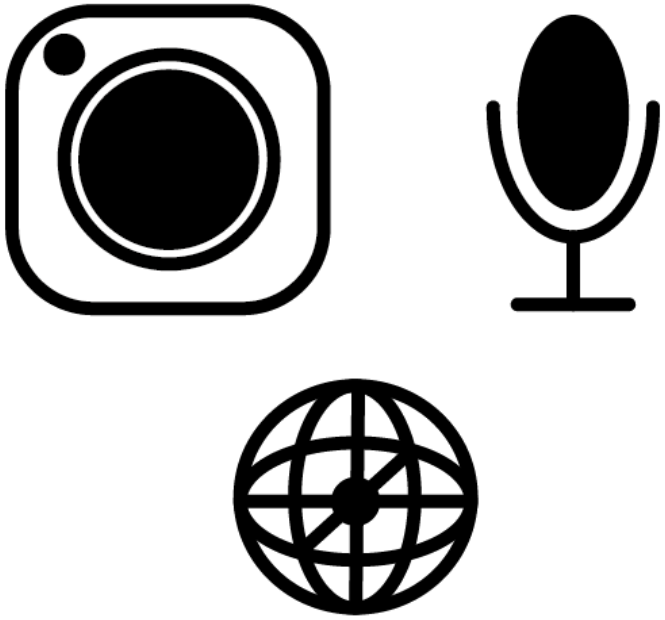


Sensordatenverarbeitung **(ab 13.1.25)**

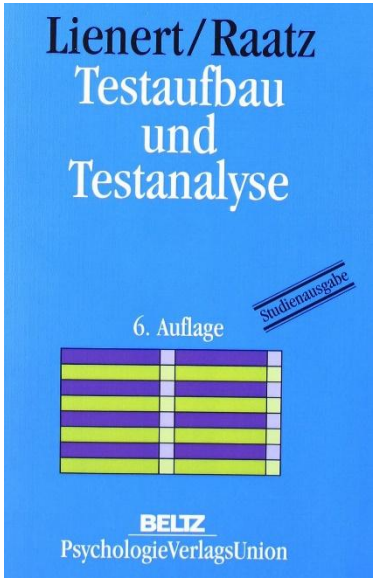
ENTWICKLUNG UND EVALUATION VON SDV-SYSTEMEN (12B)

Überblick

- Entwicklung von SdV-Systemen
- **Evaluation von SdV-Systemen**
- Beschreibende Statistik und Metriken
- Statistische Tests

Literatur

- Gustav Lienert
  - Testaufbau
  - Design von Experimenten
  - Testanalyse
    - Objektivität, Reliabilität, Validität



- Jürgen Bortz
  - Deskriptive Statistik
  - Wahrscheinlichkeitstheorie
  - Stichproben
  - Hypothese
  - Varianzanalyse
  - Multivariate Methoden

Zufallswahrscheinlichkeit

- Welcher Klassifikator ist besser (genauer)?
  - Klassifikator 1: 90% Akkuratheit bei Klassifikation von „mentale Aufgabe versus Ruhezustand“
  - Klassifikator 2: 30% Akkuratheit bei Klassifikation von 20 Handgesten

- A priori-Wahrscheinlichkeit (Ursprungswahrscheinlichkeit) wichtig
  - Bei Gleichverteilung liegt Zufallswahrscheinlichkeit bei:
    - Klassifikator 1: 50%
    - Klassifikator 2: 5%

  - Gleichverteilung bedeutet: alles Datensamples sind gleichmäßig auf alle Klassen verteilt, also
    - Beispiel 1: genauso viele Samples für „mentale Aufgabe“ wie für „Ruhezustand“ etwa 50 Datensamples x „mentale Aufgabe“ und 50 Datensamples „Ruhezustand“
    - Beispiel 2: für jede der 20 Handgesten liegt dieselbe Anzahl an Samples vor etwa 10 Datensamples „Handgeste 1“, 10 Datensamples „Handgeste 2“, …. … 10 Datensamples „Handgeste 20“

