Prediction with Expert Advice

Thoralf Mildenberger

DataLab
19. November 2013

Prediction with Expert Advice

Situation: *Forecaster* (**F**) möchte in Runde $t = 1, 2, \ldots$ Vorhersagen $\hat{p}_t \in \mathcal{D}$ über eine Zielgröße $y_t \in \mathcal{Y}$ treffen (oft, aber nicht immer $\mathcal{D} = \mathcal{Y}$).

Es werden – außer $y_t \in \mathcal{Y}$ – **keine** Voraussetzungen an y_t gemacht, insbesondere wird kein stochastisches oder statistisches Modell unterstellt.

Es stehen in jeder Runde jeweils Prognosen von *Experten* $\{f_{E,t} : E \in \mathcal{E}\}$ (mit $f_{E,t} \in \mathcal{D}$) zur Verfügung.

F erstellt zum Zeitpunkt t seine Prognose $\hat{p}_t$ auf Basis von $\{f_{E,t} : E \in \mathcal{E}\}$. Nachdem **F** die Prognose abgegeben hat, enthüllt *Natur* (**N**) den wahren Wert y_t.

F erleidet Verlust $\ell(\hat{p}_t, y_t)$.

Protokoll: Prediction with Expert Advice

Parameter: Entscheidungsraum $\mathcal{D}$, Ergebnisraum $\mathcal{Y}$, Verlustfunktion $\ell : \mathcal{D} \times \mathcal{Y} \longrightarrow \mathbb{R}^+$, Indexmenge $\mathcal{E}$ (entspricht Experten).

In jeder Runde t=1,2,. . . :

1. **N** wählt das nächste y_t und den Expertenrat $\{f_{E,t} : E \in \mathcal{E}\}$. **F** wird der Expertenrat mitgeteilt;
2. **F** wählt $\hat{p}_t \in \mathcal{D}$;
3. **N** gibt y_t bekannt;
4. **F** erleidet Verlust $\ell(\hat{p}_t, y_t)$, jeder Experte $E \in \mathcal{E}$ erleidet Verlust $\ell(f_{E,t}, y_t)$.

Prediction with Expert Advice

- **Keine** stochastischen Annahmen, **N** kann böswilliger Gegner sein!
- Weiterer Unterschied zur sequentiellen Statistik: Dort stammen die Daten i.d.R. aus einer Verteilung, am Ende muss (nach möglichst wenigen Runden) eine (möglichst gute) Entscheidung getroffen werden. Hier: Entscheidung *jede* Runde, Gesamtverlust soll klein sein.
- Die Experten funktionieren als Black Box, sie können z.B. ihre Prognose auf Basis von $y_1, \ldots, y_{t-1}$, von Insiderwissen, Astrologie o.ä. treffen. Sie können auch bewusst lügen.
- Die Experten können aber in diesem Framework durchaus auch statistische Methoden / Algorithmen sein, z.B. eine Prognose auf Basis eines Zeitreihenmodells liefern.
- In diesem Szenario: nach Abgabe der Prognose erfährt **F** y_t, seinen Verlust sowie den Verlust aller Experten. In sogenannten **multi-armed bandit**-Problemen erhält **F** nur eingeschränktes Feedback, erfährt z.B. nur seinen eigenen Verlust.

Prediction with Expert Advice

Ziel: Gesamtverlust von **F** nach n Runden

$$\hat{L}_n = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t)$$

soll "möglichst gering" sein.

Da nur die Expertenprognosen zur Verfügung stehen, ist ein *absolut* geringer Verlust i.d.R. nicht zu erreichen (die Experten könnten z.B. alle sehr schlechte Prognosen liefern).

Prediction with Expert Advice

Ausweg: Vergleiche Verlust von **F** nach n Runden mit dem Verlust des Experten $E \in \mathcal{E}$

$$R_{E,n} = \hat{L}_n - L_{E,n} = \sum_{t=1}^{n} \Big(\ell(\hat{p}_t, y_t) - \ell(\hat{F}_{E,t}, y_t) \Big)$$

(*Regret bezüglich* E).

Sinnvolleres Ziel: Maximaler Regret

$$\sup_{E \in \mathcal{E}} R_{E,n}$$

soll klein sein.

