National University of Singapore
School of Computing
IT5005 Artificial Intelligence

Introduction to Learning

1. **Linear Regression Model Fitting**.

   You are given several data points as follows:

x_1	x_2	x_3	y
6	4	11	20
8	5	15	30
12	9	25	50
2	1	3	7

   Apply the Normal Equation formula to obtain a linear regression model that minimizes MSE of the data points.

$$w = (X^T X)^{-1} X^T Y$$

2. **Examining Cost Functions**.

   For Linear Regression, there are two popular cost functions,

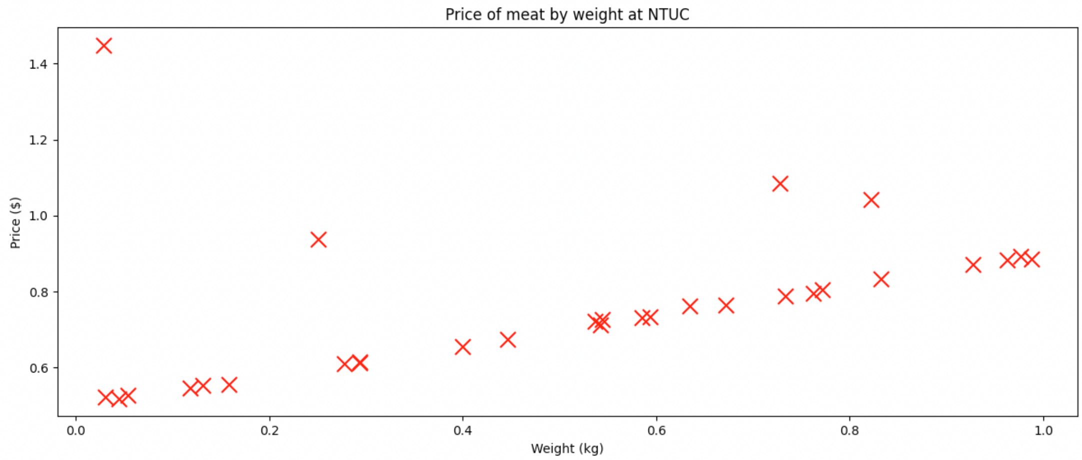
   Mean Squared Error:

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2 \tag{1}$$

   and **Mean Absolute Error**:

$$L(y, \hat{y}) = \frac{1}{2}|y - \hat{y}| \tag{2}$$

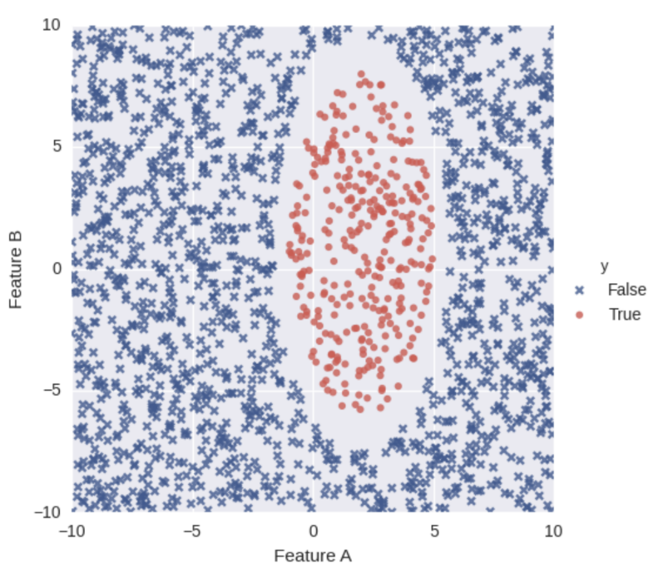
   (a) Given the scatter plot of a dataset containing the actual weight of meat at NTUC (x) and its price (y), justify your choice of cost function for this problem.



(b) Can you provide examples of cost functions that are better suited to handle outliers more effectively?

