Text Mining Final-Project by Thomas Wagner, Alexander Allen, MingYi Wang

Table of Contents

May 14 2015

Input Data

```
clear all
close all

% load 10% sample data and the test data
load DDISample.mat

% convert to normal format instead of sparse
Classp_train=full(Classp_train);
Classm_train=full(Classm_train);
Classp_test=full(Classp_test);
Classm_test=full(Classm_test);
```

PCA on data

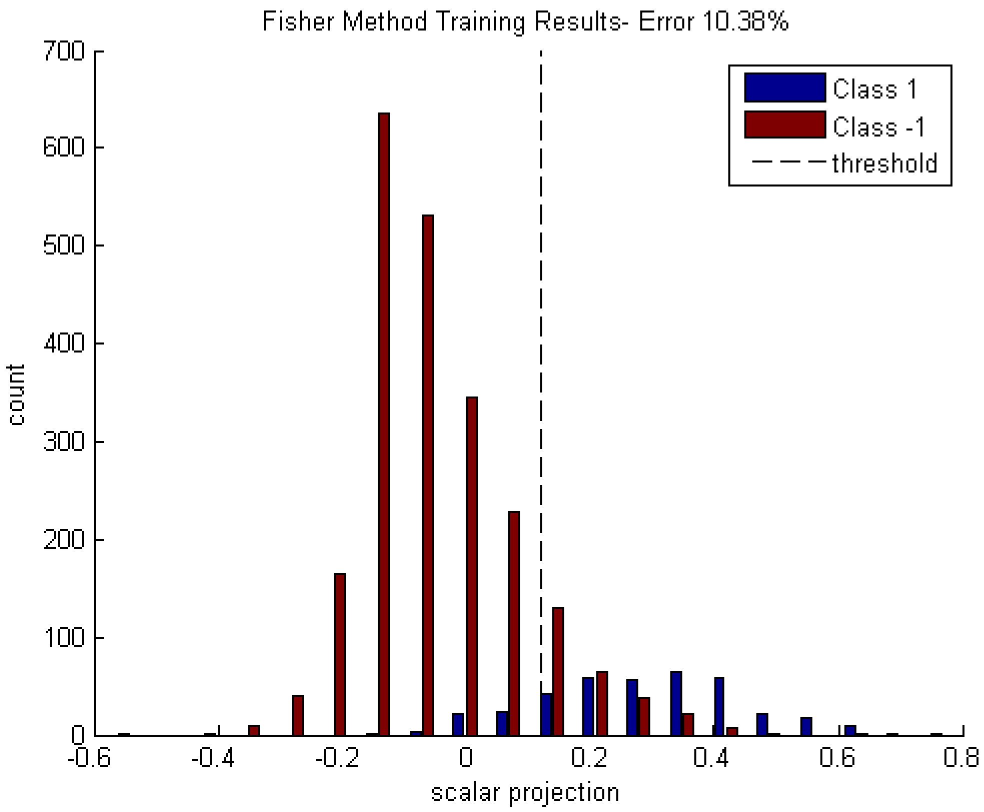
```
Train_total = [Classp_train; Classm_train];

[mp,np] = size(Classp_train);    % size for Classp
[mm,nm] = size(Classm_train);    % size for Classm
[m,n] = size(Train_total);      % size for total

train_mean = (1/m)*(ones(1,m)*Train_total);

Train_total2 = Train_total - ones(m,1)*train_mean;

[eigenvectors, scores, eigenvalues] = pca(Train_total);

trim = 337
trimmed_scores = scores(:,1:trim);
classp_scores = trimmed_scores(1:mp,:);
classm_scores = trimmed_scores(mp+1:m,:);
```

```
trim =

   337
```

Fisher

```
meanp=mean(classp_scores);
meanm=mean(classm_scores);

psize=size(classp_scores,1)
nsize=size(classm_scores,1)
Bp=classp_scores-ones(psize,1)*meanp;
Bn=classm_scores-ones(nsize,1)*meanm;

Sw=Bp'*Bp+Bn'*Bn;
wfisher = Sw\(meanp-meanm)';
wfisher=wfisher/norm(wfisher);

tfisher=(meanp+meanm)./2*wfisher
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Analyze training data  results of the Fisher Linear Discriminant

FisherPosErrorTrain = sum(classp_scores*wfisher <= tfisher);
FisherNegErrorTrain = sum(classm_scores*wfisher >= tfisher);

FisherTrainError= ((FisherPosErrorTrain + FisherNegErrorTrain)/(size(trimmed_score

% Histogram of Fisher Training Results
HistClass(classp_scores,classm_scores,wfisher,tfisher,...
    'Fisher Method Training Results',FisherTrainError);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
psize =

   380


nsize =

        2221


tfisher =

    0.1204


FisherTrainError =
            0.1038
```



Top 30 words

```
A = Train_total' * trimmed_scores * wfisher;
absA = abs(A);

words = cell(30,1);
word_values = zeros(30,1);

for i = 1:30
    [M,I] = max(absA);
    words(i,1) = featurenames(I,1);
    word_values(i,1) = A(I,1);
    absA(I,1) = 0;
end
```

