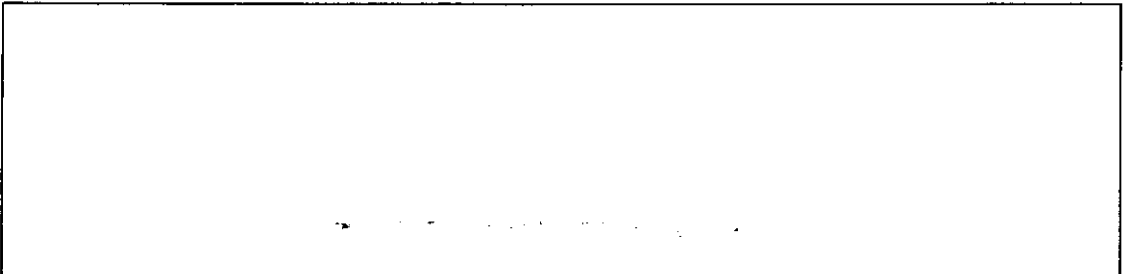
- (c) ____ Each time we draw a different test set from $P(X)$, the test accuracy curve may vary from what we see here.



- (d) ____ The variance in test accuracy will increase as we increase the number of test examples.



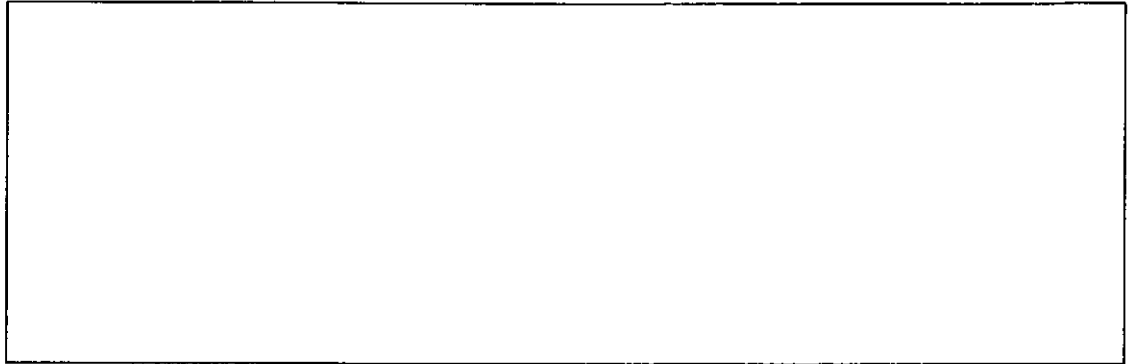
- (e) Given the above plot of training and test accuracy, which size decision tree would you choose to use to classify future examples? Give a one-sentence justification.



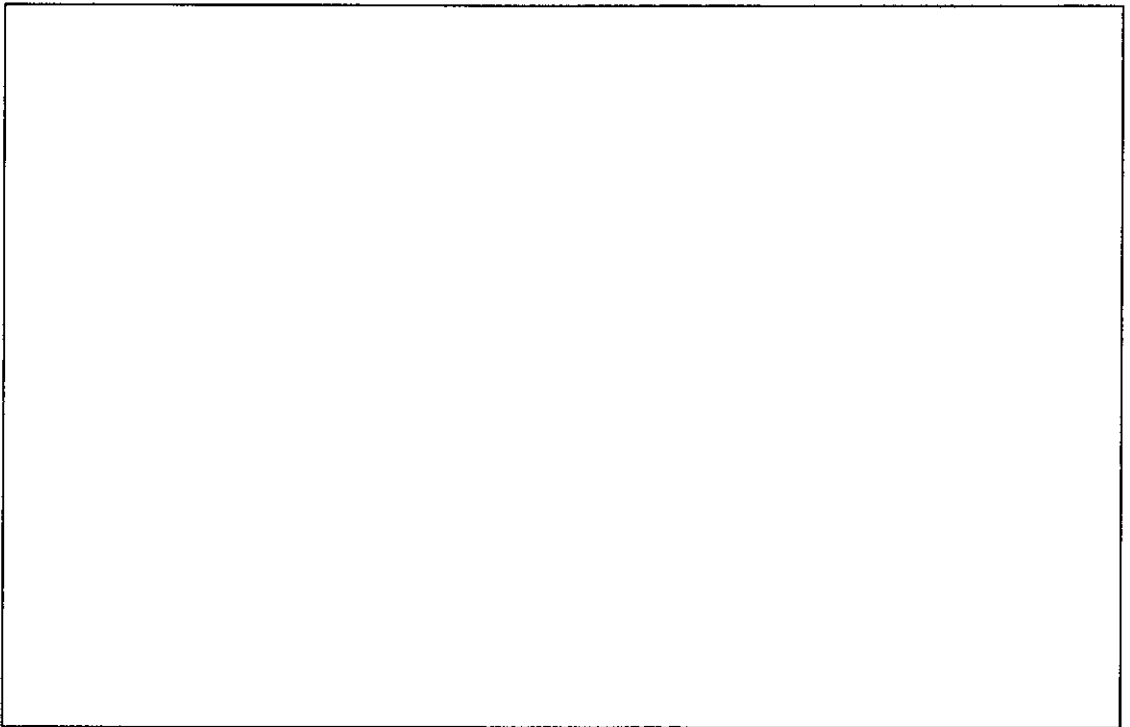
5. (a) Explain how k-fold cross-validation is implemented.

