
```
% Austin Welch
% EC503 HW6.1f
% SVM Classifier for Text Documents
% dataset: data_20news.zip
% using svmtrain, svmclassify
```

Setup

```
% clear variables/console and suppress warnings
clear; clc;
id = 'stats:obsolete:ReplaceThisWithMethodOfObjectReturnedBy';
id2 = 'stats:obsolete:ReplaceThisWith';
warning('off',id);
warning('off',id2);

% load data
disp('Loading data...');
traindata = importdata('train.data');
trainlabel = importdata('train.label');
testdata = importdata('test.data');
testlabel = importdata('test.label');
vocab = importdata('vocabulary.txt'); % all words in docs,
    line#=wordID
stoplist = importdata('stoplist.txt'); % list of commonly used stop
    words
classes = importdata('newsgrouplabels.txt'); % names of the 20 classes

% determine wordIDs in vocabulary that are not in train/test data
IDsNotInTrain = setdiff(1:length(vocab),unique(traindata(:,2)));
IDsNotInTest = setdiff(1:length(vocab),unique(testdata(:,2)));

% determine stop words' wordIDs
[~, stopIDs, ~] = intersect(vocab, stoplist);

% change stop word counts to zero
traindata(ismember(traindata(:,2),stopIDs),3) = 0;
testdata(ismember(testdata(:,2),stopIDs),3) = 0;

% add missing words to train/test data, but with zero counts
appendRows = zeros(length(IDsNotInTrain),3);
appendRows(:,1) = 1; appendRows(:,2) = IDsNotInTrain; appendRows(:,3)
    = 0;
traindata = [appendRows; traindata];
appendRows = zeros(length(IDsNotInTest),3);
appendRows(:,1) = 1; appendRows(:,2) = IDsNotInTest; appendRows(:,3) =
    0;
testdata = [appendRows; testdata];
clear appendRows;

% rearrange train/test data to dimensions (doc#, vocab#) with count
    values
Mtrain = sparse(accumarray(traindata(:,1:2), traindata(:,3)));
```

```

Mtest = sparse(accumarray(testdata(:,1:2), testdata(:,3))),

% calculate frequencies by dividing each count by the word totals
Mtrain = Mtrain ./ sum(Mtrain,2);
Mtest = Mtest ./ sum(Mtest,2);

% when removing stop words, couple docs end up with total word counts
of
% zero, which causes division by 0 when calculating frequencies and
results
% in nans. need to find these nans and replace with zeros.
Mtrain(sum(Mtrain,2)==0,:) = 0;
Mtest(sum(Mtest,2)==0,:) = 0;

Loading data...

```

Part (f) : One-versus-one OVO multi-class classification rbf kernel

```

fprintf('\nStarting part (f)...\n\n');

% all combinations and count
allPairs = combnk(1:20,2);
mChoose2 = nchoosek(20,2);

% train m(m-1)/2=190 binary SVMs for all class pairs
tic
allSVMs = cell(1,mChoose2);
fprintf('Training all binary SVM pairs with rbf kernel...\n\n');
h = waitbar(0, 'Training all binary SVM pairs...', 'Name', 'Part (e)');
for p=1:mChoose2
    waitbar(p/mChoose2);
    % select pair
    pair = allPairs(p,:);
    trainDataPair = sparse(Mtrain((trainlabel==pair(1) | ...
        trainlabel==pair(2))),:));
    trainLabelPair = trainlabel(trainlabel==pair(1) |
        trainlabel==pair(2));
    % train pair
    fprintf('training pair %3d/%d: (%d,
%d)\n',p,mChoose2,pair(1),pair(2));
    SVMStruct = svmtrain(trainDataPair, trainLabelPair, ...
        'kernel_function','rbf','autoscale','false', ...
        'kernelcachelimit', 20000);
    allSVMs{p} = SVMStruct;
end
close(h);
trainingTime = toc;
fprintf('Total training time: %0.2f seconds\n\n', trainingTime);

% test on all binary SVM pairs

