Wahrscheinlichkeitstheorie und Statistik

Serie 9 - Lösungen

MC 9-1. Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilte Zufallsvariablen mit unbekanntem Parameter $\theta \in \Theta$. Welche Aussagen über einen Schätzer $\hat{\theta} = g(X_1, \ldots, X_n)$ sind immer korrekt? (Mehrere Antworten sind möglich.)

(a) $\hat{\theta}$ muss eine messbare Funktion der Stichprobe $(X_1, \ldots, X_n)$ sein.

(b) $\hat{\theta}$ muss erwartungstreu für θ sein.

(c) $\hat{\theta}$ ist selbst eine Zufallsvariable.

(d) Wenn $\hat{\theta}$ eine konstante Funktion ist, ist $\hat{\theta}$ kein gültiger Schätzer.

Lösung: Nur (a) und (c) sind wahr.

(a) Wahr, ein Schätzer muss eine messbare Funktion der Daten sein.

(b) Falsch, ein Schätzer muss nicht erwartungstreu sein, das ist eine zusätzliche Eigenschaft.

(c) Wahr, ein Schätzer ist eine Funktion von Zufallsvariablen und damit selbst eine Zufallsvariable.

(d) Falsch, auch eine konstante Funktion (z.B. immer 42) ist formal ein gültiger Schätzer, wenn auch möglicherweise ein schlechter.

MC 9-2. Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilte Zufallsvariablen mit unbekanntem Parameter $\theta \in \Theta$. Welche Ausdrücke sind formal gültige Schätzer für θ? (Mehrere Antworten sind möglich.)

(a) $\hat{\theta}(X_1, \ldots, X_n) = 42$

(b) $\hat{\theta}(X_1, \ldots, X_n) = X_1 + \sin(X_2)$

(c) $\hat{\theta}(X_1, \ldots, X_n) = \theta + 1$

(d) $\hat{\theta}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$

Lösung: (a), (b) und (d) sind wahr.

(a) Wahr, eine konstante Funktion der Daten ist ein gültiger aber möglicherweise recht schlechter Schätzer.

(b) Wahr, eine beliebige messbare Funktion der Daten ist ein gültige Schätzer.

(c) Falsch, ein Schätzer darf nicht direkt von θ abhängen, nur von den beobachteten Daten.

(d) Wahr, der Stichprobenmittelwert ist ein klassischer gültiger Schätzer für den Erwartungswert.

MC 9-3. Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilte Zufallsvariablen. Angenommen, weniger als $n/2$ Beobachtungen X_i werden kontaminiert, indem sie durch einen sehr großen Wert M ersetzt werden.

Welche der folgenden Aussagen sind korrekt, wenn M gegen unendlich geht? (Mehrere Antworten sind möglich.)

(a) Der Stichprobenmittelwert bleibt beschränkt.

(b) Der Stichprobenmittelwert bleibt unverändert.

(c) Der Stichprobenmedian bleibt beschränkt.

(d) Der Stichprobenmedian bleibt unverändert.

Lösung: Nur (c) ist wahr.

(a) Falsch, ein großes M lässt $\bar{X}_n$ ohne Grenze wachsen.

(b) Falsch, der Mittelwert der kontaminierten Daten unterscheidet sich vom Mittelwert der unkontaminierten Daten.

(c) Wahr, der Median bleibt bis zu einer Kontamination von 50% unverändert.

(d) Falsch, wenn kleine Werte durch einen großen Wert M ersetzt werden, kann sich der Median tatsächlich verschieben – selbst wenn weniger als die Hälfte der Werte verändert wurden.

Aufgabe 9-4. Seien $X_1, \dots, X_n$ unabhängige, Zufallsvariablen mit $X_i \sim \mathcal{N}(\theta\alpha_i, 1)$, wobei $\alpha_i \neq 0$ bekannte Parameter sind, aber $\theta \in \mathbb{R}$ unbekannt ist. Bestimmen Sie den Maximum-Likelihood-Schätzer für θ.

