Final

Graded

Student

ARNAV KRISHNAKUMAR MARDA

Total Points

87.5 / 100 pts

Question 1

(no title) **7** / 10 pts

1.1 a **1** / 4 pts

+ 2 pts Correct answer

+ 2 pts Correct explanation

✔ **+ 1 pt** If they mention something like 'this p-value indicates significance' but other parts are incorrect, they will receive only 1 point.

+ 0 pts Incorrect

1.2 b **4** / 4 pts

✔ **+ 2 pts** Correct answer

✔ **+ 2 pts** Correct explanation: any answer that refers to a uniform residual plot to justify that the model needs higher complexity.

+ 0 pts Incorrect

1.3 c **2** / 2 pts

✔ **+ 1 pt** Correct answer

✔ **+ 1 pt** Correct explanation

+ 0 pts Incorrect

Question 2

(no title) **18** / 18 pts

2.1 **a** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts The number of variables is not minimal, e.g., using one-hot encoding for race

– 3 pts No binary encoding of categorical variable

– 4 pts Incorrect

– 1 pt No explanation of what each variable captures

– 1 pt The term related to intercept is incorrect

– 1 pt The non-interaction term related to `age` is incorrect

– 1 pt The non-interaction terms related to `race` are incorrect

– 1 pt Interaction terms are incorrect

– 1 pt Extra terms

2.2 **b** **10** / 10 pts

✔ **– 0 pts** Correct

– 0 pts Marker: Used three binary encoding variables (8 betas) for (a)

– 1 pt Not mentioning 'odds ratio' or 'multiplicative change'

– 1 pt Not mentioning 'over other races'

– 1 pt Not mentioning 'for 1 unit increase in age ...'

– 1 pt Incorrect explanation regarding beta 0

– 1 pt Incorrect explanation regarding beta 1

– 2 pts Incorrect explanation regarding beta 2

– 2 pts Incorrect explanation regarding beta 3

– 2 pts Incorrect explanation regarding beta 4

– 2 pts Incorrect explanation regarding beta 5

– 6 pts Missing explanation for 3 coefficients

– 2 pts Gave interpretation of β instead of e^{β}

2.3 **c** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts just mention correlation, not collinearity

– 4 pts incorrect

Question 3

(no title) **8** / 8 pts

3.1 a **5** / 5 pts

✔ **– 0 pts** correct

– 1 pt incorrect turning point 1 (i.e., x=0, y=0.5)

– 1 pt incorrect turning point 2 (i.e., x=0.5, y=0.5)

– 1 pt incorrect turning point 3 (i.e., x=0.5, y=1)

– 1 pt incorrect turning point 4 (i.e., x=1, y=1)

– 4 pts Correct shape, but exact values of turning points not shown.

3.2 b **3** / 3 pts

✔ **– 0 pts** Correct

– 1 pt Not mentioning area under curve > 0.5 indicates better than random. (If plot is wrong)

– 2 pts Not mentioning a curve hugging the top left corner is good. (If plot is wrong)

– 3 pts incorrect

Question 4

(no title) **10** / 10 pts

4.1 a **4** / 4 pts

✔ **+ 2 pts** Correct answer

✔ **+ 2 pts** correct explanation

+ 0 pts Incorrect

+ 2 pts One figure correct

4.2 b **6** / 6 pts

✔ **+ 6 pts** All correct

+ 2 pts 3 classes (3 × ...)

+ 2 pts Correct answer

+ 2 pts selecting 3 out of 5 variables greedily (5+4+3)

+ 0 pts Incorrect

Question 5

(no title) **18** / 20 pts

5.1 **a** **4** / 4 pts

✔ **– 0 pts** correct

– 4 pts incorrect

5.2 **b** **8** / 8 pts

✔ **– 0 pts** Correct. If they include a minus sign before the answer, it's also considered correct.

– 2 pts $\frac{\partial L}{\partial p}$

– 2 pts $\frac{\partial p}{\partial z}$

– 2 pts $\frac{\partial z}{\partial h_2}$

– 2 pts $\frac{\partial h_2}{\partial w_2}$

– 0.5 pts minor mistake 1 (see the solution file)

– 1.5 pts minor mistake 2 (see the solution file)

– 0 pts y(1-p)-p(1-y)

5.3 **c** **4** / 4 pts

✔ **– 0 pts** Correct

– 2 pts if you have $w_2 = w_2 + \eta \frac{\partial L}{\partial w_2}$

– 3 pts if you have $\frac{\partial L}{\partial w_2}$

– 4 pts Incorrect

5.4 **d** **2** / 4 pts

– 0 pts Correct

– 2 pts didn't mention data is not linearly separable

– 2 pts If you didn't mention that the model is linear (or logistic regression) because of linear/identity activation functions

– 4 pts If you say the above network predicts -1 or +1 with probability 0.5. This is not the reason, even if you train the network and optimize the weights, you cannot still separate the above data with a linear model.

