BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

SỰ TƯƠNG ĐƯƠNG CỦA CÁC THƯỚC ĐO ĐỘ CHÍNH XÁC

Giảng viên: Lê Bích Phượng

Bộ môn: Toán

Hà Nội, 6/2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT

BÁO CÁO HỌC THUẬT

SỰ TƯƠNG ĐƯƠNG CỦA CÁC THƯỚC ĐO ĐỘ CHÍNH XÁC

Người thực hiện **Xác nhận của Bộ môn Toán**

Lê Bích Phượng **PGS.TS Nguyễn Trường Thanh**

Hà Nội, 6/2022

SỰ TƯƠNG ĐƯƠNG CỦA CÁC THƯỚC ĐO ĐỘ CHÍNH XÁC

Đường cong ROC và các thước đo độ chính xác của các máy nhị phân

Trong chương này, chúng tôi kí hiệu Ω là một không gian đầu vào với một độ đo xác suất P_Ω trên đó, và

$$Y : \Omega \to \{0, 1\} \tag{0.1}$$

là một bài toán phân lớp nhị phân trên Ω. Ví dụ như Ω là tập hợp dân số của một địa phương nào đó, Y là một ánh xạ đi từ Ω vào tập gồm hai phần tử $\{0, 1\}$ người chẩn đoán bị mắc COVID ứng với giá trị 1 , còn người chẩn đoán không bị mắc COVID ứng với giá trị 0. Y thường được gọi là **sự thật cơ bản**

Chúng ta muốn xây dựng một *máy nhị phân*:

$$M : \Omega \to [0, 1] \tag{0.2}$$

(là một kiểm tra, toàn bộ giá trị thuộc vào đoạn [0,1]) để dự đoán giá trị của Y. Chúng ta đưa ra một mức: $\sigma \in [0, 1]$, với mỗi phần tử $x \in \Omega$, ta đặt

$$Y_\sigma(x) = 1 \text{ nếu } M(x) \geq \sigma \text{ và } Y_\sigma(x) = 0 \text{ nếu } M(x) < \sigma \tag{0.3}$$

Hiệu suất (tức là độ chính xác) của dự đoán Y_σ đối với sự thật cơ bản Y có thể được đo lường bằng hai chỉ số hiệu suất cơ bản, được gọi là

sensitivity (= tỷ lệ *dương tính thật*, có chỗ gọi là độ nhạy) $TP(\sigma)$ và **specificity** (= tỷ lệ *âm tính thật*, có chỗ gọi là độ đặc hiệu) $TN(\sigma)$,

chúng được xác định bởi các công thức sau:

$$TP(\sigma) = P(Y_\sigma = 1 | Y = 1) = \frac{P_\Omega(M(x) \geq \sigma\ ,\ Y(x) = 1)}{P_\Omega(Y(x) = 1)}, \tag{0.4}$$

$$TN(\sigma) = P(Y_\sigma = 0 | Y = 0) = \frac{P_\Omega(M(x) < \sigma\ ,\ Y(x) = 0)}{P_\Omega(Y(x) = 0)}. \tag{0.5}$$

Đường cong ROC là tập hợp tất cả các điểm, được xác định như sau:

$ROC = \{ROC(\sigma) = (1 - TN(\sigma), TP(\sigma)), \text{trong đó } \sigma \in [0, 1]\}$

còn được gọi là đường cong đặc tính hoạt động của máy thu (the receiver operating characteristic), của máy M trong tài liệu và được sử dụng trong nhiều lĩnh vực khác nhau, xem trong [**?**, **?**, **?**, **?**, **?**, **?**].

Số $FP(\sigma) = 1 - TN(\sigma)$ được gọi là tỷ lệ *dương tính giả* tại mức σ, có chỗ lại thấy gọi là "xác suất mắc sai lầm loại 2".

Đường cong ROC "đi xuống" từ điểm $ROC(0) = (1, 1)$ đến điểm $ROC(1) = (0, 0)$ trong một hình vuông đơn vị, và đường cong càng cao thì máy càng chính xác.

