technische universität
dortmund

Bachelorarbeit

Bootstrapping Ansätze zur Bestimmung von Konfidenzbändern für Verteilungsfunktionen

Dennis Richter
Monat der Abgabe

Gutachter:
Prof. Dr. Peter Buchholz
Name des Zweitgutachters

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für praktische Informatik (LS 4)
https://ls4-www.cs.tu-dortmund.de

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

1.2 Zielsetzung

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

-

1.3 Aufbau der Arbeit

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

-

Kapitel 2

Grundlagen

$$y_j = \eta(x_j, \theta_0) + \epsilon_j, \quad j = 1, 2, ..., n \text{ und } \epsilon \sim N(0, \sigma^2) \tag{2.1}$$

- Quellen: Barton 1998, Krazanowiski 1998
- geben sei eine situation, wo ein regression metamodel als repräsentation fpr die ausgabe einer simulations studie verwendet wird
- n unabhängige versuche und die beobachteten werte des simulations models seien gegeben durch y
- y ist zufallsvariable abhängig von einem design punkt x
- eine gewöhnliche darstellung ist durch ... (means of a statistical metamodel)
- ein paar beispiele die zur orientierung dienten geben
- eta bezeichnet eine deterministische funktion, namentlich regressionsfunktion
- genauer gesagt ist eta ein parmetrisches statistisches metamodel
- es wird angenommen, dass die simulation durch dieses modell repräsentiert werden kann und der erwartungswert dieses metamodels spiegelt dann den wahren erwartungswert wieder, vorausgesetzt die annahme trifft zu
- epsilon bezeichnet den für alle design punkte unabhängigen zufallsfehler mit mean 0
- oft wird angenommen dass die varianz für alle design punkte gleich ist, aber nicht zwingend
- jedoch mean(epsilon)=0 bedeutet eta gibt den erwartungs wert der statistik an
- die berechnung der regressionsfunktion ist primäres ziel der simulations studie
- angenommen das modell repräsentiert die simulation und es ex ein wahrer wert theta0, dann ist das erste problem diesen zu bestimmen bzw zu schätzen
- die mit abstand effektivste methode theta in parametrischen studien zu schätzen ist die maximum likelyhood methode, welche im nächsten abschnitt kurz vorgestellt werden soll

2.1 Maximum Likelihood Estimation

- gegeben sei eine menge von unabhängigen stichproben, erhalten von verteilungsfunktionen fi
- in unserem regressionsfall zb sind die yi werte verteilt mit... sodass wir als verteilungsfunktionen ... erhalten
- die joint distribution aus den verteilungsfunktionen bezüglich aller yi ist eine funktion in abhängigkeit von theta und gegeben der stichprobe y und wird likelihood von theta genannt
- ziel der mle ist nun diese funktion zu maximieren
- das maximum ist durch theta mit ableitung 0 gegebne,da es aber sehr umständlich ist ein produkt abzuleiten, betrachtet man stattdessen den logarithmus der likelihood, welcher sich als summe der einzelnen logarithmen schreiben lässt
- da der logarithmus eine streng monoton steigende funktion ist lässt sich der mle nun auch als maximum der loglikelihood bestimmen
- ein sehr praktischer ansatz ist nun die bestimmung des maximums mittels numerischer verfahren, die nelder mead methode liefert ein robustes suchverfahren
- wichtige erkenntnisse über den mle sind nun dass unter sehr allgemeinen voraussetzungen mle multivariat normalverteilt ist mit mean theta0(wahrer wert) und varianz matrix V(theta0)
- wobei V sich als inverse der fischer informations matrix berechnen lässt, welches widerum der erwartungswert der hessischen matrix der loglik ist
- V kann durch V(mle) angenähert werden

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

-
-
-
-
-
-
-
-
-
-
-
-
-
-

2.2 Confidence intervals

$$\mathbb{P}\left(\theta_L \leq \theta_0 \leq \theta_U\right) \geq 1-\alpha \tag{2.2}$$

$$\theta_L, \theta_U = \hat{\theta} \mp z_{\alpha/2}\sqrt{V(\hat{\theta})} \tag{2.3}$$

- wenn man ein parametrisches modell gefunden hat, welches die daten reptäsentieren soll, ist eine offensichtliche frage, wie akkurat diese schätzung ist
- oft wird ein interval bezgl des schätzers angegeben welches den wahren wert mit gewünschter wahrscheinlichkeit überdeckt
- solch ein interval heißt confidence intervall, klassische methoden zur bestimmung eines solchen intervals bauen auf asymtotischer theory und der sogenannten delta methode auf
- darstellung von ci zeigen
- ein konfindenz intervall für die varianz theta[1] in allen design punkten ist gegeben durch ... da ...
- eher interessiert uns allerdings ein confidence interval für die regressionsfunktion
- aufgrundlage von asymptotischer theorie (genauer die taylor expansion) und der delta-methode erhalten wir ein confidence interval für eta ... - herleitung von russel zeigen...
- man beachte dass die ableitung und ... durch finite difference methoden berechnet werden könnne
- nachteil dieser herangehensweise sind ... -> bootstrap
-
-

