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

Bei der statistischen Analyse von Daten ist es oft üblich aufwendige theoretische Modelle vorauszusetzen zu müssen, um aussagekräftige Ergebnisse über eine Stichprobe zu erhalten.

Die Auswahl eines statistischen Modells welches das wahre Modell so gut wie möglich repräsentiert, ist oft eine Herausforderung aber gleichzeitig ausschlaggebend für den Erfolg der Analyse.

Nur bedingt erfüllte Annahmen führen zu falschen Aussagen, zu spezifische Modelle hingegen lassen sich nicht Computer gestützt umsetzen und müssen per Hand analysiert werden.

Die Simulation der wahren Population durch die gegeben Stichprobe liefern hier einen Weg, diese Schwierigkeit zu umgehen.

Die Idee solcher sogenannten Resampling-Verfahren ist aus einer kleinen Anzahl von Stichproben beliebig viele Stichproben zu generieren, indem die ursprünglichen Daten als Schätzer für die Grundgesamtheit dienen, von der nun beliebig oft gesampelt werden kann.

Anstelle eine Verteilung vorauszusetzen, kann diese so mittels Monte-Carlo-Schätzung unter sehr allgemeinen Voraussetzungen angenähert werden.

Efron ... zeigt das sogenannte Bootstrap Verfahren statistisch exakt ist und neben den zahlreichen Anwendungsgebieten überraschend gute Eigenschaften haben kann.

Einziger Nachteil ist der zu leistende Rechenaufwand, allerdings wird die Rechenleistung von Computern immer besser und günstiger.

Bootstrap Verfahren sind sehr Einfach zu implementieren und liefern somit eine wunderbare Alternative gegenüber analytischen Verfahren, um die Verteilung einer Stichprobe zu bestimmen

beliebtes anwendungsgebiet sind konfidenz intervalle für den schätzer einer zufallsvariable...

Bootstrap-Ansätze zur Bestimmung von Konfidenz-Intervallen sind bereits gut bekannt und weit verbreitet.

Implementierungen sowohl der analytischen Ansätze als auch Bootstrap-Methoden sind eigentlich in allen gängigen statistischen Analyse-Tools vorhanden.

hat man statt einer jedoch mehrere zufallsvariablen, ist eine punktweise schätzung nicht sehr aussagekräftig...

stattessen benötigt man eine simultane berechischätung welche die schätzer aller zufallsvariablen mit gegebener wahrscheinlichkeit enthält, man spricht dann von konfidenzband

Methoden zu Bestimmung von Konfindenz-Bändern wurden bisher wenn überhaupt, eher in der Literatur behandelt.

Gerade Bootstrap finden noch kaum Anwendung bei der Bestimmung von Konfidenz-Bändern.

1.2 Zielsetzung

Ziel dieser Arbeit ist es einige solcher Ansätze vorstellen und im Kontekt von OMNeT++ umsetzten.

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

1.3 Aufbau der Arbeit

In Kapitel 2 werden dazu zuerst ein paar Grundlagen besprochen.

Die darauf folgenden Abschnitte sind in die drei Hautschwerpunkte der Arbeit unterteilt, Vorstellung der Algorithmen, Implementierung und Auswertung.

Schließlich folgt in Kapitel 6 noch ein Fazit. -

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

Kapitel 2

Grundlagen

Zuerst besprechen wir ein paar Grundlagen, die notwendig für die folgenden Kapitel sind

2.1 Simulationsstudien

Eine wesentliche Aufgabe der Statistik ist das Analysieren von realen System, um exakte Informationen zu relevanten Fragen zu erhalten.

Dabei kann es sich zum Beispiel um die Optimierung eines Herstellungsprozesses handeln.

Der Prozess in diesem Fall wird System genannt und Ziel der Analyse ist das Verhalten des Systems unter Veränderung der Eingabeparameter zu bestimmen.

Es gibt verschiedene Wege das Verhalten eines Systems zu studieren, Law... gibt einen guten Überblick zu diesem Thema.

Verschiedene Möglichkeiten ein System zu studieren werden von Law in folgender Grafik dargestellt.

Grundsätzlich kann man unterscheiden zwischen Experimenten an dem echten System und an einem Model des Systems.

Experimente an dem echten System würden bedeuten, dass man die physikalischen Gegebenheiten des System nach belieben verändern kann und die Daten für die Analyse anhand des realen Systems erheben kann

Dieser Ansatz ist immer erstrebenswert, da Unsicherheiten in der Wahl des Modells komplett eliminiert werden.

Allerdings ist dies nur in den seltensten Fällen möglich, da man in der Regel nur begrenzte Material- oder Geldressourcen zur Verfügung hat oder das System durch die Änderung zu sehr gestört wird.

