Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik

Dr. Uwe Ligges
M.Sc. Daniel Horn
M.Sc. Hendrik van der Wurp
Steffen Maletz

Übung zur Vorlesung

Computergestützte Statistik

Wintersemester 2018/2019

Übungsblatt Nr. 10

Abgabe ist Montag der 07.01.2019 an CS-abgabe@statistik.tu-dortmund.de oder Briefkasten 138

Ankündigung: Dies ist der Weihnachtszettel – ein Bonuszettel für Sie. Die erreichten Punkte zählen als Bonuspunkte und dienen lediglich dazu, ihre Gesamtpunktzahl am Ende des Semester zu verbessern. Die Punkte dieses Zettel zählen nicht, un die geforderten 50% der Punkte zu erreichen.

Aufgabe 1 **(4 Punkte)**

Implementieren Sie eine Evolutionsstrategie zur Variablenselektion.

a) (1.5 Punkte) In der Vorlesung haben Sie das grundlegende Prinzip evolutionärer Algorithmen kennen gelernt. Implementieren Sie einen Evolutionären Algorithmus zur Optimierung einer Funktion $f : \{0, 1\}^n \to \mathbb{R}$. Verwenden Sie 1-Punkt-Crossover zur Rekombination und Bit-Mutation. Eingabeparameter Ihrer Implementierung sollten sein:

   – Die Zielfunktion f sowie zugehörige Boxconstraints `lower` und `upper`.

   – Die Populationsgröße μ und die Anzahl Nachfahren λ, d.h. implementieren Sie bitte einen $(\mu + \lambda)$ EA

b) (0.5 Punkte) Testen Sie Ihre Implementierung an der Funktion $f(x) = \sum_{i=1}^{n} x_i$.

Sie haben in Kapitel 2 (sowie evtl. in der Vorlesung Lineare Modelle) lineare Modelle zur modellierung statistischer Zusammenhänge kennengelernt. Dabei wurden angenommen, dass stets sämtliche Variablen einen Einfluss auf die Zielvariable haben. Praktisch haben nicht alle Variablen einen Einfluss. Viel mehr kann die Hinzunahme von Störvariablen ohne echten Einfluss die Güte eines Modells deutlich verschlechtern. Die Güte des Modell lässt sich dabei z.B. über das adjustierte R^2 messen, welches mittels `summary(lm(...))$adj.r.squared` bestimmen lässt. Daher ist es in der Praxis häufig notwenig, eine relevante Untermenge von Variablen auszuwählen. Dies lässt sich als ein binäres Optimierungsproblem auffassen. Sei dazu weiterhin $f : \{0, 1\}^n \to \mathbb{R}$. Ist der i.te Eintrag von x 1, so wird die i.te Variable verwendet um das Modell zu bestimmen, ist der Eintrag 0, dann wird die Variable nicht verwendet. Der Wert von f entspricht dem adjustierten R^2 des angepassten Modells.

c) (0.5 Punkte) Betrachten Sie den Datensatz `fahrrad.txt`. Lesen Sie diesen ein, entfernen Sie Boebachtungen mit `NAs` sowie die Variable Datum und stellen Sie ein lineares Modell mit sämtlichen verbliebenen Einflussgrößen für Die Zielvariabe `comptage` auf. Geben Sie das adjustierte R^2 an.

d) (0.5 Punkte) Verwenden Sie die Funktion `step`, um eine optimale Variablenmenge zu bestimmen. Lesen Sie zunächst die entsprechende Hilfeseite von `step`.

e) (1 Punkt) Verwenden Sie Ihren in Aufgabenteil a) implementieren EA zur Bestimmung einer optimalen Variablenmenge. Vergleichen Sie das Ergebnis mit dem aus d).

Aufgabe 2 **(4 Punkte)**

Finden Sie heraus, was die beiden Funktionen in der Datei `schlechter_R_Code.R` berechnen. Pro Funktion können 2 Bonuspunkte erreicht werden. Gehen Sie dazu wie folgt vor:

a) Kopieren Sie die Funktionen in Ihre Abgabe.

b) Formatieren Sie den Quellcode entsprechend den bekannten Regeln neu. Achten Sie insbesondere auf sinnvolle Variablennamen.

c) Versuchen Sie anhand der auf die Eingabeparameter angewendeten Funktionen / Operationen zu erraten, welchen Typ und welche Struktur diese haben. Wenn auf den Parameter `a` der Funktion beispielsweise die Funktion `ncol` angewendet wird, dann ist `a` mit hoher Wahrscheinlichkeit eine Matrix oder ein Dataframe. Dokumentieren Sie Ihre Vermutung, indem Sie die Funktion mit einem Dokumentationskopf versehen.

d) Beheben Sie mögliche Effizienz-Probleme der Funktion. (Die Funktionen sind teilweise sehr *dumm* und ineffizient implementiert, dies soll behoben werden.)

e) Versuchen Sie den implementierten Algorithmus zu erkennen. Führen Sie hierzu die Funktion ggf. mit verschiedenen Eingaben aus und beobachten Sie die Rückgabewerte.

Aufgabe 3 **(4 Punkte)**

In der Datei `blackbox.RData` finden Sie die 4 Funktionen `f1`, `f2`, `f3`, `f4`. Minimieren Sie diese.

- `f1`: $[-10, 10]^{200} \to \mathbb{R}$ ist konvex.
- `f2`: $[-10, 10]^{10} \to \mathbb{R}$ ist stochastisch, d.h. wiederholte Auswertungen des gleichen Punktes führen zu unterschiedlichen Ergebnissen.
- `f3`: $[-10, 10]^{5} \to \mathbb{R}$ hat viele lokale Optima.
- `f4`: $\{-1000, -999, ..., 1000\}^{10} \to \mathbb{R}$, d.h. die Funktion erwartet ganze Zahlen als Eingabe.

Überlegen Sie sich jeweils, wie Sie die einzelnen Funktionen am besten optimieren können. Sie dürfen dazu sämtliche auf `CRAN`-Paketen verfügbare Pakete benutzen. Einen guten Einstieg bietet die Funktion `optim`. Beschreiben Sie jeweils Ihr Vorgehen. Es ist wichtiger, dass Sie sich sinnvolle Gedanken über die Lösung der einzelnen Probleme machen (und diese aufschreiben), als dass Sie das Optimum tatsächlich exakt bestimmen.