Computer Science
St. George Campus

19 September 2017
University of Toronto

Computing for Medicine, Phase 3, Project #1

Identifying Alzheimer's disease from picture descriptions

This assignment will give you experience with corpora (i.e., a collection of descriptions of the 'Cookie Theft' picture), Python programming, part-of-speech (PoS) tags, and very simple machine learning.

Your task is to tag sentences with a PoS tagger, gather some features from each utterance, learn models, and use these models to classify utterances as coming from people with or without dementia. Speech-based assessment is an important, still emerging topic in computational linguistics in which we try to identify clinically-relevant features and use them to help make diagnoses.

You should check the course Piazza for announcements and to post your questions:
`piazza.com/utoronto.ca/fall2016/c4m`

Introduction

Assessing for Alzheimer's disease (AD) can be an expensive and laborious process that is challenging to sustain given Canada's rapidly aging population. Even with the savings associated with early assessment, the *amortized* cost for each diagnosis is at least $6000 [4], which does not account for indirect costs such as lost time spent by patients (and families) in travel, wait times, and hours spent in assessment, which is a process so laborious (and stressful) that it is often repeated only every few years. This is unfortunate, since the high variability of symptoms in AD means that it cannot be ascertained accurately from a single assessment [10]. Repeatable, remote, and cost-effective assessment is important.

Although memory impairment is the main symptom of AD, language impairment is an important marker. There is increasing evidence that linguistic aspects of speech relate to fronto-temporal lobar degeneration and cognitive decline [8, 9]. Faber-Langendoen *et al.* [3] found that 36% of mild AD patients and 100% of severe AD patients had aphasia, according to standard aphasia testing protocols. Ahmed *et al.* [1] found that two-thirds of their participants showed subtle, but significant, changes in connected speech production up to a year before their diagnosis of probable AD. Weiner *et al.* [11], in a study of 486 AD patients, reported a significant correlation between dementia severity and a number of different linguistic measures, including confrontation naming, articulation, word-finding ability, and semantic fluency. Subsequently, Jarrold *et al.* [5] used acoustic features, PoS features, and psychologically-motivated word lists to distinguish between semi-structured interview responses from 9 AD participants and 9 controls with an accuracy of 88%. They also confirmed their hypothesis that AD patients would use more pronouns, verbs, and adjectives and fewer nouns than controls.

DementiaBank

The DementiaBank corpus is part of the larger TalkBank project [7]. These data were collected between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. Information about the study cohort is available from Becker *et al.* [2]. Participants were referred directly from the Benedum Geriatric Center at the University of Pittsburgh Medical Center, and others were recruited through the Allegheny County Medical Society, local neurologists and psychiatrists, and public service messages on local media. To be eligible for inclusion in the study, all participants were required to be

above 44 years of age, have at least 7 years of education, have no history of nervous system disorders or be taking neuroleptic medication, have an initial Mini-Mental State Exam (MMSE) score of 10 or greater, and be able to give informed consent. Additionally, participants with dementia were required to have a relative or caregiver to act as an informant. All participants received an extensive neuropsychological and physical assessment [2]. Participants were assigned to the "patient" group primarily based on a history of cognitive and functional decline, and the results of a mental status examination. Each speech sample was recorded then manually transcribed following the TalkBank CHAT protocol [6]. Narratives were segmented into utterances and annotated with filled pauses, paraphasias, and unintelligible words. From the CHAT transcripts, we keep only the word-level transcription and the utterance segmentation. Before tagging and parsing the transcripts, we automatically removed short false starts and filled pauses such as *uh, um, er*, and *ah*. In the data we provide you, only the participant ('PAR') utterances are included; the interviewer data is discarded as irrelevant.

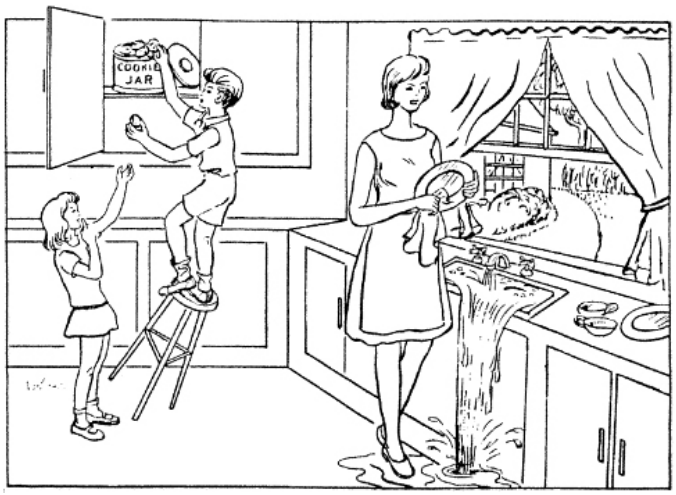

Figure 1: The "Cookie Theft" picture.