– 4 pts Incorrect

– 4 pts No attempt

✔ **– 2 pts** partially incorrect answer, lacking details

Question 6

(no title) **22** / 24 pts

6.1 **a** **10** / 10 pts

- ✔ **– 0 pts** Correct. Even if they only show the final result
- **– 1 pt** wrong coefficient in term for {}
- **– 1 pt** wrong coefficient in term for {1}
- **– 1 pt** wrong coefficient in term for {3}
- **– 1 pt** wrong coefficient in term for {1, 3}
- **– 1 pt** wrong difference in term for {}
- **– 1 pt** wrong difference in term for {1}
- **– 1 pt** wrong difference in term for {3}
- **– 1 pt** wrong difference in term for {1, 3}
- **– 1 pt** extra term
- **– 2 pts** Correct parts but incorrect final formula
- **– 8 pts** Correct formula did not use specific numbers, or numbers incorrect
- **– 10 pts** Incorrect

6.2 **b(a)** **2** / 3 pts

- ✔ **+ 1 pt** Superpixels are selected at random
- ✔ **+ 1 pt** LIME partitions the image into superpixels
- **+ 1 pt** The superpixels that weren't selected are greyed out
- **+ 0 pts** Incorrect

6.3 **b(b)** **3** / 3 pts

- ✔ **+ 1 pt** β_0
- ✔ **+ 1 pt** $\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- ✔ **+ 1 pt** what each variable captures
- **+ 0 pts** Incorrect

6.4 **b(c)** **2** / 2 pts

- ✔ **+ 1 pt** The explanation of β_0 is correct
- ✔ **+ 1 pt** The explanation of β_i is correct
- **+ 0 pts** Incorrect

6.5 **b(d)** **3** / 4 pts

– **0 pts** Correct. both equation and explanation

– **0.5 pts** $\hat{y}$ is missing or used wrong.

– **0.5 pts** $\hat{y}$ has an incorrect explanation.

– **0.5 pts** y is missing or used wrong.

– **0.5 pts** y has incorrect explanation.

✔ – **0.5 pts** π_j is missing or used wrong.

✔ – **0.5 pts** π_j has incorrect explanation.

– **1 pt** No sum (over the number of perturbed versions of the image) of the **squared** loss

6.6 **b(e)** **2** / 2 pts

✔ – **0 pts** Correct

– **1 pt** each incorrect/extra term

– **1 pt** each incorrect/extra term

Question 7

(no title) **4.5** / 10 pts

7.1 1 **2** / 2 pts

✔ **– 0 pts** Correct

– 1 pt wrong answer

– 1 pt wrong explanation

7.2 2 **0.5** / 2 pts

– 0 pts Correct

✔ **– 1 pt** wrong answer

– 1 pt wrong explanation

✔ **– 0.5 pts** partially correct explanation

7.3 3 **0** / 2 pts

– 0 pts Correct

✔ **– 1 pt** wrong answer (for answer of (d), please note: it is correct that with regularization decision boundary is smoother, but still due to relu it should be piece-wise linear (i.e., c not d))

✔ **– 1 pt** wrong explanation

– 0.5 pts Partially correct explanation

7.4 4 **1** / 2 pts

– 0 pts Correct

✔ **– 1 pt** wrong answer

– 1 pt wrong explanation

7.5 5 **1** / 2 pts

– 0 pts Correct

✔ **– 1 pt** wrong answer

– 1 pt wrong explanation

– 0.5 pts Partially correct explanation

Data Science (CS 148)
March 19, 2024

Final Exam

Duration: 120 minutes

Write your name and UID: Arnav Marda .

Note 1: If you find a question difficult, move on with the rest of the questions and come back to it in the end!

Note 2: Your final grade will be curved, if necessary.

Note 3: There are 7 questions. Only answers written in the boxes will be graded. If you need extra space, please ask for an extra sheet. Good Luck! :-)

1 Linear Regression (10 Points)

For a given data collected from 10 individuals, we use linear regression $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$ to model the effect of the dosage of a new drug (X_1) on the cholesterol level (y) for males vs females (X_2) in the data. X_2 is a binary variable, which takes $X_2 = 1$ for males and $X_2 = 0$ for females.

