**Information Extraction from Speech and Text**

**Homework # 7**

Due April 3, 2015.

Review Chapters 7 and 15 from *Statistical Methods for Speech Recognition* by Frederick Jelinek.

1. Let $p(x, y)$ be given by

\[
\begin{array}{c|cc}
Y & 0 & 1 \\
--- & --- & --- \\
X & \frac{1}{3} & \frac{1}{3} \\
0 & \frac{1}{3} & \frac{1}{3} \\
1 & 0 & \frac{1}{3}
\end{array}
\]

(a) Compute $H(X)$, $H(Y)$ and $H(X,Y)$.
(b) Compute $H(X|Y)$ and $H(Y|X)$.
(c) Compute $I(X;Y)$ from its defining formula.
(d) Compare (c) with $H(X) - H(X|Y)$ and $H(Y) - H(Y|X)$ from (a) and (b).
(e) Draw a Venn diagram relating the all the quantities above.

2. For any probability mass function $p = (p_1, p_2, \ldots, p_i, \ldots, p_j, \ldots, p_k, \ldots, p_m)$ on a discrete alphabet $\mathcal{X}$ with $m$ symbols, show that if

\[
p' = \left( p_1, p_2, \ldots, \frac{p_i + p_k}{2}, \ldots, p_j, \ldots, \frac{p_i + p_k}{2}, \ldots, p_m \right), \quad \text{then} \quad H(p) \leq H(p').
\]

In other words, show that any transfer of probability that makes a distribution more *uniform* increases entropy.

3. **Comparison of Held-Out and Good-Turing Estimates:** We will compare the held-out estimate of Section 15.2 and the Good-Turing estimate of Section 15.4 by using *text A* and *text B* from our projects. In particular, we will use text A to develop our probability estimates, and text B to check their empirical performance in terms of the *average log-likelihood* or perplexity. For this task, the alphabet $\mathcal{X}$ will be *non-overlapping pairs of consecutive letters*, so that $|\mathcal{X}| = 729$, $|A| = 15,000$ and $|B| = 2,500$. We will assume that consecutive symbols are independent, and estimate *unigram* probability distributions on $\mathcal{X}$.
(a) Let the first 12,000 symbols in text A be the development set \( \mathcal{D} \), and the remaining 3,000 the held-out set \( \mathcal{H} \). Use the procedure described in Section 15.2 to

i. begin with a provisional value of \( M = 20 \);

ii. use the counts \( c_d(x) \) in \( \mathcal{D} \) to create equivalence classes \( \Phi : \mathcal{X} \rightarrow \{0, 1, \ldots, M+1\} \), i.e. \( \Phi(x) = i \) if and only if \( c_d(x) = i \);

iii. use the counts \( r_i \) in \( \mathcal{H} \) to estimate the class-probabilities \( \lambda_i, i = 0, 1, \ldots, M+1 \);

iv. use the counts \( c_d(x) \) in \( \mathcal{D} \) to get the class membership counts \( n_i \), the relative frequency estimates \( f_d(x) \) for \( x \) with \( \Phi(x) = M+1 \), and the probability \( P_M \);

v. use the considerations in Section 15.2.3 to choose an appropriate value of \( M \) in Step 3(a)i above, then repeat Steps 3(a)ii through 3(a)iv;

vi. compute the probability estimate

\[
\tilde{P}(x) = \begin{cases} 
\lambda_i \times \frac{1}{n_i} & \text{if } \Phi(x) = i \in \{0, 1, \ldots, M\}, \\
\lambda_{M+1} \times \frac{f_d(x)}{P_M} & \text{if } \Phi(x) = M+1,
\end{cases} \quad x \in \mathcal{X}, \tag{1}
\]

and verify that it sums to unity;

vii. compute the perplexity of text B using the held-out estimate of equation (1);

viii. (alternative) just before Step 3(a)vi, pool the data-sets \( \mathcal{D} \) and \( \mathcal{H} \) back together and

use the counts \( c_d(x) \) in \( \mathcal{D} \cup \mathcal{H} \) to recalculate the equivalence classes \( \Phi : \mathcal{X} \rightarrow \{0, 1, \ldots, M+1\} \), the class membership counts \( n_i \), the estimates \( f_d(x) \) and the probability \( P_M \) in (1).

Compare the performance of the estimate with and without the optional Step 3(a)viii of merging together \( \mathcal{D} \) and \( \mathcal{H} \) for reestimating the class membership and relative frequencies.

(b) Good-Turing estimation does not require dividing the text A into \( \mathcal{D} \) and \( \mathcal{H} \). Therefore,

i. begin with a provisional value of \( M \) as determined in Step 3(a)v above;

ii. compute the counts \( c_d(x) \), equivalence classes \( \Phi : \mathcal{X} \rightarrow \{0, 1, \ldots, M+1\} \), the class membership counts \( n_i \), the estimates \( f_d(x) \) and \( P_M \) from the entire text A;

iii. using \( N \) to denote \( |A| \), compute the probability estimate

\[
\hat{P}(x) = \begin{cases} 
\frac{(i+1)n_{i+1}}{n_i N} \frac{1}{\alpha} & \text{if } \Phi(x) = i \in \{0, 1, \ldots, M\}, \\
\frac{f_d(x)}{P_M} & \text{if } \Phi(x) = M+1,
\end{cases} \quad x \in \mathcal{X}, \tag{2}
\]

where \( \alpha \) is computed, as described in Section 15.4, to ensure that \( \hat{P}(\cdot) \) sums to unity.

iv. compute the perplexity of text B using the Good-Turing estimate of equation (2).

How should one choose \( M \) in Step 3(b)i above? Experiment with a few different values.

Compare the perplexity of the best Good-Turing estimate of Part 3b with that of the best held-out estimate of Part 3a.