Datamining - Zusammenfassung, Fragen und Notizen

Julius Hülsmann

11. Januar 2016

Inhaltsverzeichnis

Diese Zusammenfassung beruht auf der Vorlesung Datamining; Wintersemester 2015/16 gehalten von Thomas Hermann und nebensächlich auf Informationen des Bereiches Statistik am Psychologischen Institut der Universität Mainz und der Vorlesung Statistik der Universität Bielefeld, Wintersemester 15/16 gehalten von Frau Gentz.

1 Kapitel 1 - Einführung

2 Kapitel 2 - Nützliche Konzepte aus der Statistik

Wie wird Zufall von Muser unterschieden?
Fragestellung: Gilt die Nullhypothese H_0?

1. Berechne Schätzer/Prüfgröße s
2. Berechne die Wahrscheinlichkeitsverteilung von s unter Annahme, dass die Nullhypothese H_0 zutrifft.
3. Berechne den Wert für s, der sich aus den Messdaten ergibt.
   $s < 0.05 \Rightarrow$ verwerfe H_0

2.1 Detektion systematischer Abweichungen

Beispiel 2.1 (Stichproben haben gleichen Erwartungswert)**.**
Gegeben *seien*

- *zwei endliche Stichproben* $\{x_i^A\}_{i=1}^{N_A}, \{x_i^B\}_{i=1}^{N_B}$.

zweier Wahrscheinlichkeitsmaße $\mathbb{P}_A$ *und* $\mathbb{P}_B$ *auf den Ereignisräumen* $(\Omega, \mathcal{F})$ *für* Ω *entweder abzählbar, dann* $\mathcal{F} = P(\Omega)$ *oder* $\mathcal{F} = \mathcal{B}_\Omega$*. Zugehörige Wahrscheinlichkeitsdichten* P_A *und* P_B*. Tatsächliche Verteilung ist unbekannt.*

Frage: *Sind die zugrundeliegenden Wahrscheinlichkeitsmaße Normalverteilt mit gleichem Erwartungswert?*
Annahme: die Varianz beider Verteilungen ist gleich $\sigma^2(P_A) = \sigma^2(P_B)$*. Diese Annahme ist zu verifizieren mithilfe des sogenannten F-Testes*

Bezeichnungen:
empirischer Mittelwert *(Erwartungswert der Stichprobe)*

$$\hat{\mu}_A := \frac{1}{N_A}\sum_{i=1}^{N_A} x_i^A, \quad \hat{\mu}_B := \frac{1}{N_B}\sum_{i=1}^{N_B} x_i^B,$$

empirische Standardabwichungen

$$\sigma(\hat{\mu}^A) := \sqrt{\frac{1}{N_A}\sum_{i=1}^{N_A}(x_i^A - \mu^A)^2}, \quad \sigma(\hat{\mu}^B) := \sqrt{\frac{1}{N_B}\sum_{i=1}^{N_B}(x_i^B - \mu^B)^2}$$

tatsächlicher Erwartungswert

$$\mu_A, \quad \mu_B, \quad \mu_{AB}$$

tatsächliche Standardabweichung

$$\sigma(P_A), \quad \sigma(P_B).$$

Schritt 0 *Stelle Nullhypothese auf:*

$$\mu_A = \mu_B.$$

Schritt 1 *Bestimme Prüfgröße. Folgende Eingenschaften sind relevant:*

- *Unabhängigkeit von der Varianz bzw. Standardabweichung,*
- *Unabhängigkeit von der Stichprobengröße.*

$$t := \frac{|\hat{\mu}_A - \hat{\mu}_B|}{\hat{\sigma}_{err}} \tag{1}$$

$$\sigma_{err} \approx \sqrt{\frac{\sigma^2(\hat{\mu}_A - \hat{\mu}_B)}{N_A + N_B}}, \tag{2}$$

da Unabhänigkeit von der Varianz gewährleistet sein soll, die beim Stichprobenumfang von $N := N_A + N_B$ durch Division mit N gewichtet wird.

$$\sigma^2_{err} \approx \frac{\sigma^2(\hat{\mu}_A - \hat{\mu}_B)}{N_A + N_B} \tag{3}$$

$$= \frac{1}{N_A + N_B}\left(\sigma^2(\hat{\mu}_A) + \sigma^2(\hat{\mu}_B)\right) \tag{4}$$

$$\approx \frac{1}{N_A + N_B}\left(\frac{\sigma^2(P_A)}{N_A} + \frac{\sigma^2(P_B)}{N_B}\right) \tag{5}$$

$$\approx \frac{1}{N_A + N_B}\left(\sigma^2_{AB} * (\frac{1}{N_A} + \frac{1}{N_B})\right) \tag{6}$$

$$= \frac{1}{N_A + N_B}\left(\frac{\sum_{i=1}^{N_A}(x_i^A - \hat{\mu}_A)^2 + \sum_{i=1}^{N_B}(x_i^B - \hat{\mu}_B)^2}{N_a + N_B - 2}\right) \tag{7}$$

Schritt 2 *Bestimmung der Wahrscheinlichkeit*

$$\mathbb{P}(t|H_0 \textit{ trifft zu }),$$

also die Wahrscheinlichkeit, dass falls die H_0-Hypothese zutrifft, die Prüfgröße t aus der Stichprobe erhalten wird.