(a) **(4 points)** If $\beta_3 = -2$, with p-value < 0.05, can we conclude that one unit increase in the dosage of the drug is more effective in lowering the cholesterol level for males? **(First answer box)**. Explain your answer briefly **(Second answer box)**.

Yes/No(s): Yes

Explanation: $y = \beta_0 + \beta_2 + (\beta_1 - 2) X_1$ for males. Thus, we can see that compared with females ($y = \beta_0 + \beta_1 X_1$), a 1 unit ~~decrease~~ increase in dosage will reduce the cholestrol level 2 units more for males than females. A p-value < 0.05 suggests that the result is statistically significant and we have 95% confidence in the result.

(b) **(4 points)** For your model, $R^2 = 0.92$ suggesting a strong linear relationship. You plot the distribution of the residuals and see that residuals have a uniform distribution. Is the linear model good for this problem? **(First answer box)**. If yes, why? If no, how do you change your model? **(Second answer box)**

(Yes / No): No .

Explanation: Uniform distribution of residuals suggests underfitting of the model. Thus, it might be better to use a quadratic model rather than a linear one.

(c) **(2 points)** For particular values of predictors, the 95% prediction interval for cholesterol is (200,235). You can request for approval of the drug *only* if you can show that you are confident about cholesterol ≤ 230. Can you ask for approval based on the above prediction interval? **(First answer box)** Briefly explain why. **(Second answer box)**

(Yes / No): NO .

Explanation: Since the 95% confidence interval is (200, 235) where $235 \geq 230$, it is not statistically correct to ask for approval based on the above prediction interval.

2 Logistic Regression (18 Points)

We use logistic regression to model the odds of a low birthweight, based on 'age-17' (X_1) and race (X_2) of the mother. Age is a real-valued variable and race is a categorical variable that can take 3 values: white, black, and others.

(a) **(4 points)** Write the logistic regression formulation to model log (use ln) odds of a low birthweight, based on whether the age and race of the mother. Include all the possible interaction terms in your model, and mention what each variable captures. *Hint: use 'minimum number' of variables possible in your model.*

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2^W + \beta_3 X_2^B + \beta_4 X_1 X_2^W + \beta_5 X_1 X_2^B .$$

If mother's race is white $\Rightarrow X_2^W = 1, X_2^B = 0$
If mother's race is black $\Rightarrow X_2^W = 0, X_2^B = 1$
If mother's race is other $\Rightarrow X_2^W = 0, X_2^B = 0$

$X_1 \rightarrow$ real ~~valued~~ valued variable for age-17 of the mother.

(b) **(10 points)** Interpret the coefficients of your model. *Note: write the interpretation for each e^{β_i} (do not write interpretation for $e^{\beta_i+\beta_j}$ or β_i).*

~~If mother white:~~

$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_2 + (\beta_1 + \beta_4) X_1$

$e^{\beta_0} \rightarrow$ base birthweight for mother of other race ~~mean~~

$e^{\beta_1} \rightarrow$ change in odds for unit change in X_1 for mother of other race.

$e^{\beta_2} \rightarrow$ ratio of base odds between white and other mother with X_1 constant.

$e^{\beta_3} \rightarrow$ ratio of base odds between black and other mother with X_1 constant.

$e^{\beta_4} \rightarrow$ ratio between change in odds per unit change in X_1 for white and other race mother.

$e^{\beta_5} \rightarrow$ ratio between change in odds per unit change in X_1 for black and other race mother.

(c) **(4 points)** Suppose we fit two other logistic regression models: The first logistic regression (LR1) models the odds of a low birthweight only based on the age of the mother; The second logistic regression (LR2) models the odds of a low birthweight only based on the race of the mother. If the coefficients for age and race in LR1, LR2 are different than the model in (a), what do you conclude?

