National University of Singapore
School of Computing
CS3244: Machine Learning
Tutorial 05

Evaluation Metrics

1. **Precision, Recall, and F_1** For a given binary classification problem, a machine learning model(M) outputs a continuous score for every input sample. Table 1 shows the results for 10 samples from the model. The actual labels can be either 1(positive label) or 0 (negative label). The model output makes the final classification decision. If a threshold, p is given, the decision is made as follows.

$$label(\mathbf{x}) = [M(\mathbf{x}) \geq p)] = \begin{cases} 1 \\ 0 \end{cases}$$

Sample - $\mathbf{x}$	Model output - $M(\mathbf{x})$	Actual label ($\mathbf{y}$)
$\mathbf{x^1}$	0.435	0
$\mathbf{x^2}$	1.257	1
$\mathbf{x^3}$	2.839	1
$\mathbf{x^4}$	4.200	0
$\mathbf{x^5}$	7.432	0
$\mathbf{x^6}$	10.237	1
$\mathbf{x^7}$	12.000	1
$\mathbf{x^8}$	14.839	0
$\mathbf{x^9}$	24.207	1
$\mathbf{x^{10}}$	77.927	1

Table 1: Data information and model outputs

   (a) For threshold, $p = 10$, find precision, recall and F_1 score.

   (b) The number of samples is increased to m. All the model predictions($M(\mathbf{x})$) are given for m samples. Here, all the model predictions are distinct. We want to use all the model outputs as thresholds to find the best threshold for the model. Propose an optimal way to find the best threshold. Comment on the running time.

   (c) We have a new set of thresholds with size $q(> m)$ instead of the model predictions. What is the new running time to find the best threshold from m thresholds?

2. **Micro- and Macro-Averaging**

   (For this question, you can try coding out the formulas and verify that the answers match!)

   We have a classifier trained to predict images of cats, dogs, and pigs. The confusion matrix for the model is given below.

		Actual		
		Dog	Cat	Pig
Predicted	Dog	10	2	1
	Cat	3	13	2
	Pig	3	4	7

(a) Create the confusion matrix for each of the individual classes, dog, cat, and pig.

(b) Calculate the micro-average confusion matrix, accuracy, precision, recall, and F_1 score. What do you notice about the precision, recall and F_1 score? Why is this so? To generate the micro-average confusion matrices, we take the sum of all the individual confusion matrices.

(c) Calculate the macro-average precision and recall.

(d) Consider the following scenario in table 2 where there is a huge class imbalance.

Class	TP	FP
A	9	1
B	100	900
C	9	1
D	9	1

Table 2: Class imbalance Data

Calculate the $Precision_{Micro}$ and $Precision_{Macro}$ and discuss the results.