Ta nhắc lại rằng, AUC là diện tích phần nằm phía dưới của đường cong ROC trong hình vuông đơn vị, và nó cũng là một thước đo phổ biến cho độ chính xác của máy. Xem hình **??** .

Một thước đo phổ biến khác cho độ chính xác của máy là: **độ chính xác có trọng tối đa** được kí hiệu MWA (xem trong, [?]) : đưa ra một trọng $w \in [0,1]$ (được định nghĩa bằng tỷ lệ giữa giá trị âm tính giả và giá trị dương tính giả), ta đặt:

$$MWA(\sigma) = w \cdot TP(\sigma) + (1-w) \cdot TN(\sigma) \tag{0.6}$$

$$MWA(\sigma) = w \cdot TP(\sigma) - (1-w) \cdot FP(\sigma) + (1-w), \tag{0.7}$$

$$MWA = \max_{\sigma \in [0,1]} MWA(\sigma). \tag{0.8}$$

Trong trường hợp riêng, khi $w = 0.5$, thì MWA trở thành **độ chính xác cân bằng tối đa** MBA (**maximal balanced accuracy**):

$$MBA = \max_{\sigma \in [0,1]} BA(\sigma), \text{ trong đó } BA(\sigma) = \frac{TP(\sigma) + TN(\sigma)}{2} \tag{0.9}$$

Một cách đơn giản, ta thấy $0 \leq AUC \leq 1$ do diện tích của hình luôn không âm, và diện tích của phần nằm dưới đường cong ROC luôn không lớn hơn diện tích hình vuông có cạnh bằng 1.

Với mọi $\sigma \in [0,1]$ thì $TP(\sigma)$ và $TN(\sigma) \in [0,1]$ do đó $BA(\sigma) \in [0,1]$, từ đó ta có $0 \leq MBA \leq 1$.

Thêm vào đó, với mọi $\sigma \in [0,1]$ thì $0 \leq w \cdot TP(\sigma) + (1-w) \cdot TN(\sigma) \leq 1$. Thật vậy ta có, vế trước của bất đẳng thức là dễ thấy vì $w, TP(\sigma), TN(\sigma) \in [0,1]$, còn vế sau thì:

Nếu $TP(\sigma) > TN(\sigma)$ thì $w \cdot TP(\sigma) + (1-w) \cdot TN(\sigma) = w \cdot TP(\sigma) + TN(\sigma) - w \cdot TN(\sigma) = w \cdot (TP(\sigma) - TN(\sigma)) + TN(\sigma) \leq TP(\sigma) - TN(\sigma) + TN(\sigma) = TP(\sigma) \leq 1$

Nếu $TP(\sigma) < TN(\sigma)$ thì $w \cdot TP(\sigma) + (1-w) \cdot TN(\sigma) = w \cdot TP(\sigma) + TN(\sigma) - w \cdot TN(\sigma) = w \cdot (TP(\sigma) - TN(\sigma)) + TN(\sigma) \leq TN(\sigma) \leq 1$

Nếu $TP(\sigma) = TN(\sigma)$ thì $w \cdot TP(\sigma) + (1-w) \cdot TN(\sigma) = TN(\sigma) \leq 1$

Như vậy, ta có $0 \leq AUC, MWA, MBA \leq 1$ đối với máy bất kỳ. Hơn thế nữa, nếu một trong các đại lượng đó bằng 1 thì có nghĩa là máy hoàn hảo, 100% chính xác.