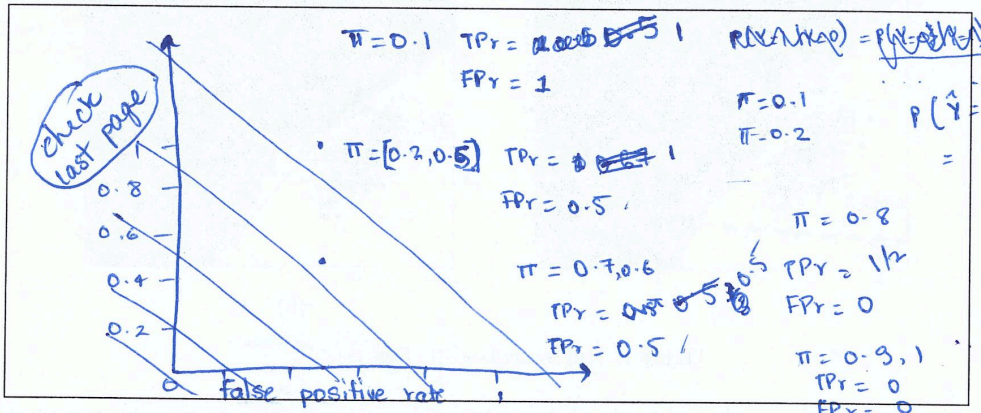
We then conclude that the variables are collinear and not unique. This leads to an incorrect model and difficulty in interpreting coefficients. Dropping one of the variables is a possible solution.

3 Classification Metrics (8 Points)

You use a binary classification model on some testing data and generate the following table listing out the set of probabilities for a positive label (Label = 1). The table then also includes the actual label.

Probability	Actual Label
.8	1
.7	0
.5	1
.1	0

(a) **(5 points)** Using the information in this table, draw out a plot of the ROC Curve (be sure to label your axes and be precise with your (X,Y) coordinate values [points will be deducted if the graph is only approximate]). *Hint: ROC curve shows true positive rate on the y-axis vs. false positive rate on the x-axis, where true positive rate is the probability that an actual positive will test positive, and false positive rate is is the proportion of all negatives that still yield positive prediction.*



(b) **(3 points)** What does the above ROC curve indicates about your model?

The model is fairly good at the classification problem since AUC will be high. The AUC is approximately 0.75 which is an indicator of a good model

4 Multi-class Logistic Regression (10 Points)

(a) **(4 points)** The following plots (Figure 1) show the decision surface of multinomial and One-vs-Rest Logistic Regressions. The lines corresponding to the individual classifiers are represented by the dashed lines (C1, C2 and C3 correspond to classifier 1, 2 and 3 respectively). Which figure corresponds to:

1. Multinomial
2. OVR

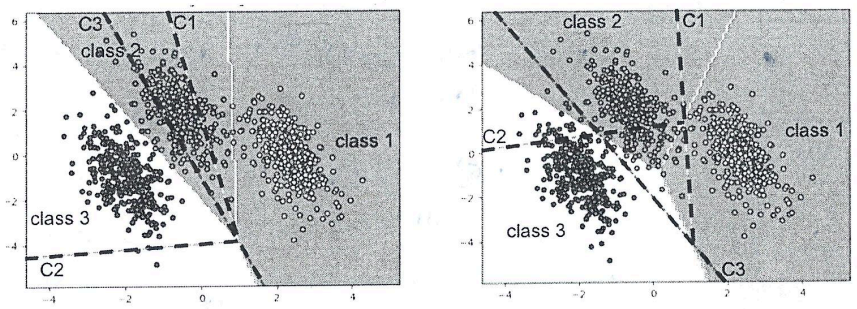
Write down either 1. or 2. in each of the answer boxes provided below.

a. 1	b. 2

Briefly explain your choice.

In fig (b), we can clearly see that each of C1, C2, C3 is trying to seperate their respective class from the other classes, thus (b) is OVR.

Explanation:
In fig (a), assuming that class 2 is the reference class, C1 seperates class 2 from 1 and C3 seperates class 2 from 3. Thus, it is multinomial.



(a) (b)

Figure 1: Multiclass Logistic Regression

(b) **(6 points)** We want to do greedy (forward) model selection (greedy step-wise variable selection) for *each of the linear classifiers* in an OvR logistic regression. If we have 3 classes, and 5 predictors, how many logistic regression models should we fit in total to choose 3 predictors for each model? Assume, we only consider the original predictors in the model selection without including any higher order terms.

Please write the final answer as a number in the left box, and show your calculation in the right box.

