CMSC389 Artificial Intelligence Blaheta

Homework 5

Due: 14 April 2020

Consider the following data set, representing 16 data points with attributes A, B, and C and "outcome" attribute R:

A	B	C	R
T	T	F	−
T	T	F	−
T	F	F	−
T	F	F	+
T	F	F	+
T	F	T	+
T	T	T	+
T	T	T	+
F	T	T	+
F	T	T	−
F	T	T	−
F	T	T	−
F	F	T	+
F	F	T	+
F	F	T	+
F	F	T	+

Problem 5.1

Give the entropy of the overall data set with respect to the random variable R; compute the information gain represented by attributes A, B, and C; and discuss what your computed gain tells us about the decision tree that would be built to predict R.

Problem 5.2

On this data set, how many decision nodes would the final decision tree have? Assuming that the data is distributed exactly like this training set, what is the best performance you could expect from the tree? How does this compare to the performance you'd expect with no tree, or the one-node only-the-root tree that you built in the previous problem? Discuss what else would be needed to further improve on that performance.

For the next two questions, consider the sentence

> The council has approved wind farms.

in light of the probability tables on the next page.

Problem 5.3

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 5.4

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.

Handin policy: As for the previous homework: pdf, docx, odt; upload one per group; use the handin script, with assignment name `hwk5`, by 11:59pm on the duedate. Everybody's name should be in/on the document. When I hand back I will send it to everyone in the group.

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 $p(\text{word}|\text{prevword})$:

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 $p(\text{word}|\text{POS})$:

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		