

Konfidenzintervall

Aufgabe: Evaluierung eines automatischen Annotations-/Klassifikationsverfahrens

→ Evaluierung durch manuelle Bewertung einer Stichprobe

Auch ein Goldstandard (manuell annotiertes Datenset) kann als Stichprobe angesehen werden

Schätzung der Erfolgsrate in der Grundgesamtheit (die “wirkliche” Erfolgsrate) durch in der Stichprobe gemessene Erfolgsrate

Konfidenzintervall

Schätzung der Erfolgsrate in der Grundgesamtheit durch Erfolgsrate in der Stichprobe

- Wie verlässlich ist die Bewertung der Stichprobe?
- Darf man sie auf die Grundgesamtheit verallgemeinern?
- Kann man Stichproben verschiedener Größe vergleichen?

Berechnung des **Konfidenzintervalls** (Vertrauensbereich)

- P = Erfolgsrate in der Grundgesamtheit
- p = gemessene Erfolgsrate in Stichprobe
- n = Umfang der Stichprobe
- z_k = Konfidenzkoeffizient. 95%: $z_k = 1,96$
- Annäherung durch Normalverteilung:
$$p_u = p - z_k \sqrt{p(1-p)/n}$$
$$p_o = p + z_k \sqrt{p(1-p)/n}$$
- P liegt mit 95% Wahrscheinlichkeit im Bereich $[p_u; p_o]$

Beispiel

Zwei Programme beantworten Synonymfragen aus dem TOEFL

Programm A wird evaluiert auf 2.000 manuell annotierten Fragen. Genauigkeit: $p^A = 74,6\%$

Programm B stehen nur 100 annotierte Fragen zur Verfügung. Genauigkeit: $p^B = 77,0\%$

Ist B tatsächlich besser als A?

Beispiel

Programm A wird evaluiert auf 2.000 manuell annotierten Fragen. Genauigkeit: 72,6%

Programm B stehen nur 100 annotierte Fragen zur Verfügung. Genauigkeit: 77,0%

B:

$$z_k \cdot \sqrt{p^B(1-p^B)/n^B} = 1,96 \cdot \sqrt{0,77 \cdot 0,23 / 100} \\ = 0,0825$$

$$p_u^B = 0,77 - 0,0825 = 0,6875$$

$$p_o^B = 0,77 + 0,0825 = 0,8525$$

P^B liegt mit 95% Wahrscheinlichkeit im Bereich 68,75% - 85,25%

Beispiel

Programm A wird evaluiert auf 2.000 manuell annotierten Fragen. Genauigkeit: 72,0%

Programm B stehen nur 100 annotierte Fragen zur Verfügung. Genauigkeit: 77,0%

A:

$$z_k \cdot \sqrt{p^A(1-p^A)/n^A} = 1,96 \cdot \sqrt{0,72 \cdot 0,28 / 2000} \\ = 0,0197$$

$$p_u^A = 0,72 - 0,0197 = 0,7003$$

$$p_o^A = 0,72 + 0,0197 = 0,7397$$

P^A liegt mit 95% Wahrscheinlichkeit im Bereich 70,03% - 73,97%

Beispiel

Vertrauensbereich verringern: Wie groß muss Stichprobe sein?

$$p_o^A \approx 0,74, p^B = 0,77$$

→ Vertrauensbereich nach unten $< p^B - p_o^A = 0,03$:

$$z_k \cdot \sqrt{p^B(1-p^B)/n} < 0,03 \rightarrow$$

$$\sqrt{n} > z_k \cdot \sqrt{p^B(1-p^B)}/0,03 = 27,5$$

$$n > 755,94$$