Minor 1

Total marks: 20 pts

Total time: 45 mins

Instructions:

- **Write your name and roll number on answer script**
- With the exception of Question 1, all your answers must be clearly motivated! *A correct answer without a proper motivation will score zero points*!
- Vectors are denoted with a single underline $\underline{a}$, and matrices by double underline, $\underline{\underline{A}}$. Scalars appear without any underline. Please follow this rule through your answer book.

1. [**5 pts**] Answer `True` or `False`.
Each correct answer scores 1 point, each incorrect answer scores −1 point and each missing answer scores 0 point.

F (a) Linear regression requires all input variables to be numerical (quantitative).

T (b) The k-nearest neighbors (kNN) algorithm is sensitive to the choice of distance metric used to measure the similarity between data points.

F (c) Logistic regression can only be used for binary classification problems, and it is not suitable for multi-class classification.

T (d) The link function in a Generalized Linear Model defines the relationship between the input and the expected value of the response variable.

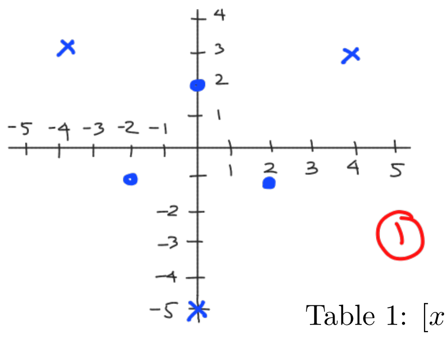
F (e) The squared hinge loss is more robust to oultiers than hinge loss.

Regression
Output is numerical
(can be compared)

2. [**1 pt**] A research team is working on predicting the time to failure of a certain metal cutting saws. They aim to develop a model that estimates the failure time in hours. However, due to the discrete nature of failure times (measured in whole hours like 1hr, 2hrs, etc.), they are debating whether to treat this problem as a regression or classification task.

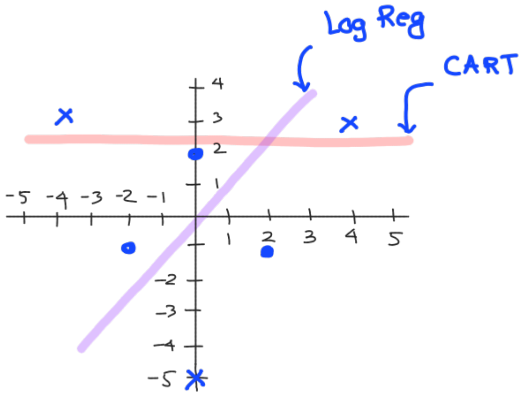
Discuss whether predicting saw failure time in minutes should be viewed as a regression or classification problem. Provide reasons for your choice.

3. [**5 pts**] Consider the dataset in Table 1.



x_1	x_2	y
−2	−1	○
2	−1	○
0	2	○
−4	3	×
0	−5	×
4	3	×

Table 1: $[x_1 \;\; x_2]^T$ is the input and y is the class label



(a) **[1 pt]** Illustrate the dataset in a graph with x_1 and x_2 on the two axes.

(b) **[4 pt]** A student has worked with the data and has tried four different classifiers

① 33% (i) Logistic regression (a line through origin misclassifies at best two pts)

① 0% (ii) kNN with $k = 1$ (the nearest neighbour of each data pt is itself; perfectly overfits)

① 50% (iii) kNN with $k = 3$ (the three 'x's are misclassified $3/6 \times 100$)

① 17% (iv) A classification tree with a single best binary split based on misclassification rate ($x_2 \leq 2.5$) best split ↓ one 'x' misclassified $1/6 \times 100$

The student has also computed the misclassification error on the training dataset for each classifier. The misclassification errors are 50%, 33%, 17% and 0%. Unfortunately, the student did not report which classifier gave what error!

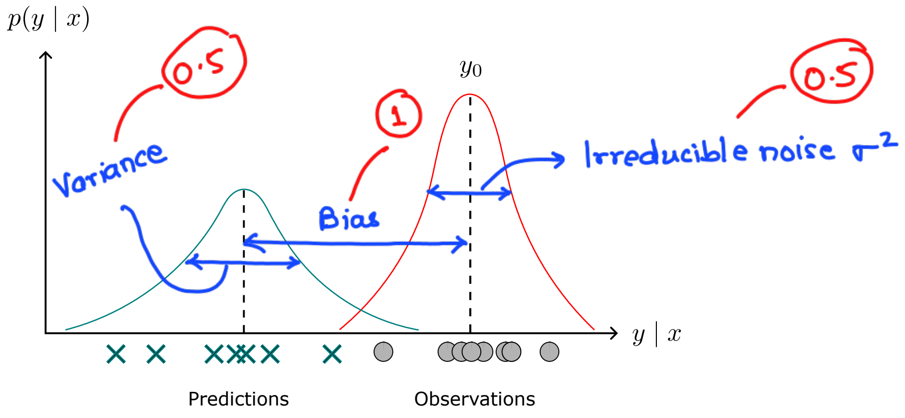
Determine which misclassification error corresponds to which classifier?

