

Medizinische Hochschule Hannover  
Institut für Biometrie

Schätzung des Vorhersagefehlers:  
Kreuzvalidierung und Bootstrap

Eine kurze Einführung

H. Hecker  
November 1997

**Contents**

1	Allgemeine Beschreibung der Fragestellung . . . . .	1
2	Voraussetzungen und Bezeichnungen . . . . .	1
3	Anmerkungen . . . . .	2
4	Schätzung des Vorhersagefehlers . . . . .	3
	4.1 Kreuzvalidierung: . . . . .	3
	4.2 Bootstrap . . . . .	4
5	Weitere Fragestellungen . . . . .	5

# Schätzung des Vorhersagefehlers: Kreuzvalidierung und Bootstrap

H.Hecker 26.2.98

## 1. Allgemeine Beschreibung der Fragestellung

In der Diskriminanzanalyse und Regressionsanalyse –im vorliegenden Zusammenhang können Baumanalysen (CART und RECPAM) und Künstliche Neuronale Netze (KNN) mit überwachtem Lernen dazugerechnet werden– wird im wesentlichen die folgenden Aufgabe behandelt:

- Voraussetzung: Es ist eine *Lernstichprobe* gegeben. Diese ist dadurch charakterisiert, daß für jeden einzelnen "Fall" (Beobachtungseinheit) sowohl *prognostische Variablen* als auch eine *Zielgröße* erfaßt wurden.
- Aufgabe: Mit Hilfe der Daten der Lernstichprobe ist eine Funktion zu spezifizieren, mit welcher aus den Werten der prognostischen Variablen *mit möglichst großer Genauigkeit* auf den Wert der Zielvariablen geschlossen werden kann. Diese Funktion wird die *Vorhersageregeln* genannt.
- Die Forderung nach größtmöglicher Genauigkeit bezieht sich auf die Anwendung der Vorhersageregeln auf *neue Fälle*, in denen *nur die prognostische Variablen gemessen* wurden und *der Wert der Zielvariablen unbekannt* ist.

Aus diesem dritten Punkt ergibt sich folgende Fragestellung:

Angenommen es ist ein Prozedur festgelegt, mit der aus der Lernstichprobe eine Vorhersageregeln erzeugt wird: Wie groß ist dann der Fehler dieser Vorhersageregeln bei der Anwendung auf neue Fälle?

Es werden hierzu einige Grundideen aus der Arbeit von Efron(1983) dargestellt. Die dort benutzte Terminologie wird etwas modifiziert.

## 2. Voraussetzungen und Bezeichnungen

- Die Lernstichprobe besteht aus der Realisierung von  $n$  unabhängig, identisch verteilten, mehrdimensionalen Zufallsvariablen  $X_1, X_2, \dots, X_n$ .
- Deren gemeinsame Verteilung sei mit  $F$  bezeichnet.
- Die Zufallsvariablen  $X_i$  setzen sich jeweils aus einem  $p$ -dimensionalen Prädiktor  $T_i = (T_{i1}, T_{i2}, \dots, T_{ip})$  und einer 1-dimensionalen Zielgröße  $Y_i$  zusammen:

$$X_i = (T_i, Y_i)$$

- Es gibt eine festgelegte Prozedur, die aus den Werten  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  der Lernstichprobe eine Vorhersageregeln erzeugen. Diese Vorhersageregeln wird mit  $\eta_{\mathbf{x}}$  bezeichnet. Sie ist auf  $p$ -dimensionale Prädiktorwerte  $t = (t_1, t_2, \dots, t_p)$  anzuwenden. Das Ergebnis wird mit  $\eta_{\mathbf{x}}(t)$  bezeichnet.
- Der "Abstand" zwischen einem "vorhergesagten" Wert  $\eta_{\mathbf{x}}(t)$  und dem Wert  $y$  der Zielgröße wird mit einer Funktion  $Q$  gemessen und mit  $Q[y, \eta_{\mathbf{x}}(t)]$  bezeichnet.

- Definition der "wahren Fehlerrate"  $Err(\mathbf{x})$  bei gegebener Lernstichprobe  $\mathbf{x}$ :

$$Err(\mathbf{x}) = E_F(Q[Y_0, \eta_{\mathbf{x}}(T_0)])$$

wobei der Erwartungswert sich auf die Zufallsvariablen  $T_0$  und  $Y_0$  bezieht, also eine von der Lernstichprobe  $\mathbf{x}$  unabhängige ("neue") Realisierung von  $X_0 = (T_0, Y_0)$  mit dem Prädiktorvektor  $T_0$  und der Zielvariablen  $Y_0$  und der Verteilung  $F$ .