3. **Linear vs Non-linear Separability**

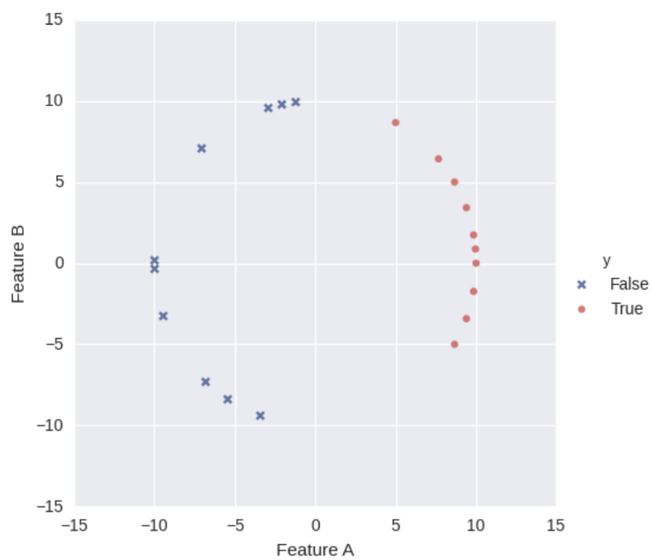
Bondrewd Workshop is a company that produces cute fluffy bunnies through experimentation and genetic mutations. Quality control for the bunnies is done manually. A group of scientists decide whether a bunny is ready to be released into the wild based on two features: **Feature A** is a bunny's cuteness score and **Feature B** is a bunny's fluffiness score. The figure below show examples of bunnies that have been released and withheld in the past. Each dot corresponds to a bunny, and responds to the following question as true or false: this bunny is ready to be released.



Knowing that you are an ML expert, Bondrewd the CEO has approached you and asked you to automate the decision making process.

(a) Define a reasonable set of features that will perfectly classify whether or not a bunny can be released into the wild.

(b) Bondrewd decides to change the production direction in the company. Bondrewed Workshop will be creating fewer, but cuter (and fluffier) bunnies. After more experiments, they have collected the examples again in the figure below.



Define a reasonable set of features that will perfectly classify whether or not a bunny can be released into the wild.

4. **Logistic Regression for Multi-Class Classification**.

Suppose you have a classification task of deciding whether an animal is a cat, a horse, or an elephant. However, you can't see the animal but you have the information about

- The weight of the animal (in kilogram)
- The length of the animal (in meter)

You, being an ML Expert, suggested to use 3 Logistic Regression models to solve this problem. After training on the training dataset, you get the following parameters:

$$w_{cat} = [4.2, -0.01, -0.12]$$
$$w_{horse} = [-20, -0.08, 35]$$

$$w_{elephant} = [-1250, 0.82, 0.9]$$

(a) You're given a list of animals with their features. Compute the probability of an animal belonging to a certain class and classify them accordingly.

Weight (kg)	Length (m)
4.2	0.4
720	2.4
2350	5.5

Table 1: List of animals with unknown class

5. **Precision, recall, F1 score and ROC curve**

Esophageal cancer is a serious and very aggressive disease. In this question, we want to look at the size of a patient's tumor and decide whether the cancer has spread to his or her lymph nodes. Using what we learnt, we use maximum dimension (mm) of esophagus tumor as input, and label 1 if the cancer had spread to their lymph nodes, and 0 otherwise.

We derived this machine learning model M which outputs a continuous score for every input sample. Figure 1 shows the results for 20 samples from the model. The actual labels can be either 1 (red, positive label) or 0 (blue, negative label). The model output makes the final classification decision. If a threshold, p is given, model M outputs label 1 if M(x) is greater than or equal to the threshold, otherwise the model outputs 0.

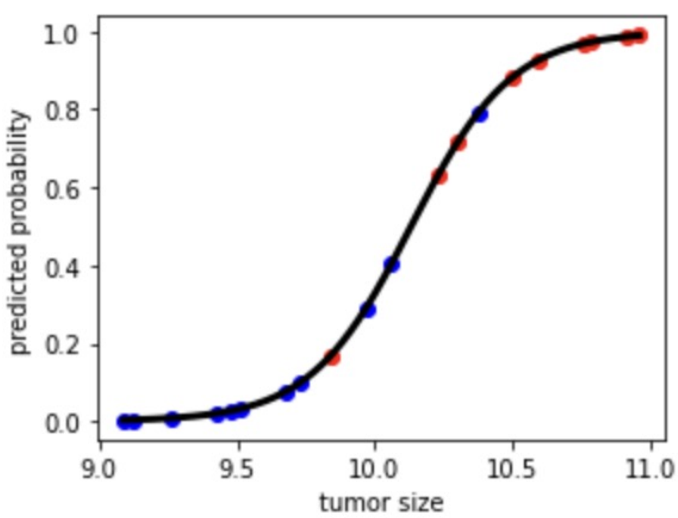


Figure 1: Model probability output and tumor size

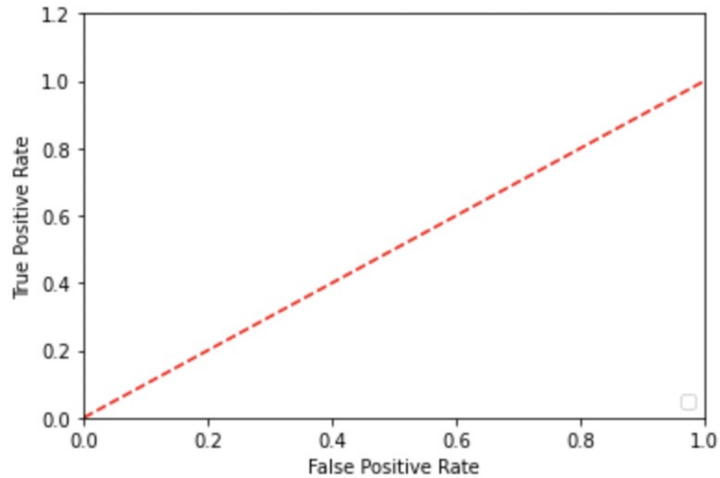


Figure 2: Blank ROC curve for convenience

(a) For the threshold, $p = 0.5$, tabulate the confusion matrix.

(b) For the threshold, $p = 0.5$, find the precision, recall and F1 score.
Hint: Precision = $\frac{TP}{TP+FP}$ *Recall =* $\frac{TP}{TP+FN}$ *F1 score =* $\frac{2TP}{2TP+FP+FN}$

(c) Based on Figure 1, derive the ROC curve
(you may draw your ROC on top on the Figure 2).
Hint: TPR = $\frac{TP}{ActualPositive}$ = $\frac{TP}{TP+FN}$ *FPR =* $\frac{FP}{ActualNegative}$ =$\frac{FP}{TN+FP}$
Hint 2: Tabulate a confusion matrix and from there, calculate the true positive rates and false positive rates. Mark out the corresponding point on the graph. Repeat this for at least four different thresholds.

(d) Based on the ROC curve you derived, decide which threshold you want to choose among p = 0.2, p = 0.5 and p = 0.8.

(e) In this question's case for detecting tumours, should we maximize precision or recall? Explain the reason for your choice.

(f) Suppose now we want to detect plagiarism instead, should we maximize precision or recall? Explain the reason for your choice.

6. **Perceptron Learning Algorithm**.

Consider the dataset shown in Table 2. Mr. Aiken would like to build a linear classification model. To this end, he initialized the weight vector as $\mathbf{w} = [0, 1, 0]$. He is keen on solving this problem by hand. Assume that Aiken is using perceptron learning algorithm with a learning rate of 0.1. You need to help him in getting started by answering the following questions:

Data Index	Feature A	Feature B	Label
1	1	0	-1
2	1	1	-1
3	1	2	1
4	2	1	1

(a) Using the initial weight vector, find the predicted label for each data point.

(b) Do the weights need update? If yes, perform one iteration of the weight update.

(c) Could Rosenblat's PLA converge to a solution? Provide the rationale.