# Homework 3

*Due: 24 March 2016*

### Problem 3.1

Consider the following situation: I wish to evaluate the probability that a storm is coming within the next hour, so I consult various weather measurements such as the current temperature, the current humidity, the current air pressure, the wind speed, the wind direction, the wind chill, the heat index, and the temperature differential and pressure differential (how much they've changed in the last hour). I have piles of data from the past, including when storms occurred. Build (by hand, on paper) some classifiers to decide whether it's more likely that a storm will or will not occur in the next hour.

a. Give the complete formula(s) (*without* making independence assumptions) that corresponds to the classifier you are building.

b. Then, give the actual formula(s) that would be used as a naïve Bayes classifier; comment on which of the independence assumptions seem more or less reasonable (and why).

c. Finally, draw out a Bayes net that accounts for the relatedness between some of the variables and alternate explanations for them.

Obviously (?), I don't expect you to compute actual probabilities and scores here since I have given you no numbers—only the symbolic formulae are required. Don't make this problem harder than it is.

*Continued on next page...*

For the next two questions, consider the sentence

> The council has approved wind farms.

## Problem 3.2

Using the probability tables on the next page, 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 3.3

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 comparable to (or higher than) the given sentence.

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→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 | | |