Dies ist also die hier primär interessierende Größe. Man beachte:  $Err(\mathbf{x})$  hängt von der Lernstichprobe  $\mathbf{x}$  ab, kann also auch als Zufallsvariable  $Err(\mathbf{x})$  angesehen werden.

- Definition der "scheinbaren" Fehlerrate (apparent error, resubstitution error):

$$\bar{err}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{\mathbf{x}}(t_i)]$$

Hier wird also der *durchschnittliche Fehler in der Lernstichprobe* berechnet, wobei die aus der Lernstichprobe  $\mathbf{x}$  erzeugte Vorhersageregeln  $\eta_{\mathbf{x}}$  auf alle Einheiten der Lernstichprobe selber wieder angewendet werden.

Diese Fehlerrate ist meist ("in der Erwartung") kleiner als die wahre Fehlerrate:  $\bar{err}(\mathbf{x})$  ist ein zu optimistischer Wert für  $Err(\mathbf{x})$ . Die Differenz wird daher von Efron mit "op" (für "optimism") bezeichnet:

- 

$$op(\mathbf{x}) = Err(\mathbf{x}) - \bar{err}(\mathbf{x})$$

Auch  $op$  ist von der Lernstichprobe abhängig und daher eine Zufallsvariable  $op(\mathbf{X})$ . Ihr Erwartungswert wird mit  $\omega$  bezeichnet:

$$\omega = E_F(op(\mathbf{X}))$$

### 3. Anmerkungen

#### Anmerkung 1:

In der Arbeit von Efron spielt i.f. dieser "Optimismus"  $op(\mathbf{x})$  die zentrale Rolle: Die Schätzung der wahren Fehlerrate  $Err(\mathbf{x})$  wird immer als *Schätzung der Differenz zur scheinbaren Fehlerrate* betrachtet:

$$\begin{aligned} Err(\mathbf{x}) &= \bar{err}(\mathbf{x}) + [Err(\mathbf{x}) - \bar{err}(\mathbf{x})] \\ &= \bar{err}(\mathbf{x}) + op(\mathbf{x}) \end{aligned}$$

und für irgendeinen Schätzer von  $Err(\mathbf{x})$  entsprechend:

$$\widehat{Err}(\mathbf{x}) = \bar{err}(\mathbf{x}) + \widehat{op}(\mathbf{x})$$

Schätzer  $\widehat{op}(\mathbf{x})$  für den "Optimismus" werden mit  $\widehat{\omega}$  bezeichnet:

$$\widehat{Err}(\mathbf{x}) = \bar{err}(\mathbf{x}) + \widehat{\omega}(\mathbf{x}) \tag{3.1}$$

#### Anmerkung 2:

Es ist vielleicht nicht immer ganz klar, ob ein Schätzer der Form  $\bar{err}(\mathbf{x}) + \hat{\omega}(\mathbf{x})$  als Schätzer des Vorhersagefehlers eine konkret vorliegenden Vorhersageregeln  $\eta_{\mathbf{x}}$  oder für deren Erwartungswert aufzufassen ist. Anzustreben ist jedenfalls ein Schätzverfahren, das den Vorhersagefehler  $Err(\mathbf{x})$  für jede Lernstichprobe  $\mathbf{x}$  (und nicht nur im Mittel über alle Lernstichproben) möglichst nahe kommt.

**Graphische Darstellung** der Zusammenhänge:

Grundgesamtheit  $\Omega$ , Lernstichprobe  $\mathbf{x}$ , Vorhersageregeln  $\eta_{\mathbf{x}}$ , Teststichprobe  $X_0$ , Vorhersagefehler  $Err(\mathbf{x})$  und Scheinbarer Fehler  $\bar{err}(\mathbf{x})$

