Einführung in die statistische Sprachverarbeitung

Detlef Prescher

May 5, 2007

Hausaufgabe 4

(i) Seien f und $\mathcal{M}$ wie folgt gegeben:

$$f(a) = 2,\ f(b) = 3,\ f(c) = 5 \text{ und } \mathcal{M} = \Big\{ p \in \mathcal{M}(\{a,b,c\}) \ \Big|\ p(a) = 0.5 \Big\}$$

Sei $\tilde{p}$ die relative Häufigkeitsverteilung auf f, sowie $\hat{p}$ eine Maximum-Likelihood-Schätzung von $\mathcal{M}$ auf f. Berechne:

$$D(\tilde{p}||\hat{p})$$

Versuche, Instanzen $p_1,\ p_2 \in \mathcal{M}$ zu finden, sodass

$$D(\tilde{p}||p_1) < D(\tilde{p}||\hat{p}) < D(\tilde{p}||p_2)$$

(ii) Wenn es Dir nicht gelingt, eine der in (i) gesuchten Instanzen p_1 oder p_2 zu finden, so beweise, dass dies prinzipiell unmöglich ist. Benutze dazu einen entsprechenden Satz aus der Vorlesung.

(iii) Berechne für die Verteilungen aus (i) alle Cross-Entropien:

$$H(\tilde{p}; p)$$

Hierbei ist p ein Platzhalter für die Verteilungen $\tilde{p}, \hat{p}, p_1$ und p_2. Zeige unter Benutzung von (i), dass rein rechnerisch für alle diese Verteilungen auch die folgende Gleichung erfüllt ist:

$$D(\tilde{p}||p) = H(\tilde{p}; p) - H(\tilde{p}).$$

(iv) Zeige oder widerlege die Behauptung, dass $D(.||.)$ eine Abstandsfunktion ist.
(Eine **Abstandsfunktion** nimmt nie einen negativen Wert an, ist symmetrisch, und erfüllt ausserdem die sogenannte Dreiecksungleichung, d.h. es gilt stets $D(a||c) \leq D(a||b) + D(b||c)$.)

(v) Die **Perplexität** eines Korpus f bezüglich einer Instanz $p \in \mathcal{M}$ is definiert als $\text{perp}(f; p) = 2^{H(\tilde{p};p)}$. Berechne die Perplexität in Abhängigkeit von der Korpuswahrscheinlichkeit $L(f; p)$ und der Korpusgrösse $|f|$, bzw. die Korpuswahrscheinlichkeit in Abhängigkeit von der Perplexität . Benutze dies, um eine schöne Interpretation für die Perplexität eines Korpus zu finden.

Abgabetermin: Montag, 14. Mai

Besprechung: Mittwoch, 16. Mai