Verzerrung und Varianz

M. Gruber

KW 49

Angenommen, wir sind in einer Lernsituation und haben ein Lernmodell. Zur Lernmenge D liefere der Lernalgprithmus die finale Hypothese $g^{(D)}$. Die Auswahl der Lernmenge ist vom Zufall $\mathbf{P}$ gesteuert. Der Erwartungswert der finalen Hypothesen bezüglich aller möglichen Lernmengen ist

$$\bar{g}(x) = \mathbf{E}_D\, g^{(D)}(x). \tag{1}$$

In der Praxis könnten wir diesen Erwarungswert durch die Größe

$$\frac{1}{K} \sum_{1 \le k \le K} g^{(D_k)}(x) \tag{2}$$

schätzen, wenn wir uns Lernmengen $D_1, \ldots, D_K$ verschaffen und K-mal lernen.

Unsere Fehlerfunktion sei die quadratische Abweichung

$$\mathrm{Err}_{\mathrm{out}}(g^{(D)}) = \mathbf{E}_x(g^{(D)}(x) - f(x))^2. \tag{3}$$

Den Ausdruck auf der rechten Seite können wir in drei Summanden zerlegen:

$$\mathbf{E}_x(g^{(D)}(x) - f(x))^2 = \mathbf{E}_x(g^{(D)}(x) - \bar{g}(x))^2 + 2\,\mathbf{E}_x(g^{(D)}(x) - \bar{g}(x))(\bar{g}(x) - f(x)) + \mathbf{E}_x(\bar{g}(x) - f(x))^2.$$

Wir interessieren uns für den Erwarungswert $\mathbf{E}_D\,\mathrm{Err}_{\mathrm{out}}(g^{(D)})$ und wenden den $\mathbf{E}_D$-Operator auf jeden der drei Summanden an.

1. $\mathbf{E}_D\,\mathbf{E}_x(g^{(D)}(x) - \bar{g}(x))^2 = \mathbf{E}_x\,\mathbf{E}_D(g^{(D)}(x) - \bar{g}(x))^2 = \mathbf{E}_x\,\mathrm{var}(x)$. Die Varianz, die wir hier sehen, ist die Varianz bezüglich D. Für spätere Zwecke vereinbaren wir die Bezeichnung $\mathrm{var} = \mathbf{E}_x\,\mathrm{var}(x)$.

2. $\mathbf{E}_D\,\mathbf{E}_x(g^{(D)}(x) - \bar{g}(x))(\bar{g}(x) - f(x)) = 0$. Um dies einzusehen, tausche man zuerst die Erwartungswertoperatoren. Dann ziehe man den Faktor $(\bar{g}(x) - f(x))$ vor den $\mathbf{E}_D$-Operator und betrachte, was jetzt noch für den $\mathbf{E}_D$-Operator übrig ist.

3. Der letzte Summand hängt nicht von D ab. Er stellt die sog. Verzerrung (*bias*) dar, die wir als $\mathrm{bias}(x) = (\bar{g}(x) - f(x))^2$ und $\mathrm{bias} = \mathbf{E}_x(\bar{g}(x) - f(x))^2$ definieren.

Somit ist $\mathbf{E}_D \mathbf{E}_x(\bar{g}(x) - f(x))^2 = \mathbf{E}_x \operatorname{bias}(x) = \operatorname{bias}$.

Insgesamt ist also

$$\mathbf{E}_D \operatorname{Err}_{\text{out}}(g^{(D)}) = \mathbf{E}_x(\operatorname{bias}(x) + \operatorname{var}(x)) = \operatorname{bias} + \operatorname{var}. \tag{4}$$

Dies ist die sogenannte *Verzerrungs-Varianz-Zerlegung (bias-variance decomposition)*. Wenn man die Wahl zwischen verschiedenen Lernmodellen hat, wählt man besser das Lernmodell mit dem kleineren Wert für $\operatorname{bias} + \operatorname{var}$.

Beispiel 1 *([1], Example 2.8) Zielfunktion ist $f(x) = \sin \pi x$, $X \times Y = [-1,1]^2$. Man soll f aus zwei Punkten des Grafen von f lernen: $D = \{(x_1, f(x_1)), (x_2, f(x_2)) \mid x_1, x_2 \in [-1,1]\}$. Man kann zwischen zwei Hypothesenmengen wählen:*

$$H_0 = \{c_1 \mid c_1 \in [-1,1]\} \quad \textit{und} \quad H_1 = \{c_1 + c_2 x \mid c_1, c_2 \in \mathbf{R}\}.$$

Welches Lernmodell ist besser?

1. *H_0 lernt f als konstante Funktion.*

   Die finale Hypothese zur Lernmenge D ist $g^{(D)}(x) = (f(x_1) + f(x_2))/2$. Sie minimiert den in-sample error der quadratischen Abweichung

   $$\operatorname{Err}_{\text{in}}(c) = \frac{(c - f(x_1))^2 + (c - f(x_2))^2}{2}$$

   für Konstanten c, denn $\frac{d}{dc} \operatorname{Err}_{\text{in}}(c) = 0$ für $c = \frac{f(x_1)+f(x_2)}{2}$.

   Der Erwartungswert der finalen Hypothesen bezüglich der Lernmengen ist

   $$\bar{g}(x) = \mathbf{E}_D\, g^{(D)}(x) = \int_{[-1,1]} \int_{[-1,1]} \frac{f(x_1) + f(x_2)}{2} \, \frac{dx_1}{2} \, \frac{dx_2}{2} = 0.$$

   Die Verzerrung an der Stelle x ist $\operatorname{bias}(x) = (\bar{g}(x) - f(x))^2 = f(x)^2$ und

   $$\operatorname{bias} = \mathbf{E}_x \operatorname{bias}(x) = \int_{[-1,1]} f(x)^2 \, \frac{dx}{2} = \frac{1}{2}.$$

   Die Varianz bezüglich D an der Stelle x ist

   $$\mathbf{E}_D(g^{(D)}(x) - \bar{g}(x))^2 = \int_{[-1,1]} \int_{[-1,1]} \left(\frac{f(x_1) + f(x_2)}{2} \right)^2 \, \frac{dx_1}{2} \, \frac{dx_2}{2} = \frac{1}{4}$$

   und

   $$\operatorname{var} = \mathbf{E}_x \operatorname{var}(x) = \int_{[-1,1]} \frac{1}{4} \, \frac{dx}{2} = \frac{1}{4},$$

Für dieses Lernmodell haben wir damit

$$\mathbf{E}_D \operatorname{Err}_{\text{out}}(g^{(D)}) = \text{bias} + \text{var} = \frac{3}{4}$$

2. *H_1 lernt f als lineare Funktion, die die Punkte der Lernmenge interpoliert. Der in-sample error ist also für jede Lernmenge null.*

   Es ist

$$g^{(D)}(x) = c_1 + c_2 x \quad \textit{mit} \quad c_2 = (y_2 - y_1)/(x_2 - x_1) \quad \textit{und} \quad c_1 = y_1 - x_1 c_2,$$

$$\bar{g}(x) \approx 0.0 + 0.775929x \quad \textit{(numerische Integration)},$$

$$\text{bias} \approx 0.206717 \quad \textit{(numerische Integration)},$$

$$\text{var} \approx 1.67628 \quad \textit{(numerische Integration)},$$

$$\text{bias} + \text{var} \approx 1.883$$

Referenzen

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook, 2012.