Metrics trong bài toán Classification

Với hai bài toán classification mà ta đã nghiên cứu là multi-label classification và multi-class classification, ta cần xây dựng bộ các metrics sao cho có thể đánh giá chính xác và khách quan nhất các mô hình machine learning.

1. Accuracy

Accuracy là metrics đơn giản nhất để đánh giá mô hình classification. Accuracy được tính bằng số lượng lời dự đoán đúng chia cho số lượng lời dự đoán đưa ra, bất kể lời dự đoán mà mô hình đưa ra thuộc lớp nào hay điểm dữ liệu đó thuộc lớp nào.

$$\text{accuracy} = \frac{\text{number of true predictions}}{\text{number of predictions}}$$

Accuracy được sử dụng đồng thời trên cả các mô hình giải quyết bài toán multi-label classification và multi-class classification. Tuy nhiên, đối với những bộ dữ liệu mất cân bằng, accuracy bộc lộ điểm yếu khi không thể đánh giá khách quan được chất lượng của mô hình.

2. Confusion matrix

Trước khi đến với các metrics đánh giá giúp giải quyết vấn đề mà accuracy gặp phải, ta đến với một công cụ trực quan hoá kết quả của mô hình classification rất hữu ích, đó là confusion matrix.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Confusion matrix là công cụ giúp trực quan hoá cả lời dự đoán của mô hình và label đúng của điểm dữ liệu đó. Đối với bài toán multi-label classification, confusion matrix là một ma trận có kích thước 2×2, trong đó, 2 cột thể hiện giá trị label đúng của các điểm dữ liệu, 2 hàng thể hiện giá trị mà mô hình dự đoán.

Từ đó, confusion matrix tạo ra 4 giá trị: True Positive (TP), False Positive (FP), True Negative (TN) và False Negative (FN).

- TP là số lượng các điểm dữ liệu mà mô hình dự đoán ĐÚNG là lớp positive tương ứng với label của chúng là postive.

- TN là số lượng các điểm dữ liệu mà mô hình dự đoán ĐÚNG là lớp negative tương ứng với label của chúng là negative.
- FP là số lượng các điểm dữ liệu mà mô hình dự đoán SAI là lớp positive nhưng với label của chúng là negative.
- FN là số lượng các điểm dữ liệu mà mô hình dự đoán SAI là lớp negative nhưng với label của chúng là postive.

Với các giá trị như vậy, hiển nhiên, ta luôn mong muốn hai giá trị TP và TN lớn và hai giá trị FP và FN nhỏ.

Đến đây, ta sẽ xét một ví dụ để nêu rõ được điểm yếu của accuracy, từ đó, nghiên cứu các chỉ số metrics mới.

Ví dụ với bài toán phân lớp bệnh nhân bị ung thư, cụ thể lớp positive (1) là có bị ung thư và lớp negative (0) là không bị ung thư. Dựa vào thực tế, số lượng bệnh nhân ung thư sẽ ít hơn nhiều so với số lượng bệnh nhân không bị ung thư, do đó, bộ dữ liệu để giải bài toán này sẽ có xu hướng chênh lệch số lượng điển dữ liệu giữa các lớp rất lớn.

Ta giả sử, bộ dữ liệu validation của ta có 100 điểm dữ liệu, trong đó có 90 điểm dữ liệu là negative (không bị ung thư) và 10 điểm dữ liệu là positive (có bị ung thư). Trong trường hợp mô hình dự đoán tất cả các bệnh nhân đều khoẻ mạnh, tức là tất cả các điểm dữ liệu đều được dự đoán là negative, ta có accuracy = 90 / 100 = 90%, một chỉ số accuracy tốt. Tuy nhiên, thực tế, mô hình này không có khả năng sử dụng trong thực tế khi tất cả các bệnh nhân cần phải được chuẩn đoán là ung thư (có label là positive), đều đã bị dự đoán nhầm là khoẻ mạnh (negative). Điều này có nghĩa là mô hình này thực sự rất rất rất kém.

Do đó, ta cần các chỉ số đánh giá khác để có thể đánh giá được khách quan mô hình này đối với bộ dữ liệu này.

2.1. Precision

Ta có precision là chỉ số đầu tiên giúp ta giải quyết vấn đề của accuracy.

