Ling 235 Homework #1

Due Wednesday, January 19, 2005

Probability

1. Which of these are true?
   (a) $P(A|B,C) \leq P(A|C)$
   (b) $P(A,B|C) \leq P(A|C)$
   (c) $P(A \cup B) \geq P(A) + P(B) - 1$

2. Suppose (in a certain genre of text) the probability that a word is a noun is 0.4, and the probability that a word is a verb is 0.2. Suppose also that the probability that the word is of Latin origin is 0.3.
   (a) Given just the above information what are the bounds on the minimum and maximum possible probability of a random word being a latinate noun.
   (b) Assuming that part of speech and latinate origin are independent, what is the probability that a random word is a latinate noun.
   (c) Suppose the probability of latinate nouns is actually 0.15. What is the probability that a random word is a noun not of Latin origin?

   Make sure you show your work, explaining the reason for each answer.

SPSS and model building

3. Build two logistic regression models from the Cedergren data in SPSS, one using only POS as an independent variable and the other using only Environment. Which one has higher data likelihood? Which one has higher classification accuracy? Is this surprising? Explain why you see the pattern you do. (Hint: examine crosstabs between Deletion and each independent variable.)

4. Build a logistic regression model from the Cedergren data in SPSS, using POS, Environment, and Class as the factors. Plot predicted versus actual deletion probability for data aggregated by the independent variables, and examine the outliers. Describe what you see. Compare the outliers to other datapoints. Do the outliers have anything in common? Is there anything about these cases that explains why these are outliers?