[7]



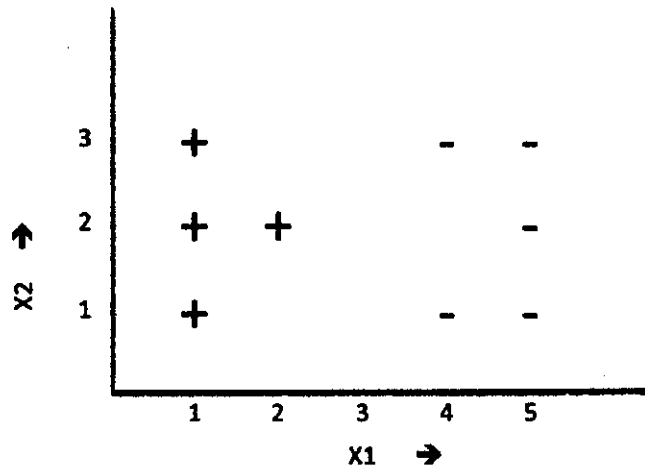


- (b) What are the advantages and disadvantages of k-fold crossvalidation relative to:
- i. The validation set approach?
 - ii. LOOCV?



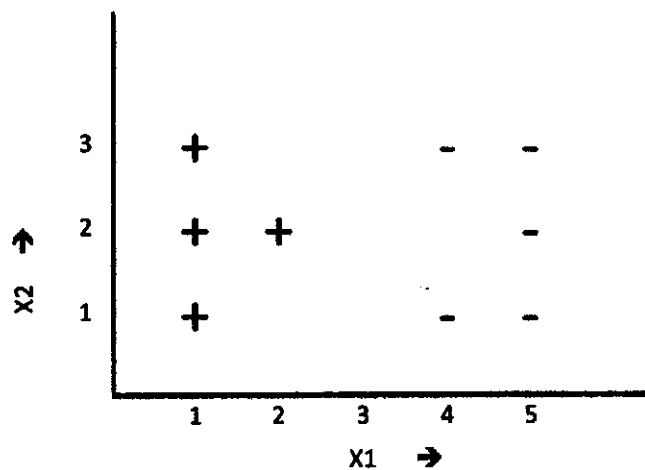
6. (a) Suppose we are using a linear SVM (i.e., no kernel), and are given the following data set.

[6]



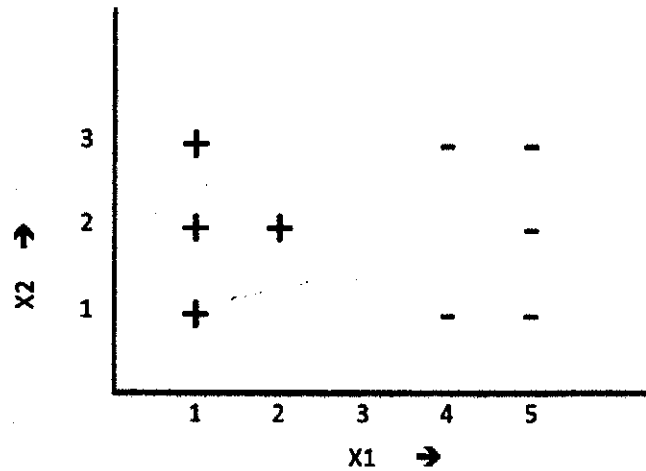
Draw the decision boundary of linear SVM in the above diagram. Give a brief explanation.

- (b) In the following image, circle the points such that removing that example from the training set and retraining SVM, we would get a different decision boundary than training on the full sample. You do not need to provide a formal proof, but give a one or two sentence explanation.



- (c) Suppose instead of SVM, we use regularized logistic regression to learn the classifier. In the previous image, circle the points such that removing that example from the training set and running regularized logistic regression, we would get a different decision boundary than training

with regularized logistic regression on the full sample.



7. Consider the following database.

[12]

Employee

ssn	ename	age	salary
314159265	Ray Diaz	35	89000
271828182	Nathan Logg	42	78000
161803398	Phyllis Bonacci	34	55000
030102999	Bryan Decimali	32	65000
141421356	Hiram Pottanoose	50	100000

Department

deptid	dname	mgrssn	budget
1	Hardware	314159265	531000
2	Firmware	314159265	420000
3	Software	161803398	678000
4	Sleepwear	141421356	88000

Works

essn	did	hours
314159265	1	20
314159265	2	20
271828182	1	10
271828182	2	20
271828182	3	5
271828182	4	5
161803398	2	10
161803398	3	30
030102999	2	20
030102999	3	20
030102999	4	20
141421356	4	40

The Works table records the hours an employee spends working for a department each week; the mgrssn field identifies the manager of a department.

(a) For each of the following SQL queries against the above database, show the resulting table.

- i.

```
SELECT ename, salary
FROM Employee
WHERE age < 40;
```
- ii.

```
SELECT dname, ename, budget, salary
FROM Employee, Dept
WHERE ssn = mgrssn
ORDER BY budget DESC;
```
- iii.

```
SELECT dname, SUM(salary)
FROM Employee, Dept, Works
WHERE ssn = essn AND deptid = did AND hours >= 20
GROUP BY deptid, dname;
```

- (b) Write the following queries in SQL against the above database schema.
- i. Retrieve the names and ages of all employees who work in both the Firmware department and the Software department.
 - ii. Retrieve the names of all employees who are not managers.
 - iii. Retrieve the name of the manager of the department with the largest budget.

Extra Page