36.	for each classifier, we have 5 + 4 + 3 = 12 models. Calculation: Since we have 3 classes, we have 3 classifiers i.e 3 × 12 = 36 total models to train

5 Neural Networks (20 Points)

Consider the following neural network with linear (Identity) activation functions in hidden nodes h_1, h_2, sigmoid output p, and binary cross entropy loss: $\mathcal{L} = \sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i)$. Assume that we are

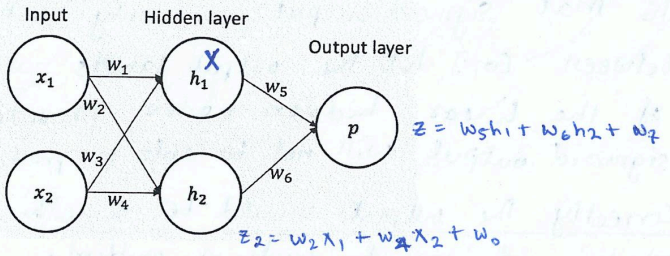


Figure 2: Neural Network

training the network with *dropout*, and in training iteration i we keep hidden neuron h_2, and drop hidden neuron h_1.

(a) **(4 points)** Which weights are getting updated in this iteration? *Hint: every training iteration consists of one forward pass and one backward pass.*

w_2, w_4 and w_6 are being updated in this iteration.

(b) **(8 points)** Use the chain rule to write the derivative of the loss w.r.t. w_2. *Hint.* $\frac{\partial \log z}{\partial z} = \frac{1}{z}$, and for a sigmoid function p, we have $\frac{\partial p(z)}{\partial z} = p(1-p)$.

$$\frac{\partial \mathcal{L}}{\partial w_2} = \frac{\partial \mathcal{L}}{\partial p} \cdot \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_2}$$

$$= \left[\frac{y}{p} - \frac{1-y}{1-p}\right] \cdot p(1-p) \cdot w_6 \cdot x_1$$

$$= x_1 w_6 \left(y(1-p) - p(1-y) \right) .$$

(c) **(4 points)** Write the gradient descent update rule for w_2 in this iteration, using η as learning rate.

$$w_2^{(i+1)} = w_2^{(i)} - \eta \frac{\partial L}{\partial w_2} = w_2^{(i)} - \eta x_1 w_6 \left(y(1-p) - p(1-y) \right) .$$

(d) **(4 points)** Consider the following points: Provide a simple argument explaining why the above neural

x_1	0	0	1	1
x_2	0	1	0	1
class	+1	-1	-1	+1

network cannot perfectly classify these points (You can choose to draw a picture if you prefer).

A possible reason as to why the model will not work is that sigmoid output can only return values between [0,1] but the output labels are the interaction of the linear hidden nodes providing input to the sigmoid output will not be able to predict the values correctly. The output would be something like  which is not ideal for such a problem.

6 Black-box Interpretability (24 points)

(a) **Shapley Value (10 points)** You want to use Shapley values to explain why a particular image is classified as a 'Cat', by a neural network. To do so, we first divide the image to 3 super-pixels. The following table shows the probability for the image to be in class 'Cat' by the neural network, when we input different combinations of super-pixels (Super-pixels kept) to the network and replace the other ones with 'gray' color. Calculate the Shapley value for super-pixel 2, indicating the contribution of this super-pixel to prediction of 'Cat'. *Note: you can keep your answer as a summation, no need to calculate the final number.*

Remember: multinomial coefficient can be calculated as $\binom{n}{m} = n!(n-m)!/n!$. *Note that we do not use this exact formula for computation of the Shapley value.*

Super-pixels kept	Probability of 'Cat'
{}	0.0
{1}	0.4
{2}	0.5
{3}	0.4
{1,2}	0.8
{1,3}	0.5
{2,3}	0.8
{1,2,3}	0.9

$$\phi_2 = \frac{0!\,2!}{3!} \times 0.5 + \frac{1!\,1!}{3!} \times 0.4 + \frac{1!\,1!}{3!} \times 0.4 + \frac{2!\,0!}{3!} \times 0.4$$

$$= \frac{1}{3} \times 0.5 + \frac{1}{6} \times 0.4 \times 2 + \frac{1}{3} \times 0.4$$

$$= \frac{1}{3}[0.5 + 0.4 + 0.4] = \frac{1.3}{3} \approx 0.433$$

(b) **LIME (14 points)** Next, we consider using LIME for the above problem. Remember that LIME trains a linear model on the perturbed versions of the image.

(a) **(3 points)** Explain briefly how LIME generates the perturbed version of the image.

LIME will first split the image into superpixels as we did in part (a). Consider 3 super-pixels. Then to generate perturbed versions of the image, LIME will randomly switch on or off the 3 superpixels in each iteration creating a perturbed image. For example, it can grey out superpixels 1, 2 and leave 3 on.

(b) **(3 points)** Write the equation for the linear model and mention what each variable/parameter captures.