Xét về mặt Tô-pô và giải tích hàm, thì chúng ta có thể nói rằng AUC, MWA, MBA là ba thước đo độ chính xác khác nhau, nhưng chúng tương đương tô-pô (*topologically equivalent*), theo nghĩa là nếu một trong các thước đo này dần đến 1, thì hai thước đo còn lại cũng dần đến 1, cũng có nghĩa là nếu máy có độ chính xác cao với một trong các thước đo này, thì cũng có độ chính xác cao đối với hai thước đo còn lại. Cụ thể hơn, chúng ta có bất đẳng thức giữa các thước đo như sau:

Mệnh đề 0.1. *Với các kí hiệu như trên:*
(i) Đối với một máy nhị phân M bất kỳ, ta có:

$$1 - 2(1 - MBA)^2 \geq AUC \geq 2MBA - 1. \tag{0.10}$$

Hơn nữa, nếu đường cong ROC của máy M là lồi, thì chúng ta có:

$$AUC \geq MBA. \tag{0.11}$$

(ii) Với mỗi trọng $w \in]0,1[$ bất kỳ, và với mỗi máy M tùy ý, ta có

$$1 - \frac{(1 - MWA)^2}{2w(1-w)} \geq AUC. \tag{0.12}$$

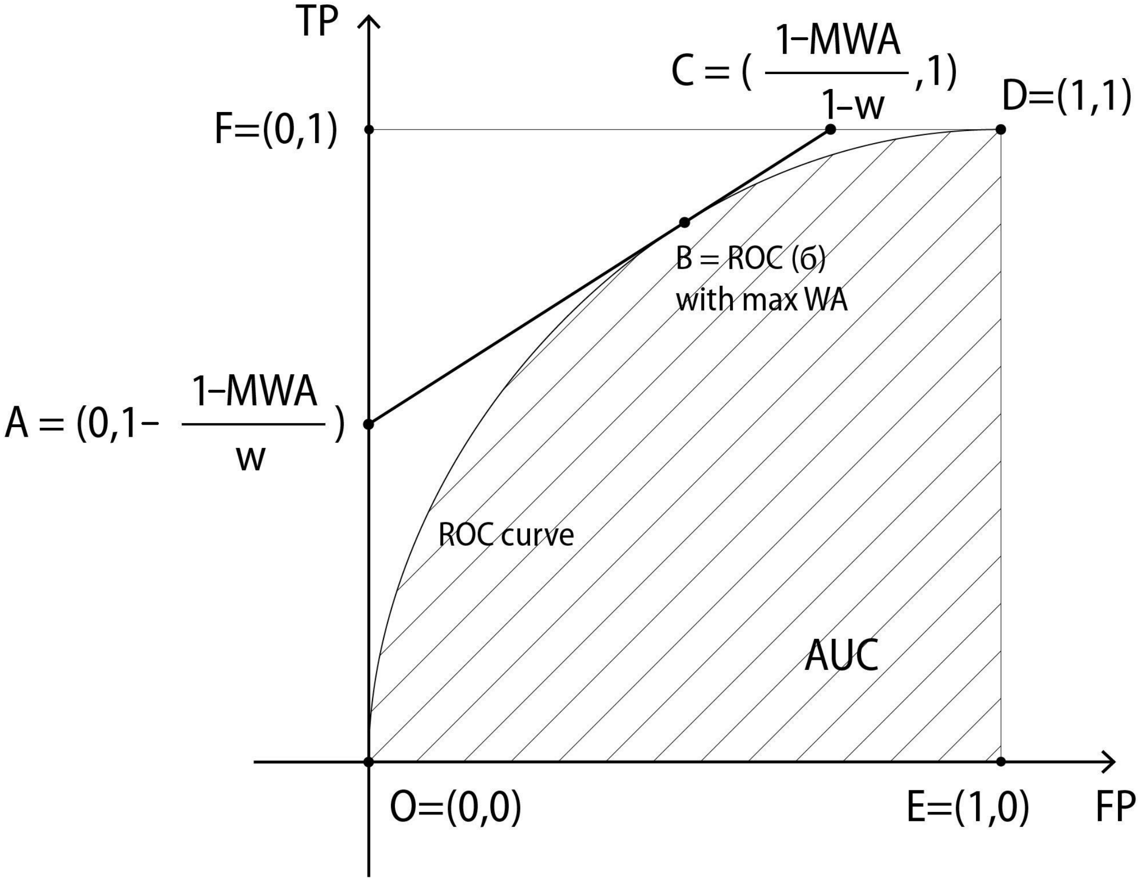
Hơn nữa, nếu đường cong ROC của máy M là lồi thì ta có

