有許多方式可以評估一 AI 是否有效，以分類器為例，常用的指標便有準確率、精確度、召回率等。但是，某項評估指標數值高，就代表分類器表現好嗎？如何正確選擇評估指標，才能真實反映 AI 效能呢？以下將以常見的二元分類器為例，討論如何藉由資料特性、評估目的及對評估指標的了解，正確選擇合適的評估指標。

There are many ways to evaluate the effectiveness of AI (Artificial Intelligence). Taking a ‘‘classifier ‘‘as an example: accuracy, precision, and recall rate are common indicators in assessing the effectiveness of a classifier. However, does a higher indicator score represent better effectiveness for a classifier? How to choose the right evaluation index that is corresponding to the AI performance accurately? Again, taking the most common example: binary classifiers, it will demostrate how to select an appropriate evaluation index relied on the characteristics of the data, the purpose after evaluation, and understanding for evaluation index.

生活中有許多屬於「二元分類」的問題，例如：應該核准或駁回貸款申請？信用卡發卡與否？是否罹患某一疾病？工廠產品是否有瑕疵？涵蓋範圍之廣，跨及各行各業。AI 二元分類器的效能，便可藉由分類（預測）結果與實際情形的差距來評估。

Many issues in life are related to the binary classification, such as: Should loan applications be approved or rejected? Does the credit card issue or not? Do you suffer from a disease? Are the factory products faulty? Speaking broadly, it covers all walks of life. The effectiveness of the binary classifier can be assessed based on gaps between prediction and the reality of the classification results.

要正確評估分類器的效能，不能只從單方面切入，還必須同時考量分類目的、對資料的了解以及評估指標的特性。舉例來說，若資料分布不均，則應試圖以上述方法改進分類過程；此外，若你的目的是盡量讓陽性樣本被正確地辨識出來，便應該選擇最能突顯陽性樣本分類效果的真陽性率、精確度、以及衡量兩者平衡的 F1 等指標；最後，你也必須清楚哪些指標不會受資料分布不均影響，將其列入候選指標。

To evaluate the performance of classifiers correctly, we should not only view unilaterally but also simultaneously consider the purpose of classification, data interpretation, and the meaning of characteristics after evaluation. For example, if the distribution of data is uneven, the classification process should be improved according to the methods mentioned above. Otherwise, if your goal is to identify the positive samples correctly as much as possible, you should select the most representative indicators that highlight the classification impacts of positive samples representing the true positive rate (TPR), accuracy, and F1-score (measures the balance between TPR and accuracy). To the end, you also need to know which indicators will not be influenced by the uneven distribution then list them as candidate indicators.

相反的，我們也不應單憑單一指標的數值妄下結論、誤判系統優劣，而應參酌前述三項條件。此篇僅簡單闡述評估指標間的相關性以及受資料分布影響的程度，在選擇評估指標前，若能閱讀相關論文或網站，對各指標的用意與限制有更多了解，才能避免誤判分類器效能。

On the contrary, we should not conclude or misjudge the system based on a single index but include the above three conditions. This paper simply is elaborated on the correlation between the evaluation indexes and the influence on the degree distribution. If you can review more relevant papers or websites before selecting the evaluation indexes and understanding the meanings and their limitations, this will avoid misjudging the performance of classifiers.