Note: *A correct pairing without a proper explanation will get zero points*

4. **[2 pts]** Consider the figure below. The figure plots $p(y \mid x)$ against $y \mid x$. The term $y \mid x$ represents either observations (shown in filled circles) or model predictions (shown in ×). The probability distribution of measurements given inputs is shown on the right, while that of model predictions is shown on the left. The measurements (denoted by filled circles) were obtained using a model:

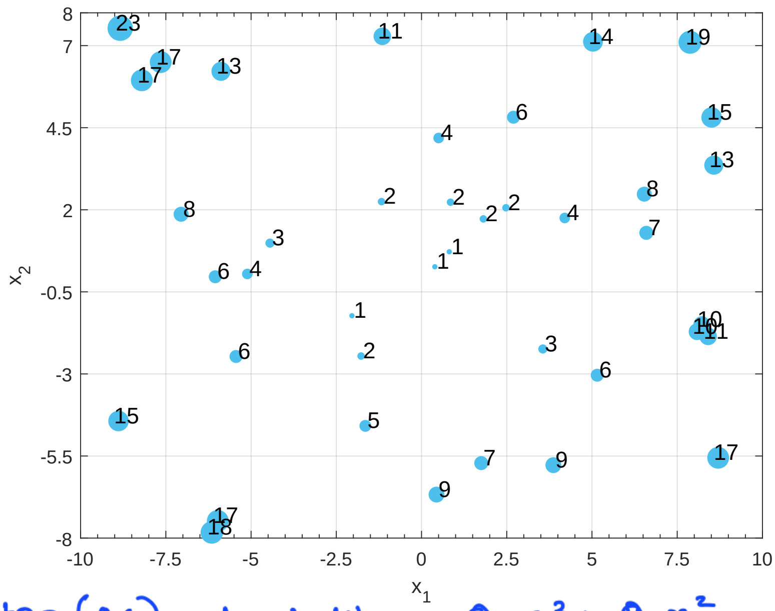
$$y = y_0(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Here $y_0(x)$ is a deterministic function.



Redraw the plot on your answer script. Sketch and denote the **bias**, **variance**, and **irreducible noise** on the plot.

5. **[2 pts]** Consider the following training dataset with inputs x_1, x_2 and output y:



Justification

The values of y are always positive irrespective of the values of x_1 and x_2.

Also, the values tend to increase gradually from center (0,0), almost like $\Theta_1 x_1^2 + \Theta_2 x_2^2$

The area of the circles is proportional to the value of y, i.e. larger values of y have larger circles. The numbers also show the value of y. By looking at the data, suggest input features for a linear regression model which you think might **avoid overfitting the data**. Justify your answer.

$\underbrace{|x_1|, |x_2|}_{②}$, or $\underbrace{x_1^2, x_2^2}_{②}$

Just writing one is not enough. x_1, x_2 are wrong choices

6. **[5 pts]** Given N input and output data points $\{(x_i, y_i)\}_{i=1}^{N}$, consider the model

$$y(x) = f(x) + \epsilon$$

where ϵ is independent zero-mean noise with variance σ^2 and

$$f(x) = \sum_{i=1}^{N} \theta_i \phi_{x_i}(x), \text{ with } \phi_{x_i}(x) = e^{-(x-x_i)^2}$$

The parameters θ are learned by minimizing the mean squared error.

(a) **[1 pt]** Is the model parametric or non-parametric? ✓ Explain.

No. of parameters scale with data ①

(b) **[2 pts]** Show that the parameters can be learned using linear regression.
Hint: Write the model as a matrix multiplication $\underline{y} = \underline{\underline{X}}\ \underline{\theta} + \underline{\epsilon}$ and give an expression for matrix $\underline{\underline{X}}$

$\{$ $\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

① $\underline{\underline{X}} = \begin{bmatrix} \phi_{x_1}(x_1) & \phi_{x_2}(x_1) \\ \phi_{x_1}(x_2) & \phi_{x_2}(x_2) \\ \vdots & \\ \phi_{x_1}(x_N) & \phi_{x_2}(x_N) \end{bmatrix}$ N×N

(c) **[2 pts]** To avoid overfitting, we use ridge regression with weight λ and learn the parameters by minimizing the cost function

$$J(\underline{\theta}) = \left\| \underline{y} - \underline{\underline{X}}\ \underline{\theta} \right\|_2^2 + \lambda \left\| \underline{\theta} \right\|_2^2$$

where $\|\cdot\|_2$ is the standard ℓ_2-norm.
Write down the expression for the prediction of a test input point, $f(x_*)$.

① $\hat{\underline{\Theta}} = (\underline{\underline{X}}^T \underline{\underline{X}} + \lambda \underline{\underline{I}})^{-1} \underline{\underline{X}}^T \underline{y}$

① $\{$ $\underline{y} = \underline{\underline{X}}\ \underline{\Theta} + \underline{\epsilon}$

① $y_* = f(x_*) = \underline{x}_*^T \hat{\underline{\Theta}} = \begin{bmatrix} \phi_{x_1}(x_*) \\ \vdots \\ \phi_{x_N}(x_*) \end{bmatrix}^T \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_N \end{bmatrix}$