Prediction with Expert Advice

Wünschenswert:

$$\sup_{E \in \mathcal{E}} R_{E,n} = o(n)$$

bzw.

$$\frac{1}{n}\left(\hat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n}\right) \overset{n \to \infty}{\longrightarrow} 0$$

(Hannan-Konsistenz; durchschnittlicher Regret pro Runde geht gegen 0).

Prediction with Expert Advice

- Unterschied zur klassischen statistischen Entscheidungstheorie: hier soll nicht Risiko etc. *absolut* klein sein, stattdessen soll Methode auf lange Sicht "nicht viel schlechter" abschneiden als die beste Methode in einer Vergleichsklasse.
- Hier: Vergleich mit dem besten Experten (sog. externer Regret). Es gibt viele weitere Maße (internal regret, switching regret), die oft zu strengeren Kriterien führen.
- Im spieltheoretischen Sinne: Vergleich mit der besten *konstanten* Strategie. Interessanter vielleicht: Vergleich mit größeren Klassen.

Im Folgenden: $\mathcal{E} = \{1, \ldots, N\}$.

Halving und Weighted Majority

Beispiel: Prognose einer Folge von Bits, Verlust = Anzahl der Fehler. $\mathcal{D} = \mathcal{Y} = \{0, 1\}$, $\ell(\hat{p}_t, y_t) = \mathbb{I}(p_t \neq y_t)$. $N \in \mathbb{N}$ Experten, zunächst: es sei bekannt, dass einer der Experten immer richtig liegt.

Gesucht: Gute Schranke für Anzahl m Fehler von **F** in unendlich vielen Runden.

Lösung: Durch Angabe eines Algorithmus. Man erhält nur obere Schranke, es könnte ja noch einen besseren Algorithmus geben.

Halving und Weighted Majority

Algorithmus Halving:

1. Setze $w = (1, \ldots, 1) \in \mathbb{R}^N$
2. In jeder Runde t:
   1. Befrage die N Experten.
   2. Falls $|\{i : f_{i,t} = 1 \text{ und } w_i = 1\}| > |\{i : f_{i,t} = 0 \text{ und } w_i = 1\}|$: $\hat{p}_t = 1$
      Sonst: $\hat{p}_t = 0$
   3. Erhalte y_t.
   4. Falls $y_t \neq \hat{p}_t$, setze $w_i = 0$ für alle i mit $f_{i,t} \neq y_t$.

Halving und Weighted Majority

Falls ein Fehler gemacht wird, muss mindestens die Hälfte aller Experten mit Gewicht 1 falsch gelegen haben. Deren Gewicht wird dann auf 0 gesetzt, d.h. bei jedem Fehler halbiert sich die Anzahl der Experten mit Gewicht 1 mindestens. Dies kann höchsten geschehen, bis nur noch ein Experte Gewicht 1 hat, also gilt für die Anzahl der Fehler m

$$m \leq \log_2 N.$$

Halving und Weighted Majority

Jetzt: Die selbe Situation, aber es wird nicht mehr vorausgesetzt, dass ein Experte immer richtig prognostiziert. Betrachte folgenden Algorithmus für ein festgewähltes $\beta \in (0,1)$:

Algorithmus Weighted Majority:

1. Setze $w = (1, \ldots, 1) \in \mathbb{R}^N$
2. In jeder Runde t:
   1. Befrage die N Experten.
   2. Falls $\sum_{i:f_{i,t}=1} w_i > \sum_{i:f_{i,t}=0} w_i$: $\hat{p}_t = 1$
      Sonst: $\hat{p}_t = 0$
   3. Erhalte y_t.
   4. Falls $y_t \neq \hat{p}_t$, setze $w_i \leftarrow \beta w_i$ für alle i mit $f_{i,t} \neq y_t$.

