Học máy và Thị giác máy tính

Đinh Viết Sang

December 20, 2022

Mục lục

Chương 1

Bias và Variance

1.1 Sai số kiểm tra kỳ vọng (expected test error)

Xét bài toán học có giám sát với tập học $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ được lấy mẫu độc lập và tuân theo phân phối giống nhau (*i.i.d.*) từ phân phối dữ liệu $\Omega(\mathcal{X}, \mathcal{Y})$, trong đó $\mathcal{X}$ là không gian dữ liệu và $\mathcal{Y}$ là không gian nhãn.

Giả sử $f : \mathcal{X} \to \mathcal{Y}$ là ánh xạ đúng biểu diễn quan hệ từ một dữ liệu đầu vào $x \in \mathcal{X}$ sang một nhãn đầu ra $y \in \mathcal{Y}$:

$$y = f(\mathbf{x}) + \epsilon \tag{1.1}$$

trong đó $\epsilon = \mathcal{N}(0, \beta^2)$ là nhiễu Gauss với kỳ vọng bằng 0.

Thông thường nhãn y không thể thu thập chính xác được do sai số khách quan trong quá trình thu thập dữ liệu liệu. Chẳng hạn, khi xét bài toán hồi quy cân nặng, chiều cao, các chỉ số sinh hoá, giá trị nhãn y có thể bị sai lệch do thiết bị đo đạc và thiết bị xét nghiệm... Trong bài toán hồi quy giá nhà, cùng một $\mathbf{x}$ (cùng diện tích, cùng tầng, cùng số phòng...) nhưng có thể có nhiều mức giá y khác nhau tuỳ thời điểm và các yếu tố khách quan khác. Kỳ vọng của nhãn y thu được ứng với dữ liệu đầu vào x chính là $f(\mathbf{x})$:

$$f(\mathbf{x}) = \mathbb{E}\big[y|\mathbf{x}\big] = \int_y y Pr(y|\mathbf{x}) \partial y \tag{1.2}$$

Giả sử ta xấp xỉ $f(\mathbf{x})$ bằng giả thuyết $h(\mathbf{x}, \mathbf{D})$ là một mô hình có tham số được huấn luyện trên tập học D. Quá trình huấn luyện h thường được quy về bài toán cực tiểu hóa rủi ro thực nghiệm (emprirical risk hoặc training error):

$$\hat{R}_D(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(h(\mathbf{x}_i, D), y_i) \tag{1.3}$$

trong đó $\mathcal{L}(h(\mathbf{x}_i, D), y_i)$ là hàm chi phí được định nghĩa trước.

Tuy nhiên, hiệu năng của mô hình phải được đánh giá trên toàn bộ phân phối dữ liệu $\mathcal{D}$, bao gồm cả những dữ liệu chưa nhìn thấy (dữ liệu test), sử dụng độ đo gọi là rủi ro kỳ vọng (expected risk hoặc test error):

$$R_D(h) \triangleq \mathbb{E}_{(\mathbf{x},y)\sim\Omega}\big[\mathcal{L}(h(\mathbf{x}, D), y)\big] = \int_{\mathbf{x}} \int_y \mathcal{L}(h(\mathbf{x}, D), y) Pr(\mathbf{x}, y) \partial y \partial \mathbf{x} \tag{1.4}$$

Giả thuyết h khi huấn luyện trên các tập học D khác nhau sẽ thu được các mô hình $h(\mathbf{x}, D)$ khác nhau ứng với các bộ tham số khác nhau. Do đó khi đánh giá giả thuyết h, ta

quan tâm tới sai số kiểm tra kỳ vọng (expected test error) ứng với mọi trường hợp khác nhau của tập học D:

$$\mathbb{E}_D\big[R_D(h)\big] = \int_D R_D(h)Pr(D)\partial D = \int_D \int_{\mathbf{x}} \int_y \mathcal{L}(h(\mathbf{x}, D), y)Pr(\mathbf{x}, y)Pr(D)\partial y \partial \mathbf{x} \partial D$$
$$= \mathbb{E}_{\mathbf{x},y,D}\big[\mathcal{L}(h(\mathbf{x}, D), y)\big] \tag{1.5}$$