Zufallswahrscheinlichkeit

- Welcher ist besser?
  - Klassifikator 1: Findet relevante Segmente mit Akkuratheit 70%
  - Klassifikator 2: Warnt vor Sturm mit 95% Akkuratheit

- A priori Wahrscheinlichkeit: hier keine Gleichverteilung
  - Relevante Segmente: 40% der Segmente
  - Stürme: 3% der Tage

Konfusionsmatrix

- Andere Darstellung von Klassifikationsergebnissen
- Zeigt an, wie viele Instanzen der Referenz-Klasse welcher Klasse im Test zugeordnet wurden (Klassifikation)

Klassifikation

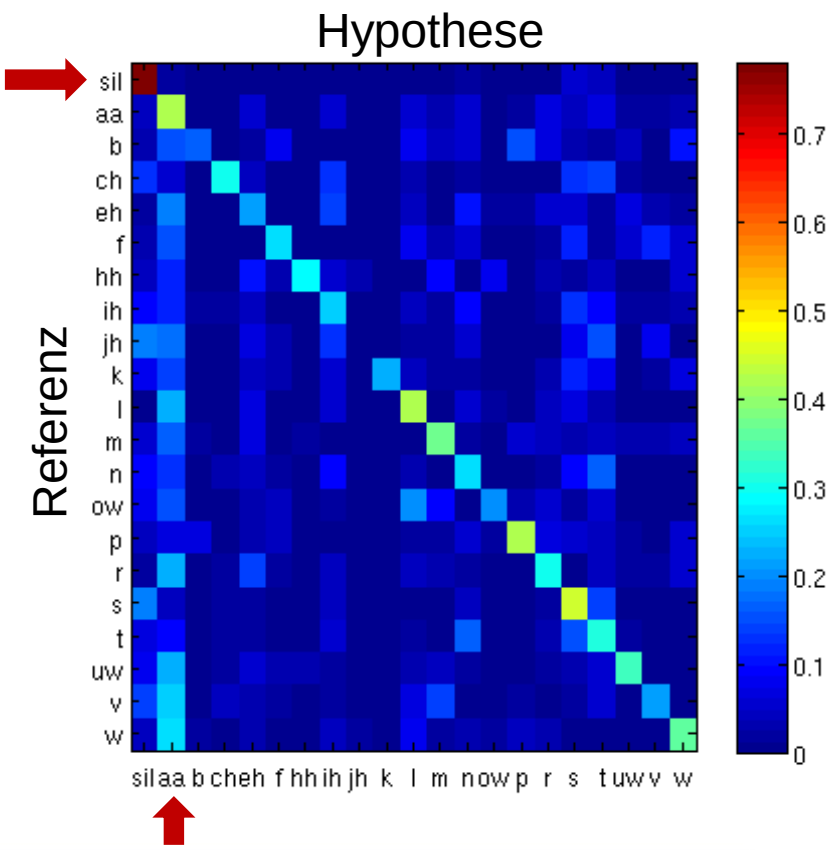
Referenz

	Klasse 1	Klasse 2
Klasse 1	26	4
Klasse 2	5	27

Konfusionsmatrix

- Gibt wichtige Informationen an, die eine detailliertere Analyse zulassen, als nur die Bestimmung der Akkuratheit
- Beispiel hier: häufige (seltene) Vertauschungen in einzelnen Klassen
  - Siehe z.B. Referenz „sil“ wird selten falsch klassifiziert
  - Siehe z.B. Modell /aa/ wird häufig (fälschlich) erkannt

    (nennt man dann ein „sink model“)



Beispiel Konfusionsmatrix

- Drei Klassenproblem
- Referenz: 1-2-3-1-2-3-1-2-3-1-2-3-1-2-3-1-2-3-1-2-3-1-2-3
- Klassifikator: 3-2-1-1-2-2-1-1-2-2-2-2-1-2-1-2-2-2-1-2-2-1-2-1
- Akkuratheit: 50% (12/24);
- A priori (Ratewahrscheinlichkeit): 33,3 %
- Vorgehen:
  - Referenz=1 auszählen:
    - 5 mal richtig als Klasse 1 klassifiziert
    - 2 mal als Klasse 2 klassifiziert, 1 mal als Klasse 3 klassifiziert
    - …

