Technische Universität Dortmund
Fakultät Statistik
Lehrstuhl Computergestützte Statistik

Uwe Ligges
Marieke Stolte
Luca Sauer
Rudi Zulauf

Übung zur Vorlesung

Computergestützte Statistik

Wintersemester 2022/2023

Übungsblatt Nr. 11

Abgabe ist Dienstag, der 10.01.2023 bis 08:00 Uhr im Moodle

Das Team der Computergestützten Statistik wünscht den Teilnehmer*innen eine ruhige Weihnachtszeit und einen guten Rutsch ins neue Jahr!

Aufgabe 1 **(4 Punkte)**

In dieser Aufgabe betrachten wir die Studie zur logistischen Regression. Dieses Mal haben wir die Implementierung und Ausführung für Sie übernommen, die Ergebnisse der Studie finden Sie in der Datei `logreg_exp.RData`. Analysieren Sie diese Ergebnisse.

a) (0.5 Punkte) Werfen Sie einen 1. Blick in die Daten (mit der Funktion `summary`). Wie sehen die Daten aus? Gibt es Auffälligkeiten?

b) Beantworten Sie die im Skript gegebenen Fragen:

   – (0.5 Punkte) Welchen Einfluss haben b, n, κ, ϵ auf die Messgrößen?

   – (0.5 Punkte) Welcher Algorithmus verwendet die geringste Anzahl an Funktionsauswertungen? Warum?

   – (0.5 Punkte) Ist ein Algorithmus besser als ein anderer?

   – (0.5 Punkte) Korreliert die Anzahl Funktionsauswertungen mit der Güte der Lösung?

   Begründen Sie Ihre Antworten und fertigen Sie insbesondere passende Grafiken an, um Ihre Antworten zu belegen. Achten Sie darauf ordentliche Grafiken zu erstellen, das Aussehen der Grafiken werden wir mit 0.5 Punkten bewerten.

c) Bei der Durchführung der Studie sind einige Probleme aufgetreten. Diese haben wir zwar bereits für Sie gelöst – wir würden dennoch gerne einmal von Ihnen wissen, wie Sie diese Probleme gelöst hätten. Geben Sie Lösungsansätze zu den beiden folgenden Problemen an. Erläutern Sie zuvor jeweils, warum die beiden Punkte jeweils ein Problem darstellen:

   – (0.5 Punkte) Wie wird β^* bestimmt?

   – (0.5 Punkte) Wie verhindert man, dass die Daten linear trennbar sind?

Aufgabe 2 **(4 Punkte)**

a) (3 Punkte) Implementieren Sie eine Funktion, die den einfachen Simplexalgorithmus der Vorlesung umsetzt. Als Eingabe sollte Ihre Funktion die zu optimierende Funktion `f`, einen Vektor mit Startwerten `init`, die Simplexgröße `t` und die maximale Anzahl zu durchlaufender Iterationen `max.it` als Argumente erhalten.

b) (1 Punkt) Testen Sie die Implementierung mit der negativen Dichte einer bivariaten Normalverteilung

$$-f(x) = -\frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}\,(x-\mu)^T\,\Sigma^{-1}\,(x-\mu)\right).$$

Dabei seien $\rho > 0$ und $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ sowie $\mu_1 = \mu_2 = 0$ und $\sigma_1 = 1$, $\sigma_2 = \sqrt{2}$. Verwenden Sie $\rho = 0.2, 0.4, 0.8$. Stellen Sie die Simplexgröße so ein, dass sie für unterschiedliche Startwerte bis auf eine Toleranz von 0.05 and das Optimum $(0, 0)$ heran kommen.

Hinweis: Das Paket `mvtnorm` kann hilfreich sein.

Aufgabe 3 **(4 Punkte)**

Betrachten Sie das NL-KQ-Beispiel aus Kapitel 4.5.4 mit wahrem Koeffizientenvektor $(\theta_1, \theta_2) = (0.5, 2)$.

a) (1 Punkt) Stellen Sie die Zielfunktion auf. Verwenden Sie dazu die Datenpunkte aus der Datei `daten_blatt11.RData`.

b) (3 Punkt) Bestimmen Sie das Optimum für jedes Paar von Startwerten aus dem Gitter von Werten $[-3, -2.95, -2.9, ..., 3.9, 3.95, 4]^2$ mit sowohl BFGS als auch Nelder-Mead (Funktion `optim`). Visualisieren Sie jeweils, zu welchen Optimum die beiden Algorithmen konvergieren, zeichnen Sie auch die Konturlinien der Funktion ein. Erzeugen Sie also Karten ähnlich zu denen aus der Vorlesung. Volle Punkte gibt es nur für schöne Graphiken!