Lösung: Die Randdichten sind $f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x_i - \theta\alpha_i)^2)$. Wegen Unabhängigkeit ist die Likelihood-Funktion

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(x_i - \theta\alpha_i)^2\Big).$$

Wir berechnen

$$\log L(x_1, \dots, x_n; \theta) = \sum_{i=1}^n \Big(-\frac{1}{2}\log(2\pi) - \frac{1}{2}(x_i - \theta\alpha_i)^2\Big).$$

Wir wollen das über θ maximieren und berechnen deshalb

$$\frac{\partial}{\partial\theta} \log L(x_1, \dots, x_n; \theta) = -\sum_{i=1}^n (x_i - \theta\alpha_i)(-\alpha_i) = \sum_{i=1}^n (x_i - \theta\alpha_i)\alpha_i.$$

Wegen $\sum_{i=1}^n \alpha_i^2 > 0$ gilt

$$\sum_{i=1}^n (x_i - \theta\alpha_i)\alpha_i = 0 \iff \sum_{i=1}^n x_i\alpha_i - \theta\sum_{i=1}^n \alpha_i^2 = 0 \iff \theta = \frac{\sum_{i=1}^n x_i\alpha_i}{\sum_{i=1}^n \alpha_i^2}.$$

Ferner gilt

$$\frac{\partial^2}{\partial\theta^2} \log L(x_1, \dots, x_n; \theta) = -\sum_{i=1}^n \alpha_i^2 \leq 0.$$

Somit ist die Funktion $\mathbb{R} \to \mathbb{R}, \theta \mapsto \log L(x_1, \ldots, x_n; \theta)$ konkav und

$$\widehat{\theta}(x_1, \ldots, x_n) = \frac{\sum_{i=1}^n x_i \alpha_i}{\sum_{i=1}^n \alpha_i^2}$$

ist der Maximierer.

Aufgabe 9-5 Schreiben Sie eine Python-Implementierung der logistischen Regression. Die Aufgabe ist, eine Modellfunktion

$$p : \mathbb{R} \to \mathbb{R}, \qquad p(x) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x))}$$

mit Parametern β_0 und β_1 zu bestimmen, welche für gegebene Daten $(X_1, Y_1), \ldots, (X_n, Y_n)$ die folgende logistische Log-Likelihood maximiert:

$$\mathrm{LL}(\beta_0, \beta_1) = \sum_{i=1}^n \left(Y_i \log p(X_i) + (1 - Y_i) \log(1 - p(X_i))\right).$$

Lösung:

```
import numpy as np
from scipy.optimize import minimize
import matplotlib.pyplot as plt

def sigmoid(z):
    return 1 / (1 + np.exp(-z))

def neg_log_likelihood(params, X, Y):
    beta0, beta1 = params
    p = sigmoid(beta0 + beta1 * X)
    # Add a small epsilon to avoid log(0)
    epsilon = 1e-10
    log_likelihood = np.sum(Y * np.log(p + epsilon) + (1 - Y) * np.log(1 - p + epsilon))
    return -log_likelihood  # We minimize negative log-likelihood

def fit_logistic_regression(X, Y):
    # Initial guess for beta0 and beta1
    initial_params = np.array([0.0, 0.0])
    result = minimize(neg_log_likelihood, initial_params, args=(X, Y), method='BFGS')
    beta0, beta1 = result.x
    return beta0, beta1

np.random.seed(0)
n = 100
X = np.random.uniform(-3, 3, n)
true_beta0 = -0.5
true_beta1 = 2.0
p_true = sigmoid(true_beta0 + true_beta1 * X)
Y = np.random.binomial(1, p_true)
```

```
beta0_est, beta1_est = fit_logistic_regression(X, Y)
print(f"Estimated parameters: beta0 = {beta0_est:.4f}, beta1 = {beta1_est:.4f}")

x_plot = np.linspace(min(X), max(X), 300)
p_est = sigmoid(beta0_est + beta1_est * x_plot)

plt.scatter(X, Y, label='Data', alpha=0.5)
plt.plot(x_plot, p_est, color='red', label='Fitted Logistic Curve')
plt.xlabel('X')
plt.ylabel('Probability of Y=1')
plt.title('Logistic Regression Fit')
plt.legend()
plt.grid(True)
plt.show()
```