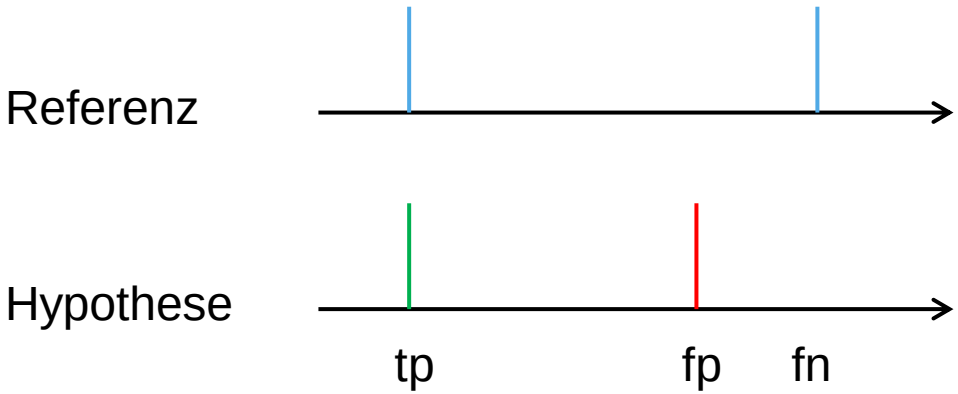
Referenz

Klassifikation	1	2	3
1	5	1	3
2	2	7	5
3	1	-	-

Metrik: Precision und Recall

- Messung der Qualität einer Erkennung
  - Vergleich mit Referenz (Ground Truth)

- Korrekt gefundene Targets (tp – true positive)
- Korrekt gefundene nicht-Targets (tn – true negative)
- Falsche gefundene Targets (fp – false positive)
- Nicht gefundene Targets (fn – false negative)



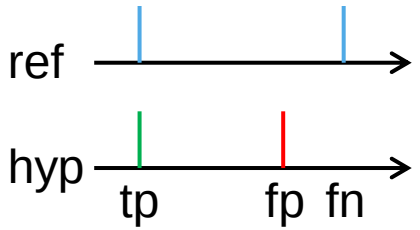
Alle anderen Zeitpunkte sind true negative (tn)

Metrik: Precision und Recall

- Precision (Deutsch: Genauigkeit)
  - Wie viele der gefundenen Targets sind auch echte Targets

$$precision = \frac{tp}{tp + fp}$$

- Recall (Deutsch: Trefferquote)
  - Wie viele korrekte Targets wurden gefunden



$$recall = \frac{tp}{tp + fn}$$

Gewichtung von Precision/Recall

- Gewichtung von Precision und Recall ist abhängig vom Task
- Was ist besser, zu oft reagieren (FP) oder Ereignisse verpassen (FN)
- Beispiel: Test auf HIV
  - **Richtig positiv (tp)**: Der Patient ist krank, und der Test hat dies richtig angezeigt: 66.933 Fälle
  - **Falsch negativ (fn)**: Der Patient ist krank, aber der Test hat ihn fälschlicherweise als gesund eingestuft: 67 Fälle
  - **Falsch positiv (fp)**: Der Patient ist gesund, aber der Test hat ihn fälschlicherweise als krank eingestuft: 82.000 Fälle
  - **Richtig negativ (tn)**: Der Patient ist gesund, und der Test hat dies richtig angezeigt: 81 Mio Fälle
  - Wenn man nur den ELISA-Test durchführe, würden 82.000 Mensch in großer Sorge sein, obwohl sie gesund sind – weitere Tests (Western-Blot)