Mô hình kỳ vọng $\bar{h}(\mathbf{x})$ là trung bình tất cả các mô hình $h(\mathbf{x}, D)$ được huấn luyện trên tất cả các tập học D khác nhau:

$$\bar{h}(\mathbf{x}) = \mathbb{E}_D\big[h(\mathbf{x}, D)\big] = \int_D h(\mathbf{x}, D)Pr(D)\partial D \tag{1.6}$$

1.2 Phân rã bias và variance

Xét trường hợp hồi quy tuyến tính với hàm mất mát bình phương $\mathcal{L}(h(\mathbf{x}, D), y) = (h(\mathbf{x}, D) - y)^2$. Khi đó sai số kiểm tra kỳ vọng có thể khai triển như sau [Bishop and Nasrabadi, 2006]:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y,D}\big[\mathcal{L}(h(\mathbf{x}, D), y)\big] &= \mathbb{E}_{\mathbf{x},y,D}\big[\big(h(\mathbf{x}, D) - y\big)^2\big] = \mathbb{E}_{\mathbf{x},y,D}\big[\big((h(\mathbf{x}, D) - \bar{h}(\mathbf{x})) + (\bar{h}(\mathbf{x}) - y)\big)^2\big] \\
&= \mathbb{E}_{\mathbf{x},y,D}\big[\big(h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big)^2\big] + \mathbb{E}_{\mathbf{x},y,D}\big[\big(\bar{h}(\mathbf{x}) - y\big)^2\big] \\
&+ 2\mathbb{E}_{\mathbf{x},y,D}\big[\big(h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big)\big(\bar{h}(\mathbf{x}) - y\big)\big] \\
&= \mathbb{E}_{\mathbf{x},D}\big[\big(h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big)^2\big] + \mathbb{E}_{\mathbf{x},y}\big[\big(\bar{h}(\mathbf{x}) - y\big)^2\big] \\
&+ 2\mathbb{E}_{\mathbf{x},y,D}\big[\big(h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big)\big(\bar{h}(\mathbf{x}) - y\big)\big]
\end{aligned} \tag{1.7}$$

Xét số hạng thứ 3 trong Công thức (1.7), do $\bar{h}(\mathbf{x}) - y$ không phụ thuộc vào D nên có thể đặt làm thừa số chung:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},y,D}\Big[\big(h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big)\big(\bar{h}(\mathbf{x}) - y\big)\Big] = \mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - y\big)\mathbb{E}_D\big[h(\mathbf{x}, D) - \bar{h}(\mathbf{x})\big]\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - y\big)\Big(\mathbb{E}_D\big[h(\mathbf{x}, D)\big] - \mathbb{E}_D\big[\bar{h}(\mathbf{x})\big]\Big)\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - y\big)\Big(\bar{h}(\mathbf{x}) - \bar{h}(\mathbf{x})\Big)\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[0\Big] \\
&= 0
\end{aligned} \tag{1.8}$$

Xét số hạng thứ 2 của Công thức (1.7), ta có:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - y\big)^2\Big] = \mathbb{E}_{\mathbf{x},y}\Big[\big((\bar{h}(\mathbf{x}) - f(\mathbf{x})) + (f(\mathbf{x}) - y)\big)^2\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)^2\Big] + \mathbb{E}_{\mathbf{x},y}\Big[(f(\mathbf{x}) - y)^2\Big] + 2\mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\big(f(\mathbf{x}) - y\big)\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)^2\Big] + \mathbb{E}_{\mathbf{x},y}\Big[(f(\mathbf{x}) - y)^2\Big] + 2\mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\big(f(\mathbf{x}) - y\big)\Big]
\end{aligned} \tag{1.9}$$

Xét số hạng thứ 3 trong Công thức (1.9), do $\bar{h}(\mathbf{x}) - f(\mathbf{x})$ không phụ thuộc vào y nên có thể đặt làm thừa số chung như sau:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},y}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\big(f(\mathbf{x}) - y\big)\Big] = \mathbb{E}_x\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\mathbb{E}_{y|x}\big[f(\mathbf{x}) - y\big]\Big] \\
&= \mathbb{E}_x\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\big(\mathbb{E}_{y|x}\big[f(\mathbf{x})\big] - \mathbb{E}_{y|x}\big[y\big]\big)\Big] \\
&= \mathbb{E}_x\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)\big(f(\mathbf{x}) - f(\mathbf{x})\big)\Big] \\
&= 0
\end{aligned} \tag{1.10}$$

