Homework 2 - Sparse generative models

Evan Gravelle October 27, 2016

1. Vocabulary Selection Methodology

My method for vocabulary selection from 20Newsgroups data uses entropy as a measure of how useful a word is for classification. The probability distribution of word frequency is found for each class. Entropy is a measure of surprise, so selecting words maximizing entropy should select words which do well in classification. Average entropy is calculated for each word, and the m words with the greatest average entropy are selected to be the new vocabulary.

2. Pseudocode

Algorithm 1: Vocabulary Selection Algorithm

Input : Full vocabulary V, articles A with classes C, number of prototypes m
Output: Vocabulary subset $\tilde{V}$, Bayes model $\tilde{P}$

Calculate marginal probabilities of each class
Calculate multinomial naive Bayes model P using word frequencies with smoothing in each document class
Calculate average entropy of each word using P
Sort words, store m words with greatest entropy
Calculate reduced multinomial naive Bayes model $\tilde{P}$ using word frequencies with smoothing in each document class
Return $\tilde{V}, \tilde{P}$

3. Experimental Results

The accuracy achieved using naive Bayes with the full vocabulary was 78.1% Unfortunately I did not have time to obtain accurate results for subsets of vocabulary.

4. Inspection of Models

Given a smaller vocabulary set $\tilde{V}$, representative words of each class were selected by iterating over the words and assigning each word to the class with highest probability that didn't already have 10 words. This is an effective method because we have already removed common words with low entropy. The representatives of each class are listed below. These words make some sense as representatives, because they would reasonable appear in articles of their class but not often otherwise.

alt.atheism: crapulous ghoddam sadie scotts whitepine wpine searchit moveable xvtdl jipping

comp.graphics: turok mlt eno dnodisplay putchar winprop iop interactivly gabi irakliotis

comp.os.ms-windows.misc: arshad einval boutilie xware xconfigure usally bagels hoswell iconname commsolutions

comp.sys.ibm.pc.hardware: tcport coskrey samos yuch eide xtpopupspringloaded kinzy xsomething dopenwin commoninteract

comp.sys.mac.hardware: coors ainman needful inman reseting simplemenu xawpositionsimplemenu wcchildren wccreate smebsb

comp.windows.x: bibcard xvttool gazooch sspkg docpart jax encyclo sinan arvai sidtool

misc.forsale: etrbom sysprog ims tonyf tical elbel intelligenz sowieso wat doofe

rec.autos: allbery xibm downie capxterm visionware carley salvador sandiway qplib lsspkg
rec.motorcycles: xmwindow myedit settext statictext staticimage jarnot vita convergent deiconify vadi
rec.sport.baseball: syslog hibition greening sprouting aub lime jello conditionals makdepend writeyourownparser
rec.sport.hockey: xscreensaver xkeyreleasedevent infores pacebes cozuelos tid hardwarecolor benj rexecd jorgen
sci.crypt: corbet kann gildas gjs archiboard giovanni pisa wlij bailgate caird
sci.electronics: savescreen bool eventhandle nxsetscreenbrightness faultline vxt papp spapp minya xcoral
sci.med: strickland calentool proteus bimail attractors openfonts typescaler currentfontmem makeafb bembo
sci.space: xblackjack basile soleil serma starynkevitch mahoney cooden bexhill ptm pmahoney
soc.religion.christian: acaird teemtalk xinet xnth multiview dgis gbytes multitasks zymos poach
talk.politics.guns: convertfont btn keyboardcommand thinnish happyboy sunf sunxk answerbook realxfishdb traversalon
talk.politics.mideast: excelan wallongong bns ungermann bwnfs xtreme hotkeys rexec sdarling pspmf
talk.politics.misc: inputfocuscolor owcolors reservecolors sinkwitz ovi sebeke lfi nsslsun nssl hexsweeper
talk.religion.misc: lxpm lxvps lxview lolgx lcps lpixrect ldd overstriking thanh xnlock

5. Critical Evaluation

I expect this method of vocabulary selection to outperform random vocabulary selection. Still, there are potentially multiple avenues for improvement in vocabulary selection. The prototype selection uses entropy as a metric for classification usefulness, where other values like variance (or linear combinations of multiple) might perform better. This method is naive in assuming that word frequencies are conditionally independent, a much more sophisticated model would try to model these probabilities. In selecting representative words, perhaps a minimum frequency threshold could be used to eliminate rare words.