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

$x_1 \rightarrow 1$ if superpixel 1 is in image, 0 otherwise

$x_2 \rightarrow 1$ if superpixel 2 is in image, 0 otherwise

$x_3 \rightarrow 1$ if superpixel 3 is in image, 0 otherwise.

(c) **(2 points)** Interpret the coefficients of the linear model in part (b).

$\beta_0 \rightarrow$ base probability of predicting cat from the model.

$\beta_1 \rightarrow$ ~~change~~ diff. in probability of predicting cat from base given that superpixel 1 is on.

$\beta_2 \rightarrow$ ~~change~~ diff. in probability of predicting cat from base given superpixel 2 is on.

$\beta_3 \rightarrow$ ~~change~~ diff. in probability of predicting cat from base given superpixel 3 is on.

(d) **(4 points)** Write the loss function for training LIME's linear model, and explain what each variable is in the loss function.

The loss function would be MSE.

$L = \Sigma_i (\hat{y} - y)^2$

where $\hat{y}$ is the predicted probability from LIME and y is the actual probability

(e) **(2 points)** Assume you trained the above linear function on perturbed super-pixels. How do you use this linear model to calculate the contribution of first and second super-pixels together, to the prediction of the class 'Cat'? Write the (parametric) equation.

$\hat{y} = \beta_0 + \beta_1 + \beta_2$.

7 Decision Boundary (10 Points)

Consider the dataset with two classes in the figure below. In each of the plots (a)-(e), one of the following classification methods has been used, and the resulting decision boundary is shown:

(1) Multi-layer Neural Network with linear activation functions

(2) Multi-layer ReLU network

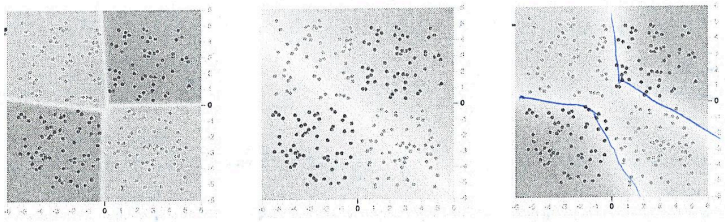
(3) Regularized Multi-layer ReLU network

(4) Multi-layer tanh network

(5) Regularized Multi-layer tanh network

Assign each of the previous methods to exactly one of the following plots (in a one to one correspondence) by matching the plots with the respective letters, and **explain briefly** why did you make each assignment in the answer box we provided.

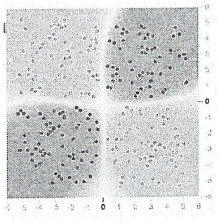
Write your choices (a, b, c ...) in the answer boxes provided below and your explanation in the corresponding explanation box.



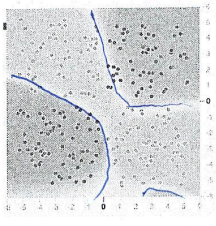
(a)

(b)

(c)



(d)



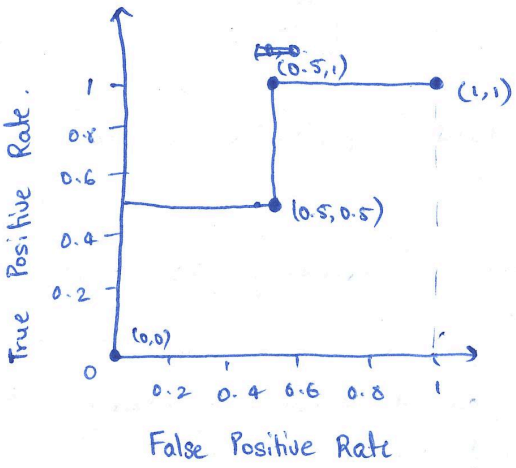
(e)

(1) b.	Explanation: Since the decision boundary is a linear with a single line.

Final Exam

(2) (c)	Explanation: The decision boundary is piecewise linear with co-efficients not tending to 0.
(3) (a)	Explanation: Piecewise linear decision boundary with coefficients tending to 0 from regularization.
(4) (e)	Explanation: Continuous ~~smooth~~ curved decision boundary with noticeable co-efficients not tending to 0.
(5)	Explanation: Smoother curved decision boundary with co-efficients tending to 0 because of regularization.

3 (a)


True Positive Rate.
1
0.8
0.6
0.4
0.2
0
(0.5, 1)
(1, 1)
(0.5, 0.5)
(0, 0)
0.2
0.4
0.6
0.8
1
False Positive Rate