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) $P(A \cup B) \geq P(A) + P(B) - 1$

   Answers:

   (a) **False.** As an extreme case, suppose you're drawing a card out of a hat. There are four cards in the hat: an ace of hearts, a king of diamonds, a queen of spades, and a jack of clubs. Let A be the event that you draw an ace, B that you draw a heart, and C that you draw a red card [the hearts and diamonds are the red suits!]. Clearly, $P(A|B, C) = 1$ but $P(A|C) = \frac{1}{2}$.

   (b) **True.** By definition,

   $$P(A, B|C) = \frac{P(A \cap B \cap C)}{P(C)}$$

   and

   $$P(A|C) = \frac{P(A \cap C)}{P(C)}$$

   Since $A \cap B \cap C$ is a subset of $A \cap C$, we have $P(A \cap B \cap C) \leq P(A \cap B)$. So $P(A, B|C) \leq P(A, C)$.

   (c) **True.** This follows directly from the property that

   $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

   and the fact that $P(A \cap B) \leq 1$ which is an axiom of probability theory (all probabilities are ≤ 1).

   Here's how to derive this property. The event space $A \cup B$ can be subdivided into three disjoint parts: $A \cap B$, $A \cap \overline{B}$, and $\overline{A} \cap B$ (remember, $\overline{A}$ is the complement of A

– the event that A does not occur). So by the property of **countable additivity**, we have

$$P(A \cup B) = P(A \cap \overline{B}) + P(\overline{A} \cap B) + P(A \cap B)$$

Now a little math trick: we add $P(A \cap B) - P(A \cap B)$ to both sides, giving us

$$P(A \cup B) + \overbrace{P(A \cap B) - P(A \cap B)}^{=0} = P(A \cap \overline{B}) + P(A \cap B) + P(\overline{A} \cap B) + P(A \cap B) - P(A \cap B)$$

Now notice that $A = (A \cap B) \cup (A \cap \overline{B})$. So we can apply the property of countable additivity again to get what we wanted:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

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.

   Answers:

   Let L be the event that a word is of Latin origin, N the event that a word is a noun, V the event that a word is a verb, and O the event that a word is another part of speech (neither a noun or a verb).

   (a) There are six possible combinations of $L, \overline{L}, N, V, O$ that may have probability mass ≥ 0. Here's a valid combination that assigns $P(L, N) = 0$:

$P(L, N) = 0$	$P(\overline{L}, N) = 0.4$
$P(L, V) = 0.2$	$P(\overline{L}, V) = 0$
$P(L, O) = 0.1$	$P(\overline{L}, O) = 0.3$

   This combination obeys all the constraints stipulated on the marginal probabilities of $P(N)$, $P(L)$, and $P(V)$. So the minimum possible value of $P(L, N)$ is 0.

   Here's a valid combination that assigns $P(L, N) = 0.3$:

$P(L, N) = 0.3$	$P(\overline{L}, N) = 0.1$
$P(L, V) = 0.0$	$P(\overline{L}, V) = 0.2$
$P(L, O) = 0$	$P(\overline{L}, O) = 0.4$

We can't set $P(L,N)$ any higher than 0.3 since otherwise the constraint $P(L) = 0.3$ would be violated. So 0.3 is the maximum bound.

Note: A very common answer was to answer this question with reference only to the constraints on $P(N)$ and $P(L)$. This is not adequate: you have to show that the other constraints don't have an indirect influence. For example, if we added the constraint that $P(\overline{L}, V) = 0$, then the upper bound on $P(L,N)$ becomes 0.1, since $P(L,V)$ must be 0.2.

(b) Given independence, we have

$$P(L,N) = P(L) \times P(N) = 0.3 \times 0.4 = \mathbf{0.12}$$

(c) Since $P(N) = P(L,N) + P(\overline{L}, N)$, we have

$$P(\overline{L}, N) = P(N) - P(L,N) = 0.4 - 0.15 = \mathbf{0.25}$$

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.)

   Answer:

   The model with POS has a higher classification accuracy, but the model with Environment has a higher likelihood. This is because for all values of Environment, the frequency of deletion is above 50% in the data, it doesn't classify the data any better than a model with only the "input" ("intercept").

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?

   Answer:

   The outliers all have small N (i.e., low counts) compared to the dataset as a whole. There are two ways to think about this (that amount to roughly the same thing). From the perspective of choosing a model to fit the data, the model doesn't care much about those outliers, because the small N doesn't contribute much to the overall likelihood of the model. So it's not that surprising that the best model actually does a poor job with some datapoints with low N.

   From the perspective of seeing whether the fitted model could reasonably have generated the data, we expect to see the worst outlier cases among the low-N datapoints because, intuitively, there is much more variability in the empirical outcome of a small sample than that of a large sample. (The technical term for this is that the sample ***variance*** is higher for low N – we'll cover this technical term later in the course.)