Validierung

M. Gruber

KW 50, Rev.1

Wir folgen [1], Lec 13.

Validierung ist eine Methode zur Schätzung des *out-of-sample errors*. Ein Teil der Lernmenge (z.B. 20%) wird zum Zwecke der Validierung als sog. Validierungsmenge abgezweigt. Eine Anwendung der Validierung besteht z.B. darin, mehrere Lernmodelle zu vergleichen und sich für das Lernmodell mit dem geringsten Validierungsfehler zu entscheiden.

Der Validierungfehler $\mathrm{Err}_{\mathrm{val}}(h)$ Die Lernmenge wird disjunkt zerlegt in eine Trainingsmenge und eine Validierungsmenge: $D = D_{\mathrm{train}} \cup D_{\mathrm{val}}$ mit $|D| = N$, $|D_{\mathrm{val}}| = K$ und $|D_{\mathrm{train}}| = N - K$. Für $h \in H$ und den Datenpunkt $[x, y]$ sei $e(h(x), y)$ der elementare Fehler (im Klassifizierungsfall: $e(h(x), y)) = [h(x) \neq y]$, im Regressionsfall: $e(h(x), y)) = (h(x) - y)^2$). Der *out-of-sample error* $\mathrm{Err}_{\mathrm{out}}(h)$ ist der Erwartungswert $\mathbf{E}\, e(h(X), Y)$ (mit Zufallsvariablen X und Y). Bezeichnen wir die zugehörige Varianz $\mathbf{Var}\, e(h(X), Y) = \mathbf{E}(e(h(X), Y) - \mathrm{Err}_{\mathrm{out}}(h))^2$ mit σ^2. Der Validierungsfehler

$$\mathrm{Err}_{\mathrm{val}}(h) := \tfrac{1}{K} \sum_{[x,y] \in D_{\mathrm{val}}} e(h(x), y)$$

ist der Mittelwert von K unabhängigen Realisierungen der Zufallsvariablen $e(h(X), Y)$. Er ist ein erwartungstreuer Schätzer für $\mathrm{Err}_{\mathrm{out}}(h)$, denn

$$\mathbf{E}\, \mathrm{Err}_{\mathrm{val}}(h) = \tfrac{1}{K} \sum_{[x,y] \in D_{\mathrm{val}}} \mathbf{E}\, e(h(X), Y) = \tfrac{1}{K} \cdot K \cdot \mathrm{Err}_{\mathrm{out}}(h) = \mathrm{Err}_{\mathrm{out}}(h).$$

Seine Varianz ist

$$\mathbf{Var}\, \mathrm{Err}_{\mathrm{val}}(h) = \mathbf{E}(\mathrm{Err}_{\mathrm{val}}(h) - \mathrm{Err}_{\mathrm{out}}(h))^2 = \mathbf{E}\Big(\sum_{[x,y] \in D_{\mathrm{val}}} \tfrac{1}{K} \left((e(h(X), Y) - \mathrm{Err}_{\mathrm{out}}(h)) \right)^2.$$

Die Summanden im rechten Ausdruck sind unabhängige Zufallsvariablen. Deshalb ist der Erwartungswert des Quadrats der Summe gleich der Summe der Erwartungswerte der Quadrate:

$$\mathbf{E}\Big(\sum_{[x,y] \in D_{\mathrm{val}}} \tfrac{1}{K} \big((e(h(X), Y) - \mathrm{Err}_{\mathrm{out}}(h)) \big)^2 = \sum_{[x,y] \in D_{\mathrm{val}}} \mathbf{E}\left(\tfrac{1}{K} \left((e(h(X), Y) - \mathrm{Err}_{\mathrm{out}}(h)) \right)^2 \right).$$

Rechts steht somit $\sum_{[x,y]\in D_{\text{val}}}(\frac{1}{K})^2\sigma^2 = \frac{\sigma^2}{K}$ und wir erhalten insgesamt

$$\mathbf{Var}\,\text{Err}_{\text{val}}(h) = \sum_{[x,y]\in D_{\text{val}}}(\tfrac{1}{K})^2\sigma^2 = \frac{\sigma^2}{K}.$$

("Varianzreduktion durch Mittelung").