$$AUC \geq 1 - \frac{(1 - MWA)}{2\min(w, 1 - w)}. \tag{0.13}$$

Chứng minh. (Xem hình **??**). Chúng ta nhớ lại rằng, số $\sigma \in [0,1]$ là một mức mà máy M đạt được độ chính xác có trọng số cao nhất nếu và chỉ nếu đường thẳng đi qua điểm $ROC(\sigma)$ và bao gồm các điểm $(FP(\sigma)+wt, TP(\sigma)+(1-w)t)$, $t \in \mathbb{R}$ là một đường nằm phía trên của đường cong ROC.

Thật vậy, các đường $\{(FP(\sigma)+wt, TP(\sigma)+(1-w)t),\ t \in \mathbb{R}\}$ với véc-tơ chỉ phương là $(w, 1-w)$ chỉ là các đường của độ chính xác có trọng hằng số.

Nếu tại điểm $B = (FP(\sigma), TP(\sigma))$ của đường cong ROC đạt được độ chính xác có trọng lớn nhất, thì không có điểm nào thuộc đường cong ROC đó có thể nằm phía trên đường dốc tương ứng $(w, 1-w)$, vì nếu nằm phía trên thì có nghĩa là có độ chính xác có trọng cao hơn.



Hình 1: Đường cong ROC và đường tiếp tuyến tại điểm mà trung bình có trọng lớn nhất.

Đường $\ell = \{(FP(\sigma) + wt, TP(\sigma) + (1 - w)t),\ t \in \mathbb{R}\}$, ở đó σ đưa ra độ chính xác có trọng lớn nhất của máy M, cắt biên của hình vuông đơn vị tại hai điểm $A = (0, 1 - \dfrac{1 - MWA}{w})$ và $C = (\dfrac{1 - MWA}{1 - w}, 1)$. Tam giác $\triangle ACF$, trong đó tọa độ của $F = (0,1)$, là tách biệt với phần nằm phía dưới của đường cong ROC, điều này nghĩa là $AUC +$ diện tích$(\triangle ACF) \leq 1$. Từ diện tích $(\triangle ACF) = \dfrac{FA \cdot FC}{2} = \dfrac{(1-MWA)^2}{2w(1-w)}$, ta thu được bất đẳng thức

$$AUC \leq 1 - \frac{(1-MWA)^2}{2w(1-w)}.$$

Mặt khác, phần nằm dưới đường cong ROC chứa toàn bộ một hình chữ nhật có các đỉnh là $(FP(\sigma), 0), (FP(\sigma), TP(\sigma), (1, TP(\sigma)), (1, 0)$.

Diện tích của hình chữ nhật đó là $TP(\sigma) \cdot (1 - FP(\sigma)) = TP(\sigma) \cdot TN(\sigma) = TP(\sigma) + TN(\sigma) - 1 + (1 - TP(\sigma))(1 - TP(\sigma)) \geq TP(\sigma) + TN(\sigma) - 1 = 2BA(\sigma) - 1$ (với mọi σ). Do đó, ta có bất đẳng thức sau

$$AUC \geq 2MBA - 1.$$

Nếu đường cong ROC là lồi, thì phần phía dưới của đường cong chứa tứ giác $OBDE$, ở đó $O = (0, 0)$, $B = (FP(\sigma), TP(\sigma)), D = (1, 1), E = (1, 0)$ (với mọi σ).

Diện tích của tứ giác này đúng bằng $BA(\sigma)$, nghĩa là bằng MBA; do đó, ta thu được bất đẳng thức $AUC \geq BA(\sigma)$ với mọi σ, nghĩa là ta có $AUC \geq MBA$.

