CSC 248/448 Assignment #2 Solutions
Due: Oct 2, 2003
CSC448 Students: Read Chapters 1, 2 and 6 of Manning and Schutze

Part I: Written Questions – Please hand in typed answers

1. *Find at least one technical error in the lecture notes.*

2. *Consider a language generated by a sequence of random variables with 2000 different word types. We have a corpus of 1 million and one word tokens (1 million bigrams) and are interested in defining a good bigram probability model.*

a) How many possible types are there in the bigram model of this language?

There are 2000^2=4,000,000 possible bigram types.

b) What is the maximum number of different bigram types that can occur in the corpus. What does this say about the lower bound on the number of bigrams that have not been seen.

If every bigram was unique there would be 1 million distinct bigrams in the corpus. That means at least 3 million bigrams have not been seen.

c) Assuming there are actually 300,000 distinct bigrams in the corpus, what is the probability estimate for the next observed element to be a bigram we have not seen before, according to the MLE, Laplace and ELE estimators?

There are 3,700,000 unseen bigram types. For MLE, the probability of seeing any one of them is 0. For Laplace, each element gets an initial count 1, so we have a count of 3700000 from the unseen bigrams. The total count is 4,000,000 for the initialization count + 1,000,000 seen bigram tokens (= 5,000,000). Thus the probability of seeing an unseen bigram next is 3700000/500000 = .74. For ELE, it's the same except the initial count is .5,, so we have 1,850,000/3,000,000) = .6167.

d) Again assuming the corpus contained 300,000 distinct bigrams. We empirically test a million new words of data and find that unseen bigrams occur 5% of the time. What value of λ in a Lidstone's estimatation would best match the data?

We want the probability of an unseen bigram to be .05. Thus we want

λ * 3.7e6 / (1e6 + λ * 4e6) = .05

which solving yields λ = 1/70 = .0143

3. *Prove that the probability function that maximizes the likelihood of a corpus is the same as the probability function that minimizes the log probability ("cross entropy") of the corpus.*

Assume that P maximizes the probability of a corpus, C, i.e., maximizes $\prod_I P(c_i)$. We know from the basic monotonic property of logs that P also maximizes $\log(\prod_I P(c_i))$. Again, using a basic property of logarithms, we know the log of a product is the sum of the logs, so P maximizes $\sum_I \log(P(c_i))$. Given that N is the size of the corpus, a constant, we also know that P maximizes $\sum_I \log(P(c_i))/N$, and hence minimizes $-\sum_I \log(P(c_i))/N$, which is the definition of corpus cross entropy.

4. *Consider the following corpus from the ABC language*
A B C B A B B C A A B C B C A B C
We want to build a good bigram model of this language using the linear interpolation technique

$P_{LI}(x \mid y) = \lambda P_{MLR}(x \mid y) + (1 - \lambda) P_{MLE}(x)$

We assume we have enough data that we can use the MLE technique for the individual distributions.

a. *Determine the MLE bigram and unigram probability models.*
We have to develop a strategy for handling the last item in the corpus (which is C), since it does not occur as the context for a bigram. Either we could eliminate it from the unigram counts, or we compute the bigrams conditional probabilities by normalizing rather than using the unigram as the denominator. We'll do the latter here.
P(A) = 5/17, P(B) = 7/17, P(C)=5/17
For bigrams, we have
Counts: AA 1 AB 4 AC 0, thus P(A | A) = .2, P(B | A)= .8 and P(C | A) = 0;
Counts: BA 1 BB 1 BC 5, thus P(A | B) = 1/7, P(B | B) = 1/7 and P(C | B) = 5/7
Counts: CA 2 CB 2 CC 0, the P(A | C) = .5, P(B | C) = .5, and P(C | C) = 0

b. *Say we have a development corpus as follows: A C B C A C B C. What is the likelihood of this corpus for the following values of λ: 1, .9 and .5. Which appears to be the best estimate?*
The likelihood of the development corpus will be
$P(C|A)^2 * P(B|C)^2 * P(C|B)^2 * P(A|C)$.
When λ=1, we are only using the bigram probabilities, and so P(C|A)=0 and thus the likelihood of the corpus is 0!
For λ=.9, $P_{LI}(C|A) = .9*P_{MLE}(C|A)+.1*P_{MLE}(C) = .029$. We can calculate the probabilities for the others similarily. Here a chart for the LI estimates for the different lambda's, and the corpus probability in each case.

Lambda	$P_{LI}(C\|A)$	$P_{LI}(B\|C)$	$P_{LI}(C\|B)$	$P_{li}(A\|C)$	Corpus Prob
1	0	.5	.71	.5	0
.9	.025	.49	.67	.48	3.3e-5
.5	.125	.47	.48	.41	3.24e-4

Thus λ=.5 appears to be the best model.
There's an important lesson here – I don't think any two students got the same numbers here – and the numbers aren't what the issue is. A good answer shows how you derived your answer and reveals the thinking process behind your solution. That's what the grades were based on.

c. *Find a better value of λ than the ones provided, or if you can't, given some evidence why this seems to be the case.*
Just doing a simple hill climbing exercise with a spreadsheet I found that λ=.34 appears to yield a good score of 3.58e-4. Numbers lower than that yield lower corpus likelihood score, as do numbers higher.

d. *(448 students) We want to find a good (ideally optimal) value for λ. Is there an analytical solution? If so, derive it. Or if you choose to estimate it with an algorithm, sketch the algorithm.*