Kết hợp các Công thức (1.7), (1.8), (1.9), (1.10), ta có phân rã cuối cùng:

$$\underbrace{\mathbb{E}_{\mathbf{x},y,D}\big[\big(h(\mathbf{x},D) - y\big)^2\big]}_{expected\ test\ error} = \underbrace{\mathbb{E}_{\mathbf{x},D}\big[\big(h(\mathbf{x},D) - \bar{h}(\mathbf{x})\big)^2\big]}_{variance} + \underbrace{\mathbb{E}_{\mathbf{x}}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)^2\Big]}_{bias^2} + \underbrace{\mathbb{E}_{\mathbf{x},y}\Big[(f(\mathbf{x}) - y)^2\Big]}_{noise} \tag{1.11}$$

Như vậy ta có các khái niệm sau:

Bias của giả thuyết $h(\mathbf{x})$ tại một điểm dữ liệu vào $\mathbf{x}$ thể hiện sự chênh lệch giữa giá trị nhãn lý tưởng $f(\mathbf{x})$ so với kỳ vọng $\bar{h}(\mathbf{x})$ của các phán đoán $h(\mathbf{x}, D)$ do các mô hình được huấn luyện trên các tập học D khác nhau đưa ra.

$$bias\big[h(\mathbf{x})\big] = \bar{h}(\mathbf{x}) - f(\mathbf{x}) = \mathbb{E}_D\big[h(\mathbf{x},D)\big] - f(\mathbf{x}) = \int_D h(\mathbf{x},D)Pr(D)\eth D - f(\mathbf{x}) \tag{1.12}$$

Nếu xét trên toàn bộ không gian dữ liệu $\mathcal{X}$, ta có khái niệm bias bình phương (*squared bias*) của giả thuyết h:

$$bias^2\big[h\big] = \mathbb{E}_{\mathbf{x}}\Big[\big(\bar{h}(\mathbf{x}) - f(\mathbf{x})\big)^2\Big] = \int_{\mathbf{x}}\left(\int_D h(\mathbf{x},D)Pr(D)\eth D - f(\mathbf{x})\right)^2 Pr(\mathbf{x})\eth\mathbf{x} \tag{1.13}$$

Variance của giả thuyết h tại một điểm dữ liệu vào $\mathbf{x}$ thể hiện sự phân tán của các phán đoán khác nhau $h(\mathbf{x}, D)$ do các mô hình được huấn luyện trên các tập học D khác nhau đưa ra.

$$variance\big[h(\mathbf{x})\big] = \mathbb{E}_D\Big[\big(h(\mathbf{x},D) - \bar{h}(\mathbf{x})\big)^2\Big] = \mathbb{E}_D\left[\Big(h(\mathbf{x},D) - \mathbb{E}_D\big[h(\mathbf{x},D)\big]\Big)^2\right] \tag{1.14}$$

Nếu xét trên toàn bộ không gian dữ liệu $\mathcal{X}$, ta có khái niệm variance của giả thuyết h như sau:

$$\begin{aligned}
variance\big[h\big] &= \mathbb{E}_{\mathbf{x},D}\big[\big(h(\mathbf{x},D) - \bar{h}(\mathbf{x})\big)^2\big] \\
&= \int_{\mathbf{x}}\int_D\left(h(\mathbf{x},D) - \int_D h(\mathbf{x},D)Pr(D)\eth D\right)^2 Pr(D)Pr(\mathbf{x})\eth D\eth\mathbf{x}
\end{aligned} \tag{1.15}$$