Halving und Weighted Majority

Sei $W_m = w_1 + \cdots + w_N$ zu dem Zeitpunkt, zu dem **F** den m-ten Fehler macht ($W_0 = N$). Beim m-ten Fehler müssen die falsch liegenden Experten zusammen ein Gewicht von mindestens $W_{m-1}/2$ haben. Ihr Gewicht wir mit dem Faktor $\beta < 1$ multipliziert, das Gewicht der anderen Experten, das höchstens $W_{m-1}/2$ beträgt, bleibt unverändert. Damit erhalten wir:

$$W_m \leq (1+\beta)\frac{W_{m-1}}{2}.$$

Wiederholte Anwendung liefert

$$W_m \leq \left(\frac{1+\beta}{2}\right)^m W_0 = \left(\frac{1+\beta}{2}\right)^m N.$$

Halving und Weighted Majority

Betrachte nun den Experten $k \in \{1, \ldots, N\}$, der bis zum Zeitpunkt m die wenigsten Fehler gemacht hat. Bezeichne die Anzahl seiner Fehler mit m^*. Nach m^* Fehlern ist das Gewicht dieses Experten β^{m^*}, und damit

$$W_m \geq \beta^{m^*}.$$

Damit erhalten wir

$$\beta^{m^*} \leq W_m \leq \left(\frac{1+\beta}{2}\right)^m N.$$

bzw.

$$m^* \log_2(\beta) \leq m \log_2\left(\frac{1+\beta}{2}\right) + \log_2 N.$$

Halving und Weighted Majority

$$m^* \log_2(\beta) \leq m \log_2 \left(\frac{1+\beta}{2} \right) + \log_2 N$$

$$\Leftrightarrow \quad -m^* \log_2(\beta) \geq m \log_2 \left(\frac{2}{1+\beta} \right) - \log_2 N$$

$$\Leftrightarrow \quad m \leq \frac{\log_2 N + m^* \log_2(1/\beta)}{\log_2 \frac{2}{1+\beta}}$$

Für festes β ist diese Schranke logarithmisch in N und linear in m^*.

Exponentially Weighted Average Forecaster

Im Folgenden: $\mathcal{D}$ konvexe Teilmenge eines Vektorraums, ℓ konvex im 1. Argument und mit Werten in $[0, 1]$.

Vorhersage durch gewichtetes Mittel der Expertenprognosen:

$$\hat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{t,i}}{\sum_{j=1}^{N} w_{j,t-1}}$$

mit Gewichten $w_{1,t-1}, \ldots, w_{N,t-1} \geq 0$.

Exponentially Weighted Average Forecaster

Exponentially Weighted Average Forecaster: Spezielle Wahl der Gewichte (für $\eta > 0$):

$$w_{i,t-1} = \exp(\eta R_{i,t-1}).$$

Es ergibt sich:

$$\begin{aligned}
\hat{p}_t &= \frac{\sum_{i=1}^{N} \exp(\eta R_{i,t-1}) f_{t,i}}{\sum_{j=1}^{N} \exp(\eta R_{j,t-1})} \\
&= \frac{\sum_{i=1}^{N} \exp(\eta(\hat{L}_{t-1} - L_{i,t-1})) f_{t,i}}{\sum_{j=1}^{N} \exp(\eta(\hat{L}_{t-1} - L_{j,t-1}))} \\
&= \frac{\sum_{i=1}^{N} \exp(-\eta L_{i,t-1}) f_{t,i}}{\sum_{j=1}^{N} \exp(-\eta L_{j,t-1})}.
\end{aligned}$$

Exponentially Weighted Average Forecaster

Einfache Updateformel: mit

$$v_{i,0} = 1$$

$$v_{i,t} = \frac{v_{i,t-1} \exp(-\eta \ell(f_{t,i}, y_t))}{\sum_{j=1}^{N} v_{j,t-1} \exp(-\eta \ell(f_{t,j}, y_t))} \quad (t \geq 1)$$

für $i = 1, \ldots, N$ gilt

$$\hat{p}_t = \sum_{i=1}^{N} v_{i,t-1} f_{i,t}.$$

Exponentially Weighted Average Forecaster

Satz (Schranke für Regret des EWA-Forecasters)

Sei ℓ konvex im ersten Argument, mit Werten in $[0,1]$. Dann gilt für alle $\eta > 0$, $n \in \mathbb{N}$, $y_1, \ldots, y_n \in \mathcal{Y}$:

$$\hat{L}_n - \min_{i=1,\ldots,N} L_{i,n} \leq \frac{\log N}{\eta} + \frac{n\eta}{8}.$$

Beweis: Sei

$$W_t := \sum_{i=1}^{N} w_{i,t} = \sum_{i=1}^{N} \exp(-\eta L_{i,t})$$

für $t \geq 1$ und $W_0 = N$.