```

```

tic
allPredictions = zeros(length(testlabel),mChoose2);
fprintf('Testing all binary SVM pairs...\n\n');
h = waitbar(0, 'Testing all binary SVM pairs...', 'Name', 'Part (f)');
for i=1:mChoose2
    waitbar(i/mChoose2);
    %pair = allPairs(i,:);
    %fprintf('testing pair %3d/%d: (%d,
%d)\n', i, mChoose2, pair(1), pair(2));
    allPredictions(:,i) = svmclassify(allSVMs{i}, Mtest);
end
close(h);
testTime = toc;
fprintf('Total test time: %0.2f seconds\n\n', testTime);

% majority vote
yPredictions = mode(allPredictions,2);
% overall CCR
CCR = sum(yPredictions==testlabel)/length(testlabel);
fprintf('Overall CCR: %0.4f\n\n', CCR);

% confusion matrix of test set
conf = confusionmat(testlabel,yPredictions)';
disp(conf);
testLabelTotals = accumarray(testlabel(:),1);

% double check that confusion matrix columns sum to label totals
fprintf('\n\nconfusion matrix column totals:\n');
disp(sum(conf))
fprintf('test data label totals:\n');
disp(testLabelTotals')
fprintf('conf mat totals - test label totals:\n');
disp(sum(conf)-testLabelTotals')
fprintf('Seems to be missing one in classification for doc #19...\n\n');

% determine most commonly classified label
[~,maxInd] = max(sum(conf,2));
mostCommonDoc = classes(maxInd);
fprintf('Most commonly classified document label: %s (label #%d)\n\n', ...
char(mostCommonDoc),maxInd);

```

Starting part (f)...

Training all binary SVM pairs with rbf kernel...

Total training time: 56.37 seconds

Testing all binary SVM pairs...

Total test time: 155.93 seconds

Overall CCR: 0.3141

Columns 1 through 13

	56	0	0	0	0	0	0	1	1	2	0
0	0										
	2	136	38	18	4	30	2	1	0	0	0
6	4										
	0	4	131	7	2	15	0	1	0	0	0
1	0										
	0	2	23	133	30	2	41	0	0	0	0
0	3										
	0	1	3	0	71	0	9	0	0	0	0
0	3										
	0	0	7	0	1	82	0	1	0	0	0
0	0										
	0	1	0	2	2	3	110	1	1	2	1
0	1										
	0	0	0	0	0	0	5	56	1	0	0
0	0										
	0	0	0	0	0	0	0	0	118	0	0
0	0										
	0	0	1	1	0	0	1	1	0	107	13
0	0										
	0	0	0	1	1	0	0	0	0	15	179
0	0										
	1	1	1	1	0	4	1	0	0	0	0
87	2										
	6	106	58	140	132	120	115	67	18	27	19
19	188										
	88	88	71	53	76	76	45	113	127	122	54
73	108										
	1	2	1	2	1	5	2	0	0	0	1
0	0										
	20	0	0	0	0	0	0	0	0	0	1
0	0										
	142	48	57	34	63	53	51	153	131	122	131
209	84										
	0	0	0	0	0	0	0	0	0	0	0
0	0										
	0	0	0	0	0	0	0	0	0	0	0
0	0										
	2	0	0	0	0	0	0	0	0	0	0
0	0										

Columns 14 through 20

2	0	13	0	3	1	12
1	3	2	1	0	1	2
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	1	0	0	0	0	0
1	0	0	0	0	0	0
4	0	0	2	0	0	0

0	0	0	1	0	0	0
0	0	0	0	0	0	0
1	0	0	0	2	0	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
37	15	25	5	5	4	5
327	162	110	12	48	37	62
0	95	1	0	0	0	0
2	0	76	0	0	0	23
18	116	170	342	294	239	135
0	0	0	0	24	0	0
0	0	0	0	0	27	0
0	0	0	0	0	0	12

confusion matrix column totals:
Columns 1 through 13

318	389	391	392	383	390	382	395	397	397	399
395	393									

Columns 14 through 20

393	392	398	364	376	309	251
-----	-----	-----	-----	-----	-----	-----

test data label totals:
Columns 1 through 13

318	389	391	392	383	390	382	395	397	397	399
395	393									

Columns 14 through 20

393	392	398	364	376	310	251
-----	-----	-----	-----	-----	-----	-----

conf mat totals - test label totals:
Columns 1 through 13

0	0	0	0	0	0	0	0	0	0	0
0	0									

Columns 14 through 20

0	0	0	0	0	-1	0
---	---	---	---	---	----	---

Seems to be missing one in classification for doc #19...

Most commonly classified document label: talk.politics.guns (label #17)

Published with MATLAB® R2017a