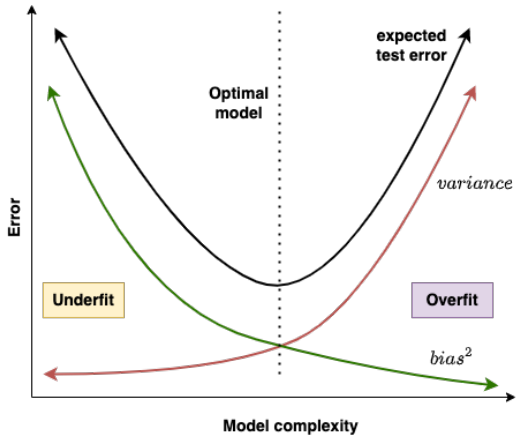
Noise là thành phần nhiễu không thể loại bỏ được (*irreducible error*). Nhiễu này không phụ thuộc vào giả thuyết nên dù mô hình hoá bài toán tốt như thế nào đi nữa cũng không

giảm được nó. Đây là thành phần ta bắt buộc phải "sống chung" do các sai số khách quan khi thu thập dữ liệu.

$$noise = \mathbb{E}_{\mathbf{x},y}\Big[(f(\mathbf{x}) - y)^2\Big] = \int_{\mathbf{x}}\int_{y}(f(\mathbf{x}) - y)^2 Pr(\mathbf{x}, y)\partial y \partial \mathbf{x} \tag{1.16}$$

1.3 Lý thuyết cổ điển về đánh đổi Bias-variance

Trong lý thuyết cổ điển về học máy, bias và variance là hai đại lượng đối nghịch nhau. Khi giả thuyết h có bias thấp thì variance sẽ cao và ngược lại. Người ta gọi đó là đánh đổi (*tradeoff*) bias-variance.



Hình 1.1: Giả thuyết chữ U cổ điển về đánh đổi bias-variance.

Underfitting là hiện tượng khi mô hình khớp dữ liệu học không tốt dẫn đến sai số cao trên cả tập học lẫn tập test. Điều này xảy ra khi mô hình quá đơn giản do có ít tham số, dẫn tới khả năng biểu diễn kém và không đủ khả năng để xấp xỉ ánh xạ đúng $f(\mathbf{x})$. Một mô hình bị underfit thường có bias cao và variance thấp. Nghĩa là dù thay đổi nhiều tập học khác nhau thì sai số vẫn cao ổn định.

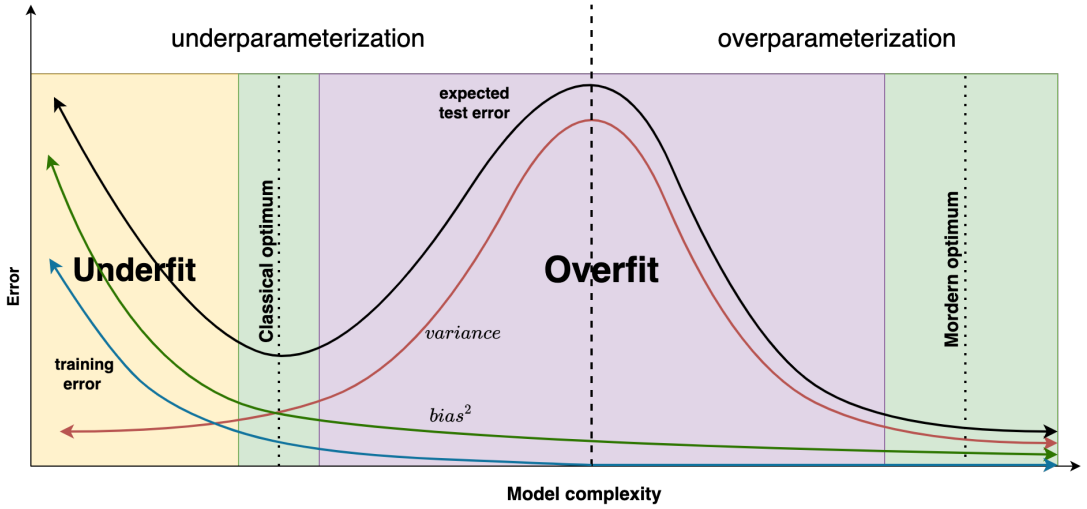
Overfitting là hiện tượng khi mô hình khớp dữ liệu học quá tốt dẫn đến sai số thấp trên tập học nhưng khi đánh giá trên tập test thì sai số lại cao hơn nhiều. Mô hình bị overfit khi khả năng tổng quát hoá (*generalization*) của mô hình kém. Điều này xảy ra khi mô hình quá phức tạp so với dữ liệu do chứa quá nhiều tham số hơn mức cần thiết. Khi đó thay vì xấp xỉ ánh xạ đúng $f(\mathbf{x})$, mô hình ghi nhớ luôn nhãn của các mẫu học trong quá trình huấn luyện mà không biểu diễn tốt quan hệ giữa dữ liệu và nhãn. Mô hình bị overfit thường có bias thấp và variance cao. Nghĩa là phán đoán của mô hình dao động rất lớn khi được huấn luyện trên nhiều tập học khác nhau, nhưng giá trị trung bình các phán đoán lại rất gần với giá trị nhãn đúng.