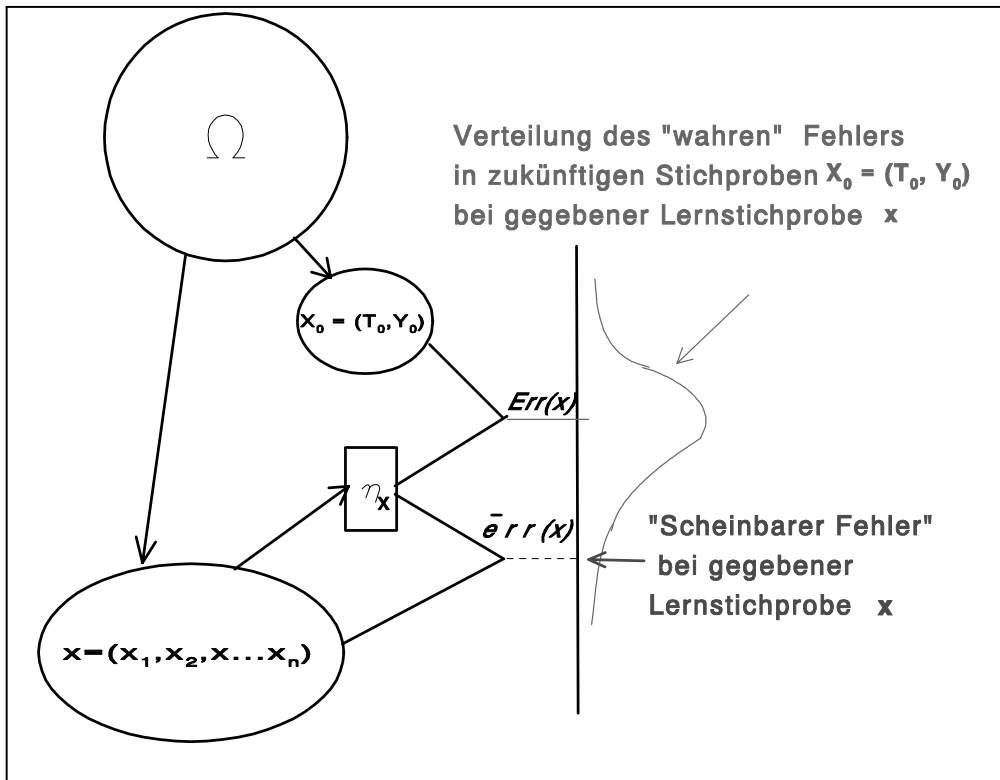


Figure 3.1:

**4. Schätzung des Vorhersagefehlers**

**4.1. Kreuzvalidierung:**

$$\hat{\omega}^{(CV)}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{\mathbf{x}_{(i)}}(t_i)] - \bar{err}(\mathbf{x})$$

Dabei bezeichnet  $\mathbf{x}_{(i)}$  die Stichprobe, die aus der Lernstichprobe  $\mathbf{x}$  dadurch entsteht, daß die Beobachtungseinheit Nr.  $i$  entfernt wird.

$\eta_{\mathbf{x}(i)}$  ist die aus dieser reduzierten Lernstichprobe erzeugte Vorhersageregeln. Die Schätzung der wahren Vorhersagefehler ist daher

$$\begin{aligned}\widehat{Err}^{(CV)}(\mathbf{x}) &= \bar{err}(\mathbf{x}) + \widehat{\omega}^{(CV)}(\mathbf{x}) \\ &= \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{\mathbf{x}(i)}(t_i)]\end{aligned}$$

Sie ist also gleich dem Durchschnitt der Fehler, die entstehen, wenn jede Beobachtungseinheit 1 mal als Teststichprobe und gleichzeitig jeweils der Rest als Lernstichprobe benutzt wird.

## 4.2. Bootstrap

Die Grundidee des bootstrap besteht in Folgendem:

Es sei die Situation gegeben, daß aus einer Stichprobe  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ , gezogen aus einer Grundgesamtheit mit der (meist multivariaten) Verteilungsfunktion  $F$ , eine bestimmte Funktionen  $g(\mathbf{X})$  gebildet wird. Dann kann man die Verteilung von  $g(\mathbf{X})$  dadurch schätzen, daß man zunächst die Verteilungsfunktion  $F$  durch eine geschätzte Verteilungsfunktion  $F^*$  ersetzt. Die Verteilung von  $g(\mathbf{X})$  wird dann unter der Annahme bestimmt, daß statt  $F$  die geschätzte Verteilung  $F^*$  zugrundeliegt.

