Homework 4

Due: 29 March 2022

v20220322-2300

- Include work in symbolic form (e.g. $p(X = \text{foo}|Y = 3)$). I should be able to tell where every number came from. You can abbreviate (e.g. $p(\text{foo}|3)$) as long as it's clear what is meant.
- Sanity-check your answers. If your answer seems crazy but you can't find the error, at least make it clear you know there's a problem.

Problem 4.1

A standard deck of playing cards has 52 cards in four suits (two red, two black), each suit with cards numbered 2–10 and "face cards" labelled jack, queen, king, and ace. A "pinochle deck" has 48 cards: each of the four suits has only 9 and 10 plus the face cards, and each card appears twice in the deck.[1]

Assuming in each case that the relevant deck is well-shuffled, identify the following probabilities. Show your work by making it clear where each number in your probability comes from.

a. p(R=ace | D=standard): In a standard deck, the probability of drawing any card whose rank is 'ace'

b. p(R=ace | D=pinochle): In a pinochle deck, the probability of drawing any card whose rank is 'ace'

c. p(F=true, C=red | D=pinochle): the probability of drawing a card whose rank is one of the face cards and whose suit is one of the red suits, from a pinochle deck

d. p(R=A, S=H | F=true): the probability of drawing the ace of hearts given that the card drawn is a face card

e. If there are 3 standard decks and 1 pinochle deck in a pile, and you pick one at random before drawing a card from that deck, what is p(F=true), i.e. the overall probability of drawing a face card?

[1] I swear I am not making this up. It's pronounced "PEE-nuckle".

Problem 4.2

I have many decks of playing cards, most of which are standard, but some are pinochle decks—for the purposes of this problem, let's say I have two pinochle decks and ten standard decks. Without counting the cards, it's hard to tell at a glance whether you've accidentally grabbed a pinochle deck.

a. If I grab a deck completely at random and draw a card from it, what is the probability that the card is a 5?

b. If I grab a deck completely at random and draw a card from it, what is the probability that the card is a jack?

c. If I grab a deck completely at random and draw a card, and the card is a jack, what is the likelihood that I've grabbed a pinochle deck?

d. If I grab a deck completely at random and draw two cards from it, some pairs give me certainty: if either card is a 4, for instance, or if both cards are the jack of diamonds. But if one is a jack and the other is a king, does that give me any knowledge about the deck? Why or why not? (Note: I'm not looking for exact numbers on this part, because they're subtle and a bit gross. Focus on the analysis.)

Again, don't forget to show your work.

For the next two questions, consider the sentence

> The council has approved wind farms.

in light of the probability tables on the back page.

Problem 4.3

Using the tables, give an appropriate probability for the sentence according to each of the following models:

a. unigram model (word probabilities out of context)

b. bigram model (previous word as context/1st-order Markov model)

c. POS-driven model (parts of speech as hidden states in HMM)

You will need to make certain assumptions to complete this; make sure you state them (and, of course, show your work).

Problem 4.4

For each of the three models (unigram, bigram, POS-driven), give another six-word utterance that is A) not good English (the worse, the better) but which B) has a probability substantially higher than the given sentence, according to the given probabilities, and justify your answer (i.e. explain why the probability would be a lot higher).

Collaboration policy: group work! If you work with other people on this homework, hand in one copy and put all your names on top. There will be a revision cycle for this.

These are actual statistics (rounded and with just a few modifications), trained from the primary training section of the WSJ Penn Treebank,

These are the overall frequencies of each word:

the	.05
council	.000075
has	.0035
approved	.00015
wind	.00002
farms	.000015

These are a selection of relevant bigram probabilities p(word|prevword):

the→approved	.00005	approved→has	.00001
the→council	.0004	approved→the	.15
the→wind	.00006	approved→wind	.00000
council→has	.00000	wind→approved	.00000
has→approved	.0006	wind→council	.00000
has→the	.015	wind→farms	.00000

(Note that bigram pairs that did not occur in the training are reported as .00000 in the above table.)

These are a selection of part-of-speech probabilities for the given words p(word|POS):

D→the	.6	V→has	.03
N→council	.0002	V→approved	.001
N→wind	.000025	V→wind	.00006
N→farms	.00004	V→farms	.000008
N→the	.00015		

These are a selection of POS probabilities given previous POS:

D→D	.0015	N→V	.15
D→N	.65	V→D	.2
D→V	.03	V→N	.15
N→D	.0075	V→V	.15
N→N	.3		