Wir können also davon ausgehen, dass

$$\text{Err}_{\text{out}}(h) \leq \text{Err}_{\text{val}}(h) + O\left(1/\sqrt{K}\right)$$

ist. Je größer K ist, desto genauer schätzt der Err_{val} den Err_{out}. Andrerseits werden wir in der Generalisierungstheorie lernen: je kleiner der Umfang $N-K$ der Trainingsmenge ist, desto größer ist die Wahrscheinlichkeit eines großen Generalisierungsfehlers.

Einfache Anwendung Die empfohlene Kompromiss für die Aufteilung der Lernmenge in Trainings- und Validierungsmenge ist $K = N/5$ ("Erfahrungswert").

Anwendung "Modellselektion" Gegeben seien M Lernmodelle mit den Hypothesenmengen $H_1, H_2, \ldots, H_M$ und eine gemeinsame Lernmenge $D = D_{\text{train}} \cup D_{\text{val}}$. Auf D_{train} wird gelernt. Die finale Hypothese aus H_k sei g_k^- für $k = 1, \ldots, M$. Die Validierungsfehler $\text{Err}_{\text{val}}(g_k^-)$ werden berechnet. Sei k_* der Index mit dem kleinsten Validierungfehler. Im k_*-ten Lernmodell wird nun auf ganz D gelernt. Ergebnis ist die finale Hypothese g_{k_*} mit

$$\text{Err}_{\text{out}}(g_{k_*}) \approx \text{Err}_{\text{out}}(g_{k_*}^-) \leq \text{Err}_{\text{val}}(g_{k_*}^-) + O\left(\sqrt{(\ln M)/K}\right).$$

Der Faktor $\ln M$ stammt aus der Generalisierungstheorie, mit der wir uns bald beschäftigen werden. Streng genommen dürfte die Validierungsmenge nicht erneut zum Lernen herangezogen werden. Wenn man dies dennoch tut, muss man wissen, dass man sich damit eine "optimistische Verzerrung" (*optimistic bias*) einhandelt.

Das sogenannte K-Dilemma besteht darin, dass man für $\text{Err}_{\text{out}}(g_{k_*}) \approx \text{Err}_{\text{out}}(g_{k_*}^-)$ ein möglichst kleines K braucht, aber für $\text{Err}_{\text{out}}(g_{k_*}^-) \approx \text{Err}_{\text{val}}(g_{k_*}^-)$ ein möglichst großes.

Es gibt jedoch einen Ausweg...

Kreuzvalidierung

- *Leave-one-out cross validation* (K=1)

  Hat man eine Lernmenge $D = \{[x^{(1)}, y^{(1)}], \ldots, [x^{(N)}, y^{(N)}]\}$ und M Lernmodelle mit Hypothesenmengen $H_1, H_2, \ldots, H_M$, so führt man pro Lernmodell die folgenden Schritte durch:

  1. Man bildet

     $$D_n = D \setminus \{[x_n, y_n]\}$$

für $n = 1, \ldots, N$ und ermittelt auf jedem D_n eine finale Hypothese g_n^-,

2. man berechnet für den Validierungsfehler

$$\mathrm{Err}_{\mathrm{val}}(g_n^-) = e(g_n^-(x_n), y_n) =: e_n$$

für jede finale Hypothese g_n^-,

3. man berechnet den Kreuzvalidierungs-Fehler *(cross validation error)* des Lernmodells

$$\mathrm{Err}_{\mathrm{cv}} = \frac{1}{N} \sum_{1 \leq n \leq N} e_n.$$

Das Lernmodell mit dem kleinsten $\mathrm{Err}_{\mathrm{cv}}$ gewinnt.

Der Rechenaufwand dieses Verfahrens ist hoch. Für große N (z.B. $N = 10^6$) stößt man an Grenzen. In diesem Fall empfiehlt sich die Vorgehensweise des nächsten Punkts.

- 10-*fold cross validation*

  Man bildet Blöcke der Größe $K = N/10$ und lässt beim Training jeweils einen Block aus. Auf diesem wird der Validierungsfehler bestimmt. Das arithmetische Mittel aller Validierungsfehler ist der Kreuzvalidierungsfehler. Dieser wird als Kriterium für die Wahl des Sieger-Lernmodells herangezogen.

Referenzen

[1] Y. S. Abu-Mostafa. Learning from Data.