Die am meisten Eingesetzte Variante ist die Simulation, bei der Versucht wird das reale Model so genau wie möglich mathematisch abzubilden und durch Computer gestützte Simulation beliebig viele Daten zu erheben.

Die Versuche der Simulation sind iid was für die späteren Methoden ein wesentlicher Faktor ist

Ein Modell welches das reale Modell repräsentiert wird Simulationsmodell genannt und kann beliebig komplex sein.

In unserem Fall betrachten wir ...

Ein einfaches Simulationsmodell welches eine ...Queue repräsentiert ist das M/M/1-Modell, diese sind sehr Verbreitet in der Statistik und gut bereits recherchiert.

Law gibt eine gute Einführung und unser Szenario orientiert sich stark daran.

Nachdem die Daten durch die Simulation erhoben wurden, wollen wir das Verhalten des Systems unter Veränderung der Parameter (etwa. der arivalrate oder servicetime) analysieren und durch eine Formel angeben.

Solch eine Formel die wiederum das Simulationsmodell vereinfacht repräsentiert ist auch ein mathematisches Modell und wird Metamodell genannt (bzw. Regressions Metamodell falls...)

Barton gibt eine gute Einführung für die Bestimmung solcher Metamodelle

Wir nehmen an, dass das Simulationsmodell das reale Modell korrekt repräsentiert, das Metamodell ist dann ... in Abhängigkeit von einem Parametervektor

Die Berechnung der Regressionsfunktion und Analyse der Methoden ist für uns primäres ziel der Simulationsstudie

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

2.1.1 M/M/1-Modell

Barton, R.R., 1998. Simulation metamodels:

Law, A.M. and Kelton, W.D. 1991. Simulation modeling and analysis:

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

- 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

2.2 Parameter-Schätzer

Hat man ein mathematisches Modell (bzw. regression metamodell) bestimmt, welches die Verteilung einer Datenmenge beschreiben soll, ist die Frage nun, was ist der wahre Wert des Parameters unter der Annahme, dass das Modell korrekt ist?

Der Parameter ist eine Zufallsvariable, die unter sehr allgemeinen Voraussetzungen multivariat normalverteilt um den wahren Wert ist.

Bei der linearen Regression wir oft die Kleinste-Quadrate-Methode verwendet.

Chernick, M.R. 1999.Bootstrap Methods, A Practitioners Guide 3.1.1.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans 9.1.

2.2.1 Kleinste-Quadrate-Schätzer

2.2.2 Maximum-Likelihood-Schätzer

- 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

2.3 GoF-Tests

2.3.1 Kolmogorov-Smirnov Test

2.3.2 Anderson-Darling Test

2.3.3 Chi-Quadrat Test

2.4 Konfidenzintervalle

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

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

- 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 deltamethode 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

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

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

2.5 Konfidenzbänder

Banks, J. 1998. Handbook of simulation: - 7.4 Multivariate Estimation

- 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

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

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

2.6 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

Sobald ein Regressionsmodell den Daten angepasst wurde ist es üblich, ein Konfidenz-Interval anzugeben, welches die Genauigkeit der Regressionsfunktion anzeigt.

Für die einzelnen Zufallsvariablen ... sind solche Bereichsschätzer sehr einfach zu bestimmen.

Der mittels Maximum-Likelihood-Methode geschätzte Parameter ... für die Standardabweichung liefert bereits das Konfidenz-Intervall ...

2.7 Resampling Verfahren

2.7.1 Jackknife

2.7.2 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

Banks, J. 1998. Handbook of simulation: - 7.2.3: Quantile Estimation

Was sind Voraussetzungen, die durch Bootstrap abgelöst werden?

Es werden im wesentlich zwei Ansätze vorgestellt, die Resampling einsetzen, um Konfidenzbänder zu bestimmen.

Der erste Ansatz Setzt Bootstrap ein, um

3.1 Standard Bootstrap

3.2 Bootstrapping der Residuen

3.3 Parametrsches Bootstrap

3.4 Wild Bootstrap

3.5 Bayes'sches Bootstrap

3.6 Resampling von $\partial R(\alpha)$

Banks, J. 1998. Handbook of simulation: - 5: Random Variate Generation (S. 143 ff)

Kapitel 4

Implementierung

4.1 OMNeT++

4.1.1 Überblick

4.1.2 Simulationen

4.2 Parameterstudien in OMNeT++

4.2.1 Datenerfassung

4.2.2 Auswertung

4.2.3 Darstellung

4.3

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 27. Februar 2021

Muster Mustermann