Die Schätzung von  $F^*$  von  $F$  erfolgt mit Hilfe der Lernstichprobe:

**Definition: Die Verteilung  $F^*$  ist diskret mit der Wahrscheinlichkeiten  $\frac{1}{n}$  für jeden der Stichprobenwerte  $x_1, x_2, \dots, x_n$ .**

Aufgrund dieser Festlegung läßt sich die Verteilung von  $g(\mathbf{X})$  beliebig genau durch Simulation bestimmen:

1. Bilde eine Stichprobe  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_n^*)$  vom Umfang  $n$  aus  $F^*$  und bilde daraus  $g(\mathbf{X}^*)$ . Bei dieser Stichprobenziehung hat jedesmal jeder der Werte  $x_1, x_2, \dots, x_n$  die Wahrscheinlichkeit  $\frac{1}{n}$ , gezogen zu werden (das bedeutet: "Ziehen mit Zurücklegen").
2. Führe diesen Schritt 1 sehr oft durch und bilde daraus die empirische Verteilung von  $g(\mathbf{X}^*)$  (und dalle daraus interessierenden Parameter).

Die Anwendung dieses Prinzips auf die Schätzung des "Optimismus"  $op(\mathbf{x})$  geht wie folgt:

1. Bootstrap-Schätzung für  $\bar{err}(\mathbf{x})$ : Aus einer bootstrap-Stichprobe  $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)$  bilde die "scheinbare Fehlerrate"  $\bar{err}(\mathbf{x}^*)$ :

$$\begin{aligned}\bar{err}(\mathbf{x}^*) &= \frac{1}{n} \sum_{i=1}^n Q[y_i^*, \eta_{\mathbf{x}^*}(t_i^*)] \\ &= \sum_{i=1}^n P_i^* Q[y_i, \eta_{\mathbf{x}^*}(t_i)]\end{aligned}$$

mit

$$P_i^* = \frac{|\{x_j^* \mid (j = 1, \dots, n), x_j^* = x_i\}|}{n}$$

2. Die Fehlerrate von  $\eta_{\mathbf{x}^*}$  Unter der Annahme, daß  $F^*$  die zugrundeliegende Verteilung ist, kann die Fehlerrate von  $\eta_{\mathbf{x}^*}$  unmittelbar berechnet werden; sie ist

$$Errr^*(\mathbf{x}^*) = \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{\mathbf{x}^*}(t_i)]$$

denn in einer neuen Beobachtung erscheint jeder Wert  $x_i = (t_i, y_i)$  mit der Wahrscheinlichkeit  $\frac{1}{n}$ , und er ist verbunden mit dem Fehler  $Q[y_i, \eta_{\mathbf{x}^*}(t_i)]$ .

3. Insgesamt folgt daraus für den bootstrap-Schätzer für  $\hat{\omega}^{(BOOT)}(\mathbf{x})$  für  $op(\mathbf{x})$  :

$$\hat{\omega}^{(BOOT)}(\mathbf{x}) = E_{F^*} \sum_{i=1}^n \left(\frac{1}{n} - P_i^*\right) Q[y_i, \eta_{\mathbf{x}^*}(t_i)]$$

Der Erwartungswert  $E_{F^*}$  über die Verteilung  $F^*$  wird dabei durch den Mittelwert der bootstrap-Zufalls-Stichproben ermittelt.

Damit lautet die Schätzung für den *Vorhersagefehler*:

$$\begin{aligned} \hat{Errr}^{(BOOT)}(\mathbf{x}) &= \bar{err}(\mathbf{x}) + \hat{\omega}^{(BOOT)}(\mathbf{x}) \\ &= \bar{err}(\mathbf{x}) + E_{F^*} \sum_{i=1}^n \left(\frac{1}{n} - P_i^*\right) Q[y_i, \eta_{\mathbf{x}^*}(t_i)] \end{aligned}$$

## 5. Weitere Fragestellungen

- Weitere bootstrap-Versionen der Fehlerschätzung
- Vergleich verschiedener Schätzer in Hinblick auf Genauigkeit, Verzerrung
- Spezifizierung für logistische Regression, COX-Regression(?). Dabei
  - Einbezug der Variablenselektion in die Fehlerschätzung
  - Einbezug der Fehlerrate in die Modellwahl (z.B. in CART, RECPAM, KNN)
- Software Realisierungen, Anwendung in Beratung und Auswertung

### Literatur:

Efron, Bradley (1983). *Estimating the error rate of a prediction rule: Improvement on cross-validation*.

Journal of the American Statistical Association, **78**: 316-331.

**Key Words:** Analysis of variance [ANOVA]; Bootstrap; Logistic regression