BTU Cottbus-Senftenberg
Lehrstuhl für Wahrscheinlichkeitsrechnung & Statistik
Lorenz Richter

7. September 2017

Angewandte mathematische Statistik

4. Aufgabenblatt

1. Aufgabe (Regression I)

Gegeben sei jeweils das lineare Modell $Y = X\beta + \epsilon$, mit $Y \in \mathbb{R}^n, \beta \in \mathbb{R}^p$ und $\epsilon \sim \mathcal{N}(0, \Sigma)$.

(a) Wir betrachten die lineare Regression und nehmen zunächst unkorreliertes Rauschen an, d.h. $\Sigma = \sigma^2 \mathbb{1}$. Simulieren Sie $n = 1000$ Datenpunkte $X_i \sim \mathcal{U}([0, 10])$, wählen Sie ein $\beta \in \mathbb{R}^2$ Ihrer Wahl und berechnen Sie Y. Schätzen Sie nun $\hat{\beta}$ und vergleichen Sie Ihr Ergebnis mit der R-Methode `lm()`. Plotten Sie die Daten sowie die Regressionsgerade.

(b) Erstellen Sie nun korreliertes Rauschen, etwa durch die Kovarianzmatrix $\Sigma = A^\top A + \mathbb{1}$ und der Funktion `mvrnorm()`, wobei $A = (a_{ij})$ mit $a_{ij} \sim \mathcal{U}([-1, 1])$, und wiederholen Sie die obigen Schritte.

(c) Wir betrachten nun wieder unkorreliertes Rauschen sowie den polynomiellen Zusammenhang $y = f(x) = 2x^3 + x^2 - 8x + 1$. Stellen die zugehörige Designmatrix auf und berechnen Sie $\hat{\beta}$. Vergleichen Sie Ihr Ergebnis mit dem einer linearen Regression.

2. Aufgabe (Regression II)

Laden Sie die Daten `regressiondata.csv` und versuchen Sie, die vierte Variable durch die zweite mittels linearer Regression zu erklären. Untersuchen Sie die Signifikanz des Zusammenhangs und interpretieren Sie Ihr Ergebnis.

3. Aufgabe (ANOVA)

Simulieren Sie Daten in drei Gruppen, und zwar jeweils $X_{i1}, \ldots, X_{in_i} \sim \mathcal{N}(\mu_i, \sigma^2)$ für $i = 1, 2, 3$ mit beliebig gewählten Gruppengrößen n_i und Mittelwerten μ_i. Testen Sie die Nullhypothese $H_0 : \mu_1 = \mu_2 = \mu_3$ mittels der Varianzanalyse, indem Sie die entsprechende Designmatrix aufstellen und den Schätzer $\hat{\beta}$ des linearen Modells berechnen. Mit diesem können Sie schließlich die Varianzen innerhalb und zwischen den Gruppen bestimmen und zu einem Schluss kommen. Vergleichen Sie Ihr Ergebnis mit der R-Funktion `anova()`.

4. Aufgabe (Logistische Regression)

Laden Sie den Datensatz `creditcardfraud.csv` und versuchen Sie, die Variable `class` $\in \{0, 1\}$ vorherzusagen.

(a) Standardisieren Sie hierfür zunächst Ihre Daten und unterteilen Sie sie in einen Trains- und einen Testdatensatz. Auf den Trainingsdaten berechnen Sie die Regressionskoeffizienten $\hat{\beta}$ mittels der R-Funktion `glm()` und dem zusätzlichen Parameter `family=binomial(link='logit')`. Mit $\hat{\beta}$ können Sie schließlich Wahrscheinlichkeitswerte für die Klassenzugehörigkeit der Testdaten ausrechnen und diese Werte mittels der ROC-Kurve in Bezug auf die wahren Klassenwerte evaluieren – hierfür ist die R-Bibliothek `ROCR` hilfreich.

(b) Wiederholen Sie die Aufgabe, indem Sie nun $\hat{\beta}$ mit einem selbst implementierten Gradientenverfahren schätzen.