Giả thuyết chữ U (*U-shape*) trong lý thuyết học máy cổ điển cho rằng khi tăng dần độ phức tạp của mô hình, sai số kiểm tra kỳ vọng sẽ giảm dần khi mô hình chuyển từ trạng thái underfit sang trạng thái tối ưu (*optimal model*) với sai số kiểm tra kỳ vọng thấp nhất. Sau đó nếu tiếp tục tăng độ phức tạp, mô hình sẽ dần mất trạng thái tối ưu và chuyển sang trạng thái overfit. Khi độ phức tạo càng tăng, trạng thái overfit càng nghiêm trọng, nghĩa là sai số kiểm tra kỳ vọng càng cao. Các nghiên cứu gần đây chỉ ra rằng giả thuyết

này chỉ phù hợp khi mô hình ở chế độ thiếu tham số hoá (*underparameterization*). Còn khi mô hình có số tham số lớn hơn nhiều so với kích thước tập học (*overparameterization*) thì giả thuyết hình chữ U như Hình 1.1 không còn đúng nữa.

1.4 Hiện tượng hạ kép (double descent)

Hiện tượng hạ kép (*double descent*) lần đầu tiên được phát hiện bởi Belkin và cộng sự trong [Belkin et al., 2019]. Lý thuyết học máy cổ điển cho rằng mô hình càng lớn càng dễ bị quá khớp và có khả năng tổng quát hóa kém, nhưng Belkin và cộng sự đã chỉ ra rằng các mô hình càng lớn càng tốt hơn. [Belkin et al., 2019] phát hiện ra rằng bức tranh cổ điển hình chữ U về sự đánh đổi bias-variance sẽ bị phá vỡ tại thời điểm khi sai số huấn luyện xấp xỉ 0 mà họ gọi là ngưỡng nội suy (*interpolation threshold*). Trước ngưỡng nội suy, sự sự đánh đổi bias-variance được đảm bảo, và độ phức tạp của mô hình càng tăng sẽ dẫn đến hiện tượng học quá khớp, làm tăng sai số kiểm tra. Tuy nhiên, sau ngưỡng nội suy, họ nhận thấy rằng lỗi kiểm tra bắt đầu giảm xuống khi tiếp tục tăng độ phức tạp của mô hình.



Hình 1.2: Hiện tượng hạ kép trong mạng nơ-ron và nhiều mô hình học máy cổ điển khác: Nửa trái là bức tranh cổ điển về sự đánh đổi bias-variance ở chế độ thiếu tham số hoá; Nửa phải là bức tranh hiện đại khi các mô hình được tham số hóa quá mức.

Trong chế độ nội suy khi mô hình bị tham số hoá quá mức (số tham số mô hình nhiều hơn số mẫu học), Rocks and Mehta [2022] thử nghiệm với hồi quy tuyến tính và mạng nơ-ron hai lớp ẩn đã chỉ ra rằng bias và variance đơn điệu giảm khi tăng độ phức tạp của mô hình (xem Hình 1.2). Khi số tham số đủ lớn, mô hình sẽ thoát khỏi overfitting, chuyển sang trạng thái "**benign overfit**" và cho khả năng tổng quát hoá tốt trên tập kiểm tra.

Tài liệu tham khảo

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

J. W. Rocks and P. Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1):013201, 2022.