ANLP Tutorial Exercises for Week 1 (v1.5)

Adam Lopez, Sharon Goldwater
(School of Informatics, University of Edinburgh)

Goals

This exercise sheet has two main goals:

1) To check that you have read the material on probability theory (or are already familiar with it) and are learning to solve problems on your own. If this material is entirely new to you, you may not have completely absorbed all of it by now (this will take some time), but you should definitely try to work through these exercises and post questions if you have trouble. We will use probability throughout the course.

2) To give you practice thinking about ambiguity and morphology.

Deadline, solutions, and getting help

You should complete the exercises in this sheet *before the start of Week 2*. There is no required submission, but the deadline will help keep you on track with the pace of the course.

If you have questions while working through the problems, we encourage you to post them to the class Piazza forum.

Many of the exercises have numeric answers, and for these you can check your answers automatically using Gradescope (see the link on Learn).

For all the exercises, we will release a solution sheet with some additional hints and explanations on Monday the week after releasing the questions. Please look over it and if you still have questions, post to Piazza or bring your questions to the drop-in help hour.

Exercise 1

Suppose we have a group of first- and second-year students, each of whom comes from either the UK, Germany, or China. We choose a student uniformly at random from this group. Let F = "the student is in first year", U = "the student is from the UK", C = "the student is from China", and G ="the student is from Germany". For this group of students, $P(U) = 0.6$, $P(G) = 0.3$, $P(C) = 0.1$, $P(F \mid U) = 0.7$, $P(F \mid G) = 0.5$, $P(F \mid C) = 0.6$.

a) What is the probability that we've chosen a first-year student?

b) What is the probability that the student is from the UK if it is a first-year student?

Solutions

a) Use the rule of total probability:

$$\begin{aligned} P(F) &= P(U)\ P(F \mid U) + P(G)\ P(F \mid G) + P(C)\ P(F \mid C) \\ &= (.6)(.7) + (.3)(.5) + (.1)(.6) \\ &= .63 \end{aligned}$$

b) Now use Bayes' Rule:

$$\begin{aligned} P(U \mid F) &= \frac{P(F \mid U)\, P(U)}{P(F)} \\ &= \frac{(0.7)(0.6)}{.63} \\ &= 2/3\ (\approx 0.67) \end{aligned}$$

Exercise 2

(Exercise 4.6 from the Basic Probability Theory tutorial)

Suppose I want to choose a two-word name for a rock band via a random process. The first word will be chosen with uniform probability from the set {`yellow`, `purple`, `sordid`, `twisted`}. The second word will be either `quake`, with probability 1/3, or `revolution`, with probability 2/3.

a) Determine the sample space for this experiment. How many outcomes are in it?

b) If I choose the first and second words independently, what is the probability that my rock band will be called `yellow revolution`?

c) Instead, I decide to condition the second word on the first. If the first word is a color, then I'll pick `quake` as the second word with probability 1/6. Assuming I still want the overall probability of a name with `quake` to be 1/3, what is the probability of `quake` if the first word is *not* a color?

(Hint: use a joint probability table to help you answer this question. You may not need to fill in the entire table.)

Solutions

a) The sample space is all ordered two-word pairs using the words provided, and contains 8 possible pairs.

b) $P(\texttt{yellow revolution}) = P(\texttt{yellow})P(\texttt{revolution}) = (1/4)(2/3) = 1/6$, or around 0.17.

c) Here's one way to solve the problem. Let C be the event that the first word is a color. Since there are two color words out of four equally likely words, $P(C) = 1/2$. Also, let Q be "second word is `quake`" and R be "second word is `revolution`". The initial JPT contains all the marginals: we were given $P(R)$ and $P(Q)$, we just computed $P(C)$, and $P(\neg C) = 1 - P(C)$. Also, we add $P(C, Q) = P(C)P(Q \mid C) = (1/2)(1/6) = 1/12$:

	Q	R	
C	1/12		1/2
$\neg C$			1/2
	1/3	2/3	

Now we can fill in $P(Q, \neg C) = P(Q) - P(Q, C) = 1/4$. And finally, compute $P(Q \mid \neg C) = \frac{P(Q,\neg C)}{P(\neg C)} = \frac{1/4}{1/2} = 1/2$.