Precision được tính bằng việc lấy số lượng dự đoán ĐÚNG positive (TP) của mô hình chia cho TỔNG số dự đoán positive của mô hình (TP + FP).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Với ví dụ nói trên, ta có precision được tính bằng việc lấy số lượng dự đoán đúng bệnh nhân ung thư chia cho tổng số lượng lời dự đoán bệnh nhân ung thư của mô hình. Khi xem xét đến precision, với trường hợp mô hình đoán tất cả các bệnh nhân là khoẻ mạnh, precision = 0 / 0 = nan. Khi đó, ta ngay lập tức có thể nhận ra được vấn đề của mô hình thông qua chỉ số precision.

Tuy nhiên, vậy là chưa đủ, nếu mô hình chỉ đưa là rất ít lời dự đoán positive, hay cụ thể ví dụ mô hình chỉ đưa ra đúng 1 lời dự đoán bệnh nhân ung thư và nó đúng. Lúc này, precision = 1 / 1 = 100%, một chỉ số precision tốt, nhưng vẫn còn 9 bệnh nhân ung thư nữa không được dự đoán ra mà mô hình lại không làm được.

Do đó, ta cần một chỉ số đánh giá khác để bổ trợ cho precision trong trường hợp này.

2.2. Recall

Ta có recall là chỉ số giúp bổ trợ cho precision giúp đánh giá mô hình khách quan hơn.

Recall được tính bằng việc lấy số lượng dự đoán ĐÚNG positive (TP) của mô hình chia cho TỔNG số điểm dữ liệu thực sự là postive (TP + FN).

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Với ví dụ nói trên, ta có recall được tính bằng việc lấy số lượng dự đoán đúng bệnh nhân ung thư chia cho tổng số lượng bệnh nhân thật sự bị ung thư. Từ đó, nếu mô hình chỉ đưa ra đúng 1 lời dự đoán bệnh nhân ung thư và nó đúng và precision = 100% thì recall lại rất thấp. Cụ thể, chỉ có 1 bệnh nhân bị ung thư được dự đoán ra, trong khi 9 bệnh nhân còn lại thì không, recall = 1 / 10 = 10%.

Với việc quan sát thêm recall, ta có thể dễ dàng đánh giá được chính xác chất lượng của mô hình machine learning.

Tuy nhiên, việc quan sát đồng thời cả precision và recall đôi lúc gây ra khó khăn, câu hỏi đặt ra là ta sẽ chọn mô hình có chỉ số precision tốt hay mô hình có chỉ số recall tốt?

2.3. F1 score - F score

Ta có $\mathrm{F1}$ score là chỉ số giúp kết hợp được precision và recall.

$\mathrm{F1}$ được tính bằng sự kết hợp của cả giá trị precision và giá trị recall, do đó, $\mathrm{F1}$ chỉ cao khi cả precision và recall đều cao, còn bất kỳ một trong hai chỉ số thấp thì $\mathrm{F1}$ sẽ thấp.

$$\mathrm{F1} = \frac{2 * \mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Tuy nhiên, trong một số trường hợp, mặc dù ta muốn quan sát đồng thời cả precision và recall, nhưng ta lại ưu tiên precision hơn một chút hoặc ưu tiên recall hơn một chút. Ta có thể sử dụng dạng khái quát của $\mathrm{F1}$ là F_β.

$$\mathrm{F}_\beta = \frac{(1 + \beta^2) * \mathrm{precision} * \mathrm{recall}}{\beta^2 * \mathrm{precision} + \mathrm{recall}}$$

Trong đó, β là giá trị do ta lựa chọn nhằm cân đối giữa việc ưu tiên precision hay ưu tiên recall.

- Với trường hợp ta muốn ưu tiên precision, ta lựa chọn $0 < \beta < 1$. β càng nhỏ, càng ưu tiên precision.
- Với trường hợp ta muốn ưu tiên recall, ta lựa chọn $1 < \beta < \infty$. β càng lớn, càng ưu tiên recall.
- Với trường hợp cân bằng, ta chọn $\beta = 1$, ta có chỉ số $\mathrm{F1}$.

2.4. Specificity

Một chỉ số tương tự như recall, nhưng hoạt động với lớp negative, đó là specificity. Tuy nhiên, chỉ số này ít được sử dụng trong thực tế.

$$\mathrm{specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

2.5. Confusion matrix đối với bài toán multi-class classification