Cuối cùng, bất đẳng thức $AUC \geq 1 - \dfrac{(1-MWA)}{2\min(w, 1-w)}$ đạt được trong trường hợp đường cong ROC là lồi và trọng w bất kỳ, là hệ quả trực tiếp từ dãy các bất đẳng thức sau

$$AUC \geq dt(OBDE) \geq \min(dt(OADE), dt(OCDE))$$

và các đẳng thức $dt(OCDE) = 1 - \dfrac{(1-MWA)}{2(1-w)}$, $dt(OADE) = 1 - \dfrac{(1-MWA)}{2w}$. □

Ghi chú. Đối với một số bất đẳng thức trong Mệnh đề 0.1, ta giả sử đường cong ROC là lồi. Độ gần lồi của đường cong ROC đã từng được quan sát thực nghiệm trong nhiều chuyên khảo và bài báo trong thời gian dài; (Định lý 3) nói về độ lồi của đường cong ROC tương đương với các điều kiện hợp lý tự nhiên (rationality) khác trên máy (cụ thể là, "giá trị sigmoid" càng cao thì khả năng sự kiện đúng càng cao). Điều này đã từng được chỉ ra trong [**?**]

Trong phần **??**, chúng tôi chỉ ra rằng máy xác suất thực (theo nghĩa là máy tự nhiên nhất), cũng là máy chính xác nhất và đường cong ROC của nó tự động lồi.

Kết quả của chúng tôi giúp giải thích tại sao hầu hết các đường cong ROC gặp trong thực tế gần như lồi (bởi vì máy móc phải gần như tối ưu theo nghĩa nào đó nếu quy trình học máy tạo ra chúng hiệu quả).

Ghi chú. Nếu chúng ta tham số hóa một hàm sigmoid Σ bằng một ánh xạ tăng tùy ý $f : [0, 1] \to [0, 1]$, thì chúng ta thu được một hàm sigmoid mới $\Sigma' = f \circ \Sigma$ mà có đường cong ROC giống với đường cong ROC của Σ, qua một phép tham số hóa bởi f. Nghĩa là, $ROC_{\Sigma}(\sigma) = ROC_{\Sigma'}(f(\sigma))$ với mọi $\sigma \in [0, 1]$.

Đặc biệt, phép tham số hóa làm thay đổi các giá trị sigmoid nhưng không làm thay đổi các giá trị của độ đo hiệu suất $AUC, MBA,$ và MWA của hệ thống. □

Ghi chú. Một số tác giả cũng sử dụng trung bình hình học, (xem trong [**?**]) (the geometric mean) $GM = \sqrt{TP.TN}$ của "độ nhạy"(sensitivity) (TP) và "độ đặc hiệu"(specificity) (TN) như một thước đo độ chính xác cho bài toán dự báo nhị phân. Các bất đẳng thức số học tự nhiên

$(a+b)^2/4 \geq ab \geq (a+b)^2/2 - 1$ (đối với mọi số dương $a, b \leq 1$) , liên hệ giữa độ chính xác theo trung bình hình học và độ chính cân bằng $BA = (TP + TN)/2$. Đặc biệt, nó nghĩa là độ chính xác trung bình hình học (cực đại) cũng là một thước đo độ chính xác tốt như MBA và AUC, tức là chúng tương đương về mặt Tô-pô. □

Ghi chú. Cho một phân phối xác suất ban đầu $P = P_\Omega$ trên không gian Ω, mà nó không cân bằng theo nghĩa $P(Y=0) \neq P(Y=1)$ (không cân bằng giữa các trường hợp dương tính và các trường hợp âm tính), chúng ta có thể thay đổi nó thành một phân phối xác suất mới cân bằng, $\hat{P}$, được định nghĩa bởi công thức sau:

$$\hat{P}(A) = \frac{1}{2}\left[\frac{P(A \cap \{Y=0\})}{P(Y=0)} + \frac{P(A \cap \{Y=1\})}{P(Y=1)}\right].$$

