

Datamining - Zusammenfassung, Fragen und Notizen

Julius Hülsmann

11. Januar 2016

Inhaltsverzeichnis

1	Kapitel 1 - Einführung	2
2	Kapitel 2 - Nützliche Konzepte aus der Statistik	3
2.1	Detektion systematischer Abweichungen	3
2.2	χ^2 -Verteilung und Student- t -Verteilung	5
2.3	Test auf unterschiedliche Verteilung	5
2.3.1	Der χ^2 -Test	5
2.3.2	Der Kolmogorov-Smirnov-Test	5
2.4	Detektion von Abhängigkeiten zwischen Zufallsvariablen . . .	5
2.5	Lineare Korellation	5
2.5.1	Geometrische Interpretation als Winkelmaß	5
2.5.2	Geometrische Interpretation als Optimierungsfehler . .	5
2.5.3	Statistische Interpretation von r	5
2.6	Nichtparametrische Korrelation	5
2.6.1	Spearman's Rang-Korrelationskoeffizient	5
2.6.2	Kendall's τ	5
2.7	identifikation von Ausreißern	5
2.7.1	Ausreißerererkennung	5
2.7.2	Ausreißerbehandlung	5

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.
 $s < 0.05 \Rightarrow$ verwirfe 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 (Ω, \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 Standardabweichungen

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

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 Eigenschaften sind relevant:

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

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

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

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

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

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

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

$$\approx \frac{1}{N_A + N_B} \left(\sigma_{AB}^2 * \left(\frac{1}{N_A} + \frac{1}{N_B} \right) \right) \quad (6)$$

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

Schritt 2 Bestimmung der Wahrscheinlichkeit

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

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

2.2 χ^2 -Verteilung und Student- t -Verteilung

2.3 Test auf unterschiedliche Verteilung

2.3.1 Der χ^2 -Test

2.3.2 Der Kolmogorov-Smirnov-Test

2.4 Detektion von Abhängigkeiten zwischen Zufallsvariablen

2.5 Lineare Korellation

2.5.1 Geometrische Interpretation als Winkelmaß

2.5.2 Geometrische Interpretation als Optimierungsfehler

2.5.3 Statistische Interpretation von r

2.6 Nichtparametrische Korrelation

2.6.1 Spearman's Rang-Korrelationskoeffizient

2.6.2 Kendall's τ

2.7 identifikation von Ausreißern

2.7.1 Ausreißererkennung

2.7.2 Ausreißerbehandlung