-
-
-
-
-
-
-

2.3 Simultaneous Confidence Intervals

$$\mathbb{P}\left(y_L(x) \leq \eta(x, \theta_0) \leq y_U(x)\right) \geq 1 - \alpha \quad \forall x \in \mathbb{R} \tag{2.4}$$

$$y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp z_{\alpha/2} \sqrt{\left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}^T V(\hat{\theta}) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}} \tag{2.5}$$

- Quellen: Miller 1981
- interessanter für uns, als die schätzung von confidence intervallen für die einzelnen design punkte, ist eine schätzung von confidence intervallen, die für alle werte simultan gilt
- gesucht ist ein band welches mit gewünschter wahrscheinlichkeit die gesamte regressionsfunktion überdeckt
- beispiele von miller 1981 nenen ...
- eine einfaches und conservatives confidence band erhält man duch anwendung der taylor reihen expansion und asymptotischer theorie
- beispiel von russel zeigen ...
- für kleine n aller dings erhält man durch diesen ansatz oft fälschlich höhere werte für die konfidence als der eigentlich berechnete konfidence
- bootstrap kann in diesem fall helfen
-
-
-
-
-
-
-
-
-

2.4 Likelihood based Confidence Region

$$\mathbb{P}\left(y_L(x) \leq \eta(x, \theta_0) \leq y_U(x) \quad \forall x \in \mathbb{R}\right) \geq 1 - \alpha \tag{2.6}$$

$$y_L(x), y_U(x) = \eta(x, \hat{\theta}) \mp \sqrt{\chi_p^2(a) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}^T V(\hat{\theta}) \left(\frac{\partial \eta(x, \theta)}{\partial \theta}\right)_{\hat{\theta}}} \tag{2.7}$$

2.5 Coverage Error

- die qualität der confindence bereiche wir oft in form sogenannter coverage error beschrieben
- diese können durch asymptotische theorie bestimmt werden, auch für die bootstrap versionen
- coverage error ist in der regel O(1 / sqrt n) aber kann oft auf O(1 / n) durch balanced ci reduziert werden
- coverage error kommt hauptsächlich vom bias, da der effekt entgegengesetzte ist links und rechts von null hebt er sich auf, falls die ci balanciert werden
- coverage error kann als maß für den unterschied zwischen erreichter und gewünschter überdeckungswahrscheinlichkeit dienen

-

-

-

-

-

-

-

-

-

-

2.6 Basic Bootstrap

2.7 Parametric Bootstrap

for $j = 0$ to B **do**
  for $i = 0$ to n **do**
    ziehe ein Sample y_{ij} von $F(.)$
  end for
  berechne die Statistik $s_j = s(y_j)$
end for

Algorithmus 2.1: Basic-Sampling Methode

Eingabe: zufälliges Sample $y = (y_1, y_2, ...y_n)$ von $F(.)$
  erstelle die EDF $F_n(.|y)$
  for $j = 0$ to B **do**
    for $i = 0$ to n **do**
      ziehe ein Sample y^*_{ij} von $F_n(.|y)$
    end for
    berechne die Statistik $s^*_j = s(y^*_j)$
  end for
  erstelle die EDF $G_n(.|s*)$

Algorithmus 2.2: Bootstrap-Sampling Methode

Kapitel 3

Vorstellung der Algorithmen

3.1 parametrisches Bootstrapping

3.2 nicht-parametrisches Bootstrapping

Kapitel 4

Implementierung

4.1 Parameterstudien in OMNeT++

4.2 Testszenario

4.3 TODO

Kapitel 5

Auswertung

5.1 Analytisches Verfahren

Eine Referenz [1].

5.2 parametrisches Bootstrapping

5.3 nicht-parametrisches Bootstrapping

5.4 Vergleich

Kapitel 6

Schlussteil

6.1 Fazit

6.2 Ausblick

Anhang A

Weitere Informationen

Abbildungsverzeichnis

Algorithmenverzeichnis

Literaturverzeichnis

[1] Aggarwal, Alok und Jeffrey Scott Vitter: *The Input/Output Complexity of Sorting and Related Problems.* Communications of the ACM, 31(9):1116–1127, 1988.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 14. Februar 2021

Muster Mustermann