Exercise 3

Suppose I decide to probabilistically generate a "word", which can contain only characters in the set $\{\texttt{a}, \texttt{b}, \texttt{c}, \texttt{d}, \texttt{e}\}$. To generate the word, I start by generating the character `a`. Then, with probability q, I generate a single `b` and stop, otherwise I choose one of the other four characters (`a`, `c`, `d`, `e`) uniformly at random and keep going. I continue this process, always either generating a single `b` with probability q and stopping, or choosing one of the other four characters uniformly at random and continuing.

a) If L is a random variable representing the length of a word, give an equation for $P(L = n)$.

b) If $q = 0.3$, what is the probability that my generated word has 4 characters?

c) If $q = 0.3$, what is the probability of generating the word `aaab`?

d) If $q = 0.3$, what is the probability of generating the word `aaab` given that I generate a 4-character word?

Solutions

a) The equation is $P(L = n) = (1 - q)^{n-2}(q)$. (Note the first character is "free", thus the $n - 2$; also the fact that there are multiple different continuation characters is irrelevant, only the stopping prob matters.)

b) The probability of a 4-character word is $(0.7)^2(0.3) \approx 0.15$.

c) There are many different possible 4-character words, so $P(\texttt{aaab})$ is much lower. After the first character, $P(a) = 0.7/4$, so $P(\texttt{aaab}) = (0.7/4)^2(0.3) \approx .0092$.

d) Note that $P(\texttt{aaab}, L = 4) = P(\texttt{aaab})$, so

$$\begin{aligned} P(\texttt{aaab}\,|, L = 4) &= \frac{P(\texttt{aaab})}{P(L = 4)} \\ &= \frac{(0.7/4)^2(0.3)}{(0.7)^2(0.3)} \\ &= 1/16 \ (= 0.0625) \end{aligned}$$

Exercise 4

Newspaper headlines are often ambiguous, sometimes leading to unintended humour. In each of the example headlines below, identify one type of ambiguity that we discussed in lecture. If it is a lexical ambiguity, provide at least two definitions of the ambiguous word. If it is a syntactic ambiguity, paraphrase the sentence two different ways to illustrate the different possible meanings of the sentence.

a) Scientists count whales from space.

b) March planned for January.

c) Medics help dog bite victim.

d) Paramilitary head seeks arms.

Solutions

You may find that there are multiple types of ambiguity in each sentence, so the ones below are not necessarily an exhaustive list, merely illustrative examples.

a) Syntactic ambiguity with possible paraphrases:

- Scientists count whales by viewing them from space (e.g. with satellite cameras).
- Scientists count whales that are in space.

b) "March" is lexically ambiguous, meaning:

- a procession organized as a protest.
- the third month of the year.

c) Syntactic ambiguity with possible paraphrases:

- Medics help someone who was bitten by a dog.
- Medics help a dog to bite someone.

d) There are (at least!) two ambiguous words here:

- "head" can mean:

  - chief of an organisation
  - the upper part of the human body, containing the brain, mouth, and sensory organs
- "arms" can mean:
  - weapons
  - each of the two upper limbs of the human body from the shoulder to the hand.

Exercise 5

a) Give the lemma for any word(s) in following the sentence where the lemma is different from the word itself.

The students are arriving in Edinburgh after a long trip.

b) Consider the following set of Spanish words and their English translations. Using the translations as a hint, decompose each word into a pair of morphemes, and for each morpheme that you find, give its translation. Since we don't expect you to know Spanish, it is ok if the morphemes that you find don't exactly match their dictionary form.

hablo	'I speak'
trabajas	'you work'
hablas	'you speak'
escucho	'I listen'
trabajo	'I work'
escuchas	'you listen'

Solutions

a) Lemmatised words in **bold**:
The ***student be arrive*** *in Edinburgh after a long trip.*

b) Make a table that groups words into rows and columns according to their translations. Once you've placed each word into the table, likely morpheme boundaries should be apparent:

person →	1st	2nd
English lemma ↓	'I'	'you'
'listen'	*escuch+o*	*escuch+as*
'speak'	*habl+o*	*habl+as*
'work'	*trabaj+o*	*trabaj+as*

You can take the first part of each Spanish word to be the root, and the second to be the affix indicating person. In this case, Spanish words undergo some changes during affixation; since you haven't observed the lemma form of the words, I wouldn't expect you to know this, but I've included the true Spanish lemmas in the table below.

found morpheme	meaning	Spanish lemma
escuch	'listen'	*escuchar*
habl	'speak'	*hablar*
trabaj	'work'	*trabajar*
-o	1st person	
-as	2nd person	