Exponentially Weighted Average Forecaster

Zunächst gilt:

$$\begin{aligned}
\log \frac{W_n}{W_0} &= \log \left(\sum_{i=1}^{N} \exp(-\eta L_{i,n}) \right) - \log N \\
&\geq \log \left(\max_{i=1,\ldots,N} \exp(-\eta L_{i,n}) \right) - \log N \\
&= -\eta \min_{i=1,\ldots,n} L_{i,n} - \log N.
\end{aligned}$$

Exponentially Weighted Average Forecaster

Weiter gilt für alle $t = 1, \ldots, n$:

$$\begin{aligned}
\log \frac{W_t}{W_t - 1} &= \log \frac{\sum_{i=1}^N w_{i,t}}{\sum_{j=1}^N w_{j,t-1}} \\
&= \log \frac{\sum_{i=1}^N \exp(-\eta \ell(f_{i,t}, y_t)) \exp(-\eta L_{i,t-1})}{\sum_{j=1}^N \exp(-\eta L_{j,t-1})} \\
&= \log \frac{\sum_{i=1}^N w_{i,t-1} \exp(-\eta \ell(f_{i,t}, y_t))}{\sum_{j=1}^N w_{j,t-1}}
\end{aligned}$$

Exponentially Weighted Average Forecaster

Anwendung der

Hoeffding-Ungleichung

Für ZV $X \in [a, b]$, $s \in \mathbb{R}$ gilt:

$$\log \mathbb{E}(\exp(sX)) \leq s\mathbb{E}X + \frac{s^2(b-a)^2}{8}.$$

auf diskrete ZV $X = \ell(f_{i,t}, y_t)$ mit Werten in $[a, b] = [0, 1]$ und $s = -\eta$ liefert:

$$\log \frac{\sum_{i=1}^N w_{i,t-1} \exp(-\eta \ell(f_{i,t}, y_t))}{\sum_{j=1}^N w_{j,t-1}} \leq -\eta \frac{\sum_{i=1}^N w_{i,t-1} \ell(f_{i,t}, y_t)}{\sum_{j=1}^N w_{j,t-1}} + \frac{\eta^2}{8}$$

Exponentially Weighted Average Forecaster

Wegen der Konvexität von ℓ im ersten Argument gilt:

$$-\eta \frac{\sum_{i=1}^N w_{i,t-1}\ell(f_{i,t}, y_t)}{\sum_{j=1}^N w_{j,t-1}} + \frac{\eta^2}{8} \leq -\eta\ell\left(\frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}}, y_t\right) + \frac{\eta^2}{8}$$
$$= -\eta\ell(\hat{p}_t, y_t) + \frac{\eta^2}{8}.$$

Weiter ist

$$\log \frac{W_n}{W_0} = \sum_{t=1}^n \log \frac{W_t}{W_t - 1}$$
$$\leq \sum_{t=1}^n \left(-\eta\ell(\hat{p}_t, y_t) + \frac{\eta^2}{8}\right)$$
$$= -\eta \hat{L}_n + \frac{\eta^2}{8} n.$$

Exponentially Weighted Average Forecaster

Zusammen mit der unteren Schranke ergibt sich:

$$-\eta \min_{i=1,\dots,n} L_{i,t} - \log N \leq \log \frac{W_n}{W_0} \leq -\eta \hat{L}_n + \frac{\eta^2}{8} n.$$

bzw.

$$\hat{L}_n - \min_{i=1,\dots,n} L_{i,t} \leq \frac{\log N}{\eta} + \frac{\eta}{8} n$$

□

Ableiten und Nullsetzen der rechten Seite liefert $\eta = \sqrt{8 \log(N)/n}$ und damit die obere Schranke $\sqrt{(n/2) \log N}$.

Literatur

Cesa-Bianchi, N., Lugosi, G. (2005), *Prediction, Learning, and Games*, Cambridge University Press, Cambridge.