Your tasks

1. Gathering feature information

In this section, you will complete the function `extract_features`. Each utterance is represented by a special **identifying code (fileID)** such as `096-0-112v-1_8_PAR`, where the first three digits identifies the person, the last digit represents the position of the utterance within a sequence in the conversation, and you can safely ignore other codes.

We have already separated the words (`(fileID).stem`), performed PoS tagging (`(fileID).pos`), and performed grammatical parsing (`(fileID).pars`). Note that, because of ambiguities in these tasks, errors *do* exist in these data, as is typical in practice. All these data are available in aptly-named `Controls` and `Dementia` folders in `Data`.

The `extract_features` function takes a list of (automatically-generated) files, and *one* of these folders as arguments. The function returns a $D \times 9$ matrix (i.e., '`features`'), where D is the number of fileIDs in the folder. I.e., each fileID is represented by a `vector` of the 9 numbers you compute.

For each fileID (which is represented by three files!), you will extract 9 features. Specifically:

1. **Count number of words in utterance**: count the number of lines in the respective `.stem` file.
2. **Count average number of characters per word in utterance**: divide the sum of all characters in the respective `.stem` file by the number of words in that file.
3. **Compute Honoré's statistic on utterance**: This is $100 \log_{10} N / \left(1 - \left(\frac{V_1}{V}\right)\right)$, where N is the Feature (1) above, V_1 is the number of these that occur exactly once, and V is the total number of *unique* words in the respective `.stem` file. If $V_1 = V$, return 0.
4. **Compute the parse tree depth**: Count the longest sequence of `')'` characters in the respective `.pars` file.
5. **Count the number of 'CC' instances in the respective `.pos` file.**: The number of coordinating conjunctions is a proxy for syntactic complexity.
6. **Count the number of 'VBG' instances in the respective `.pos` file.**: Atypical use of gerund verbs is associated with dementia.
7. **Count the number of 'VBZ' and 'VBP' instances in the respective `.pos` file.**: Atypical use of verbs which can be used as auxiliaries is associated with dementia.
8. **Count the average Age of Acquisition (AoA) of words**. We extracted the spreadsheet from Norms.csv into the dictionary variable `norms`. For each word in the respective `.stem` file, look it up in `norms`; if it exists, add the first field of the result to a running total; if it doesn't exist, add 0. Divide by the total number of words *with* an entry in `norms`. If none of the words occur in `norms`, return 0.
9. **Compute (# NN + NNS + NNP + NNPS) / (# PRP + PRP$)** in the respective `.pos` file. If `(# PRP + PRP$)` is 0, return 0.

Note that the meanings of PoS tags are given in Tables 1a and 1b of the appendix.

2. Classifying feature vectors

You will now complete the function `classify`, which takes the feature matrices for the participants with and without dementia, respectively. From these, we automatically define 'classes' for each utterance (i.e., `ALLclass`, where 1: Alzheimer's, and 0: Controls). The function is already set up to do 5-fold cross validation, given random permutations of their concatenation. Within that for-loop, create an SVM model with:

```
my_model = svm.SVC(kernel="linear")
```

Then, train a model with

```
c_train = np.ravel( c_train ) # reformat the numpy array
my_model.fit( A , c_train )
```

where `A` is the slice of `all_data` using only the indices in `i_train`. Finally, use that trained model to make predictions on test data using

```
my_predictions = my_model.predict( B )
```

where `B` is the slice of `all_data` using only the indices in `i_test`. .

Given (TP): the number of Alzheimer's utterances classified as Alzheimer's; (TN): the number of Control utterances classified as Control; (FP): the number of Control utterances classified as Alzheimer's; and (FN): the number of Alzheimer's utterances classified as Control, compute:

accuracy : $(TP + TN)/(TP + TN + FP + FN)$, stored in `accuracies[fold]`

sensitivity : $TP/(TP + FN)$, stored in `sensitivities[fold]`

specificity : $TN/(TN + FP)$, stored in `specificities[fold]`

Report the averages and variances of each of these, over the 5 folds.

Note: we're taking a severe shortcut, for simplicity. By randomizing all *utterances*, there is a very good probability that *some* individual person will be represented in both the training and testing set, in different utterances, in any fold. In practice, this needs to be avoided, to mimic the scenario that people are assessed given systems trained only with *other* people. So, in 'real life', the cross-fold validation is more complex.