ELISA-Test	HIV positiv	HIV negativ	Summe
HIV-Test positiv	66.933	82.000	148 933
HIV-Test negativ	67	81.851.000	
Summe	67.000		82.000.000

Evaluation von SdV-Systemen

Vergleichende Evaluation

- Selten reicht nur eine Evaluation des eigenen Verfahrens
- Besser: Vergleich mit (sortiert nach wie empfehlenswert)
  - Etablierter Benchmark oder Community Challenge
  - Mit und ohne eine entwickelte Verbesserung eines Verfahrens
  - Konkurrenzverfahren mit verfügbarer Implementierung
  - Gold-Standard: ein (einfaches) meist langsames Verfahren, dass anerkannt die bestmögliche Lösung liefert (besonders als Vergleich für schnelle Verfahren)
  - Baseline (Blei-Standard): ein einfaches Verfahren, für das jedes ambitionierte Verfahren besser sein sollte
  - Konkurrenzverfahren selbst implementiert
    (Problem: Aufwand, vielleicht schlecht implementiert)

Statistische Analyse: Beispiel

- Beispiel:
  - Zufriedenheit Bewertung von 0 bis 5 für zwei Systeme A und B
  - Gesammelten Daten für die Systeme:
    - A: [2,3,1,1,0,2,3,…]
    - B: [3,4,3,2,1,2,4,…]

- Wie analysiert man welches System besser ist:

- Erster Versuch: Berechne den Mittelwert:

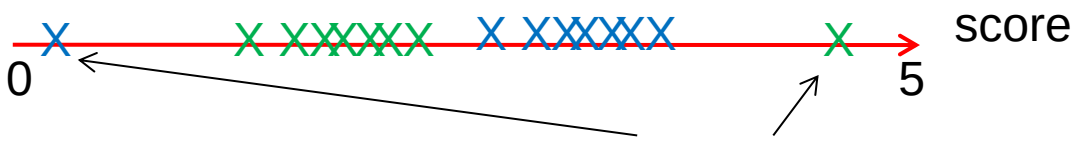
- Ist A besser als B?

Problem 1: Outliers & Stichproben Größe

- Vorstellung ist, dass dieses Bild …

- …durch dieses hervorgerufen wurde:

- Tatsächlich sind die Daten aber so:



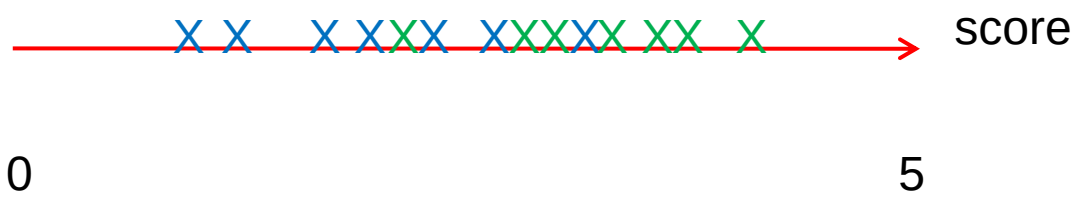
- Outlier kommen häufig in Biosignalen vor (z.B. verrauschte Daten)
- Große Auswirkungen besonders auf kleine Datensätze

Problem 2: Varianz

- Annahme: Wir filtern Outlier heraus
- Dieses Bild,

- Könnte also von dieser Population kommen:

- Was passiert, aber wenn Daten tatsächlich so aussehen:



- Hohe Varianz im Sample
- Wie verlässlich sind Schlüsse, die aus diesen Daten gezogen werden?

Problem 3: Unabhängigkeit

- Annahme: Outlier und Varianz sind gefiltert
- Stichproben Verteilung:

- Können wir nun schließen, dass B besser ist?
- Fehler können auch im Experimentdesign stecken
- Wenn A zum Beispiel immer zuerst getestet wurde
  - Probanden haben aus A gelernt und sind nun bei B erfahrener

  → Stichproben A und B sind nicht unabhängig

- Bei umgekehrter Reihenfolge dann vielleicht diese Verteilung: