National University of Singapore
School of Computing
CS2109S: Introduction to AI and Machine Learning
Semester 2, 2024/2025

Tutorial 5
Classification and Logistic Regression

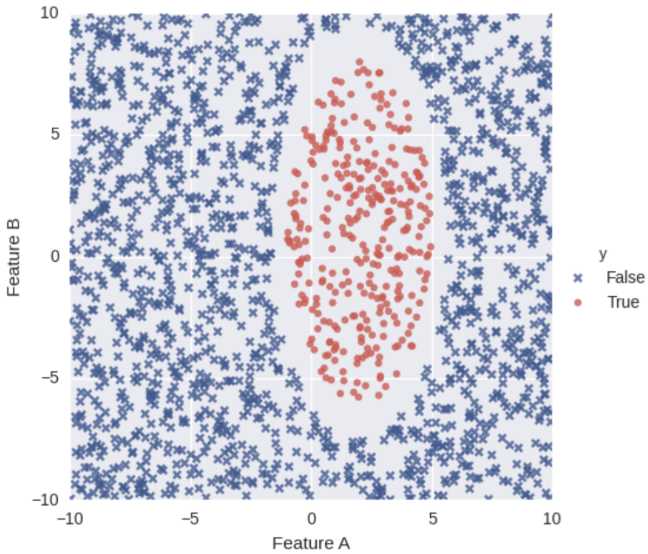
Summary of Key Concepts

In this tutorial, we will discuss and explore the following learning points from Lecture:

1. Classification
   (a) Linear vs Non-linear Separability
2. Logistic Regression
   (a) Loss function
   (b) Choosing suitable metric
   (c) Multi-class classification
3. Performance measures
   (a) Confusion matrix
   (b) Precision vs Recall

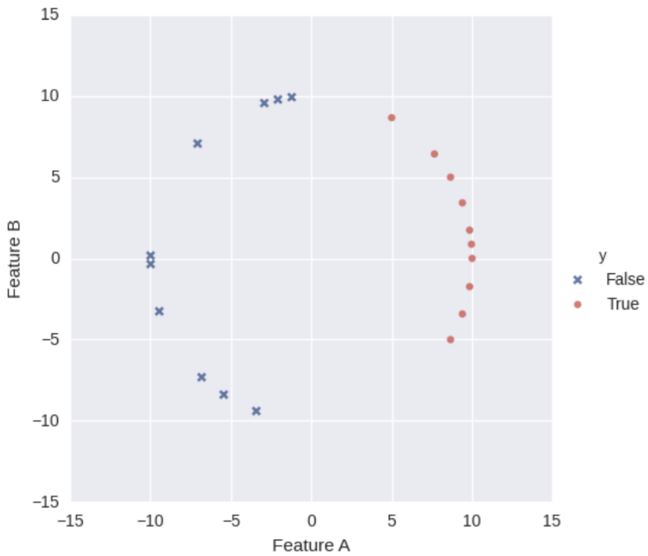
A Linear vs Non-linear Separability

Bondrewd Workshop is a company that produces cute fluffy bunnies through experimentation and toy bunnies. Quality control for the bunnies is done manually. A group of engineers decide whether a bunny is ready to be released 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 a student in CS2109S, Bondrewd the CEO has approached you and asked you to automate the decision making process.

1. Define a set of features that will perfectly classify whether or not a bunny can be released. Here are some examples of sets of features:
   (a) (A, B)
   (b) (A)
   (c) (B)
2. 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 set of features that will perfectly classify whether or not a bunny can be released.

B Loss Function of Logistic Regression

In lecture, we discussed about Logistic Regression model which has the following hypothesis:

$$h_w(x) = \frac{1}{1 + e^{-w^T x}} \tag{1}$$

$h_w(x)$ could be interpreted as a probability p assigned by the model such that $y = 1$. The probability of $y = 0$ is therefore $1 - p$.

1. Write down the probability p as a function of x and calculate the derivative of $\log(p)$ with respect to each weight w_i.
2. Write down the probability $1 - p$ as a function of x and calculate the derivative of $\log(1 - p)$ with respect to each weight w_i.
3. Using results from Questions B(1) and B(2), derive $\frac{\partial L}{\partial w_i}$, where L is the loss function of logistic regression model.

$$L = -y \log(h_w(x)) - (1 - y) \log(1 - h_w(x)) \tag{2}$$

C Precision, recall, and F1 score

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 a 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, the model M outputs label 1 if $M(x)$ is greater than or equal to the threshold, otherwise the model outputs 0.

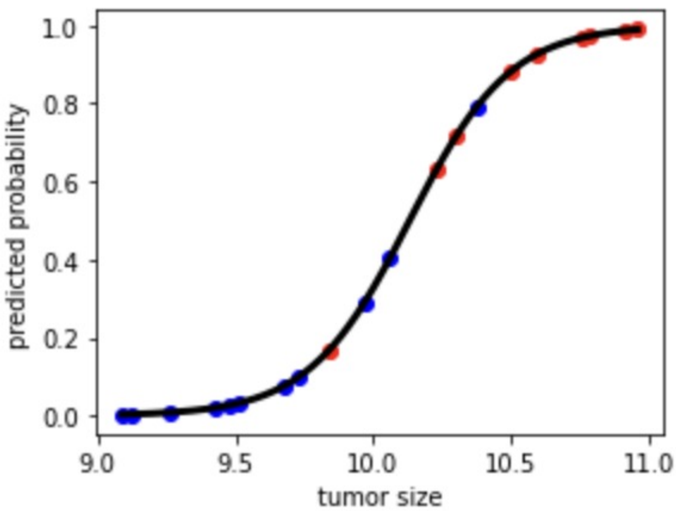


Figure 1: Model probability output and tumor size

1. For the threshold, $p = 0.5$, tabulate the confusion matrix.
2. 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* = $\frac{2TP}{2TP+FP+FN}$
3. **(Optional)** In this question's case for detecting tumours, should we maximize precision or recall? Explain the reason for your choice.
4. **(Optional)** Suppose now we want to detect plagiarism instead, should we maximize precision or recall? Explain the reason for your choice.
5. **Bonus:** We are helping banks with credit card fraud detection in PS3. In such a case, should we maximize precision or recall? Explain the reason for your choice.

D 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 three Logistic Regression models to solve this problem. After training on the training dataset, you get the following parameters:

$$\begin{aligned} w_{cat} &= [4.2, -0.01, -0.12]^T \\ w_{horse} &= [-20, -0.08, 35]^T \\ w_{elephant} &= [-1250, 0.82, 0.9]^T \end{aligned}$$

1. 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

Hint: The hypothesis of Logistic Regression has the following formula

$$h_w(x) = \frac{1}{1 + e^{-w^T x}}$$

2. What if we want to extend the classification task to classify other animals? Can we train a new model while keeping the weights of the previous models?

E Evaluating Logistic Regression

1. Which of the following evaluation metrics are the least appropriate when comparing a logistic regression model's output with the target label? Explain your answer.

   (a) Accuracy

   (b) Binary Cross Entropy Loss
   Binary Cross Entropy Loss is commonly used as an evaluation metric for logistic regression. It is also used as the cost function for logistic regression:

   $$L = -ylogh_w(x) - (1-y)log(1-h_w(x))$$

   Simply put, it is the negative log of the predicted probabilities.

   (c) Mean Squared Error
   Mean Squared Error (MSE) is commonly used as an evaluation metric for linear regression, but can be used for logistic regression as well. Here, it is defined as the squared difference between the predicted probability and the target label (0 if negative and 1 if positive, or vice versa) :

   $$MSE = \frac{1}{2}(y - h_w(x))^2$$

2. **Bonus:** What about Mean Absolute Error (MAE)? *Similar to MSE, Mean Absolute Error (MAE) is defined here as:*

   $$MAE = \frac{1}{2}|y - h_w(x)|$$