References

[1] S. Ahmed, A.-M.F. Haigh, C.A. de Jager, and P. Garrard. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, 136(12):3727–3737, 2013.

[2] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle. The natural history of Alzheimer's disease. *Archives of Neurology*, 51:585–594, 1994.

[3] K. Faber-Langendoen, J.C. Morris, J.W. Knesevich, E. LaBarge, J.P. Miller, and L. Berg. Aphasia in senile dementia of the Alzheimer type. *Annals of Neurology*, 23(4):365–370, 1988.

[4] D. Getsios, S. Blume, K.J. Ishak, G. Maclaine, and L. Hernandez. An economic evaluation of early assessment for Alzheimer's disease in the United Kingdom. *Alzheimer's & Dementia*, 8:22–30, 2012.

[5] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M.L. Gorno-Tempini, and J. Ogar. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36, 2014.

[6] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

[7] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland. AphasiaBank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

[8] S.V.S Pakhomov, G.E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D.S. Knopman. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and behavioral neurology : official journal of the Society for Behavioral and Cognitive Neurology*, 23(3):165–177, 2010.

[9] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090, 2011.

[10] K. Rockwood, S. Fay, L. Hamilton, E. Ross, and P. Moorhouse. Good days and bad days in dementia: A qualitative analysis of variability in symptom expression. *International Psychogeriatrics*, 26(8):1239–1246, 2009.

[11] M.F. Weiner, K.E. Neubecker, M.E. Bret, and L.S. Hynan. Language in Alzheimer's disease. *The Journal of Clinical Psychiatry*, 69(8):1223–1227, 2008.

Appendix 1: Tables

Table 1a: The Penn part-of-speech tagset—words

Tag	Name	Example
CC	Coordinating conjunction	*and*
CD	Cardinal number	*three*
DT	Determiner	*the*
EX	Existential *there*	*there [is]*
FW	Foreign word	*d'oeuvre*
IN	Preposition or subordinating conjunction	*in, of, like*
JJ	Adjective	*green, good*
JJR	Adjective, comparative	*greener, better*
JJS	Adjective, superlative	*greenest, best*
LS	List item marker	*(1)*
MD	Modal	*could, will*
NN	Noun, singular or mass	*table*
NNS	Noun, plural	*tables*
NNP	Proper noun, singular	*John*
NNPS	Proper noun, plural	*Vikings*
PDT	Predeterminer	*both [the boys]*
POS	Possessive ending	*'s, '*
PRP	Personal pronoun	*I, he, it*
PRP$	Possessive pronoun	*my, his, its*
RB	Adverb	*however, usually, naturally, here, good*
RBR	Adverb, comparative	*better*
RBS	Adverb, superlative	*best*
RP	Particle	*[give] up*
SYM	Symbol (mathematical or scientific)	+
TO	*to*	*to [go] to [him]*
UH	Interjection	*uh-huh*
VB	Verb, base form	*take*
VBD	Verb, past tense	*took*
VBG	Verb, gerund or present participle	*taking*
VBN	Verb, past participle	*taken*
VBP	Verb, non-3rd-person singular present	*take*
VBZ	Verb, 3rd-person singular present	*takes*
WDT	*wh*-determiner	*which*
WP	*wh*-pronoun	*who, what*
WP$	Possessive *wh*-pronoun	*whose*
WRB	*wh*-adverb	*where, when*

Table 1b: The Penn part-of-speech tagset—punctuation

Tag	Name	Example
#	Pound sign	£
$	Dollar sign	$
.	Sentence-final punctuation	*!, ?, .*
,	Comma	
:	Colon, semi-colon, ellipsis	
(	Left bracket character	
)	Right bracket character	
"	Straight double quote	
‘	Left open single quote	
“	Left open double quote	
’	Right close single quote	
”	Right close double quote	

Table 2: Conversion from raw text to tagged text

Raw text:

```
Meet me today at the FEC in DC at 4. Wear a carnation so I know
it's you. <a href="Http://bit.ly/PACattack" target="_blank"
class="tweet-url web" rel="nofollow">Http://bit.ly/PACattack</a>.
```

Tagged text:

```
Meet/VB me/PRP today/NN at/IN the/DT FEC/NN in/IN DC/NN at/IN 4/NN ./.
Wear/VB a/DT carnation/NN so/RB I/PRP know/VB it/PRP 's/POS you/PRP ./.
```