(520|600).666
Information Extraction from Speech and Text

Homework # 4

Due March 28, 2024.

In class, we discussed linear interpolation for smoothing a bigram language model, namely

$$P(w|v) \;=\; \gamma f(w|v) + (1-\gamma) f(w)\,,$$

where $f(\cdot|\cdot)$ and $f(\cdot)$ denoted the appropriate relative frequency estimates, and γ was chosen so as to maximize the probability of some held-out data.

This homework considers *alternative* strategies for smoothing a bigram language model by directly modifying the *counts* observed in the training data. In particular, let $C(v, w)$ denote the count of a bigram $\langle v, w\rangle$ in the <u>*training text*</u>, and let $C^*(v, w)$ be the modified count. For some constant $\theta > 0$, consider the three cases

(i) $C^*(v, w) = C(v, w) + \theta$,

(ii) $C^*(v, w) = C(v, w) + \theta C(w)$, and

(iii) $C^*(v, w) = C(v, w) + \theta C(v) f(w)$.

In each case, the smoothed bigram probability is calculated as

$$P^*(w|v) \;=\; \frac{C^*(v, w)}{\sum_{w' \in \mathcal{V}} C^*(v, w')}.$$

Let $N(v, w)$ denote the count of a bigram $\langle v, w\rangle$ in the <u>*held-out text*</u> $\mathcal{H}$.

1. Derive an expression for the θ that maximizes the log-probability

   $$P(\mathcal{H}) \;=\; \sum_{v \in \mathcal{V}} \sum_{w \in \mathcal{V}} N(v, w) \log P^*(w|v)$$

   of the held-out text in each of the three cases (i), (ii) and (iii) above.

2. Show that if $N(v, w) = C(v, w)$ for all bigrams $\langle v, w\rangle$, then the optimal value is $\theta = 0$ in each case. Why is this an expected result?

3. Show, in each case, that P^* may be written as the linear interpolation of a bigram and a lower order language model, though not necessarily $f(w)$.

   $$P^*(w|v) \;=\; \gamma f_2(w|v) + (1-\gamma) f_1(w)\,,$$

   i.e., identify f_1, f_2 and γ, and discuss the merits/drawbacks of each smoothing strategy.

After finishing the homework, carefully review *all sections* of Chapter 4 again.