Compute Test Scores

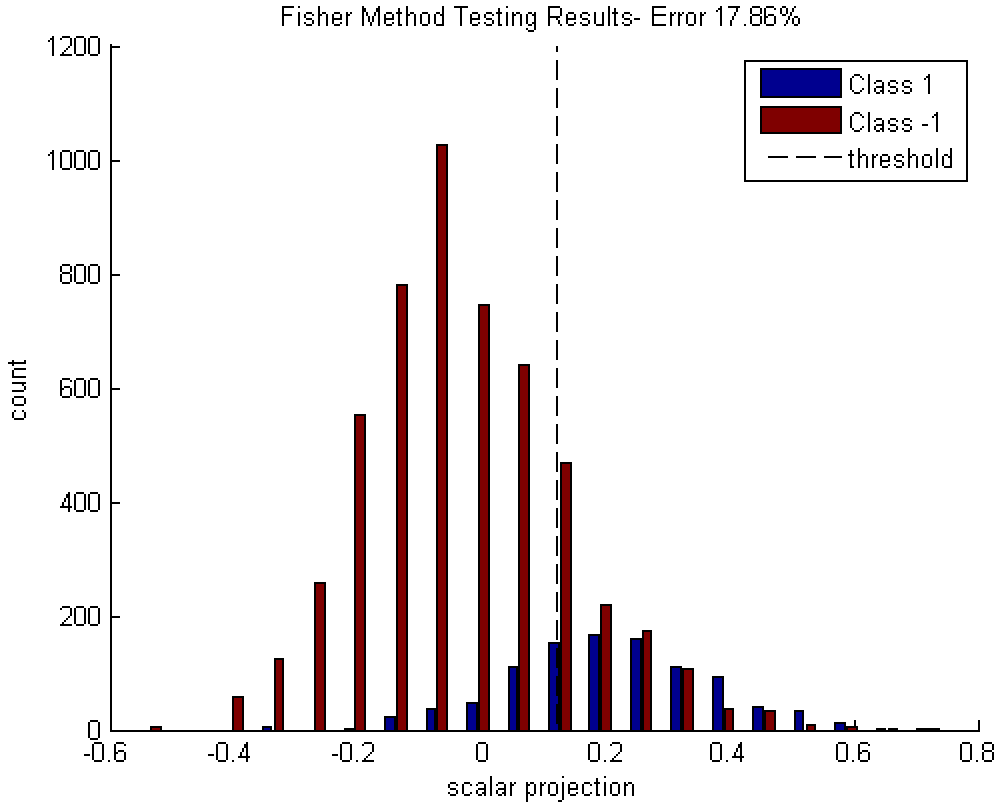
```
Test_total = [Classp_test; Classm_test];

[mp_test,np_test] = size(Classp_test);    % size for Classp
[mm_test,nm_test] = size(Classm_test);    % size for Classm
[m_test,n_test] = size(Test_total);      % size for total
```

```
Test_total2 = Test_total - ones(m_test,1)*train_mean;
Classp_test2 = Test_total2(1:mp_test,:);
Classm_test2 = Test_total2((mp_test+1):end,:);

Classm_test_scores = Classm_test2 * eigenvectors;
Classp_test_scores = Classp_test2 * eigenvectors;

scores_test_total = [Classp_test_scores; Classm_test_scores];


trimmed_scores_test = scores_test_total(:,1:trim);
classp_test_scores = trimmed_scores_test(1:mp_test,:);
classm_test_scores = trimmed_scores_test(mp_test+1:m_test,:);
```

Fisher on Test

```
FisherPosErrorTest = sum(classp_test_scores*wfisher <= tfisher);
FisherNegErrorTest = sum(classm_test_scores*wfisher >= tfisher);

FisherTestError= ((FisherPosErrorTest + FisherNegErrorTest)/(size(trimmed_scores_t

% Histogram of Fisher Testing Results
HistClass(classp_test_scores,classm_test_scores,wfisher,tfisher,...
    'Fisher Method Testing Results',FisherTestError);
```

```
FisherTestError =

    0.1786
```



Fisher Method Testing Results- Error 17.86%

Experimentation results

```
%RESULTS size = 50     21.45% training, 21.91% testing
%RESULTS size = 100    17.69% training, 20.28% testing
%RESULTS size = 150    15.30% training, 18.42% testing
%RESULTS size = 200    13.99% training, 18.40% testing
%RESULTS size = 250    12.76% training, 18.48% testing
%RESULTS size = 300    11.8% training, 18.24% testing
%RESULTS size = 320    11.1% training, 18.15% testing
%RESULTS size = 330    11.8% training, 17.94% testing
%RESULTS size = 337    10.38% training, 17.86% testing
%RESULTS size = 340    10.34% training, 18.29% testing
%90.33 variance explained and elbow is visible
%RESULTS size = 400    9.34% training, 18.69% testing
```

```
%Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
%this error occurs using as low as size 50
```

Published with MATLAB® R2014a