Dễ dàng chứng minh được rằng, đường cong ROC được tham số hóa (với một máy đã cho $M = \Sigma \circ \phi : \Omega \to [0,1]$) đối với $\hat{P}$ trùng với đường cong ROC đối với P. Thật vậy đối với $\sigma \in [0,1]$ bất kỳ, mức dương đúng tại σ là $TP(\sigma) = \dfrac{P(\Sigma > \sigma, Y=1)}{P(Y=1)} = 2\hat{P}(\Sigma > \sigma, Y=1) = \dfrac{\hat{P}(\Sigma > \sigma, Y=1)}{\hat{P}(Y=1)} = \widehat{TP}(\sigma)$, và tương tự đối với $TN(\sigma)$. Vì vậy,trong nghiên cứu về độ chính xác, không mất tính tổng quát, người ta có thể giả sử rằng phân phối xác suất là cân bằng đối với Y theonghĩa là $P(Y=0) = P(Y=1) = 0.5$. □

Phương pháp biểu quyết để cải thiện độ chính xác

Bây giờ chúng ta hãy thảo luận về chủ đề thứ hai. Lúc đầu, giả sử rằng, chúng ta có một nhóm n **các chuyên gia độc lập** hoàn toàn, mỗi chuyên gia đều có một độ chính xác (cost-wise) tỉ lệ chính xác $p > 1/2$ và **tỉ lệ lỗi** $q = 1 - p < 1/2$.

Phân bố về số lượng các chuyên gia có dự đoán đúng là phân bố nhị thức $P(k) = C_n^k p^k q^{n-k}$ $(0 \leq k \leq n)$. Theo định lý giới hạn trung tâm, khi n đủ lớn thì phân phối nhị thức xấp xỉ bằng **phân phối chuẩn** với trung bình là np và phương sai là npq. Do vậy xác suất S_n có ít nhất $n/2$ dự đoán đúng (*trong số n các chuyên gia*) thì xấp xỉ

$$S_n \cong \Phi(\frac{\sqrt{n}(p-1/2)}{\sqrt{pq}}) \quad \text{where} \quad \Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-x^2/2}dx$$

S_n cũng là điểm số chính xác dự kiến của quyết định tập thể bằng cách biểu quyết. (*1 chuyên gia = 1 phiếu bầu*). Cụ thể, $\lim_{n\to\infty} S_n = 1$.

Tuy nhiên trong thực tế, **chúng ta không thể có các chuyên gia thực sự độc lập**. Họ có “những điểm mù phổ biến” (**common blind spots**). *Một ví dụ hư cấu:một người sao Hỏa cải trang hoàn hảo thành một con người trên Trái đất. Không ai có thể phát hiện ra anh ta.*

Để đơn giản, chúng tôi xem xét một mô hình chỉ có hai loại điểm mù phổ biến, đó là “ nửa mù” **half-blind** (các quyết định ngẫu nhiên), và “mù hoàn toàn” (**completely blind**) (*mọi người bị tẩy não và tin vào những lời dối trá*): Tập hợp tổng số được chia thành 3 phần như sau:

$$\Omega = \Omega_{\text{blind}} \cup \Omega_{\text{hb}} \cup \Omega_{\text{l}}$$

Trên Ω_{blind} mọi người đều sai, trên Ω_{hb} các quyết định kiểu như ngẫu nhiên, và trên Ω_{l} (tập có thể học được) các chuyên gia là độc lập và có tỷ lệ sai sót q (như trên đã nói đến). Tiếp theo, chúng tôi có công thức gần đúng cho điểm số (tỉ số) chính xác của quyết định tập thể bằng cách biểu quyết:

$$S_n \cong \Phi(\frac{\sqrt{n}(p-1/2)}{\sqrt{pq}})(1-\mathbf{P}_{\text{blind}}-\mathbf{P}_{\text{hb}})+\frac{1}{2}\mathbf{P}_{\text{hb}}$$

Cụ thể, $\lim_{n\to\infty} S_n = 1-\mathbf{P}_{\text{blind}}-\frac{1}{2}\mathbf{P}_{\text{hb}}$