SEIS 763: Machine Learning
Garth Mortensen, mort0052@stthomas.edu

Graduate Program in Software
SEIS 763: ML
Assignment #5 (100 points)
Due Date: July 9th, 2018

The dataset on the Blackboard contains various measurements (i.e. size, center, etc) from thousands of bacterium under microscope. The last column with non-zero values indicate the bacterium are interesting enough for further study. Otherwise (i.e. last column with zero values), those bacterium are not interesting candidates for further study. Convert this dependent variable to binary values. Standardize predictors first using Z score.

Perform an analysis on this dataset using the Support Vector Machine method.

Answer the following questions:

1. **How many support vectors did you find?**

There were 228 support vectors.

2. List top 3 records that have the smallest ** **absolute**** values from w T • X + b calculation.

yhat = theta^{T}X

```
[label, score] = predict(SVMModel, XAll_scaled)

absolute = abs(score); %this strangely returns 2 identical columns
S = sortrows(absolute);
S(1:3,1:1)
```

ans = 0.0034, 0.0048, 0.0050

3. What are the "w^T • X + b" values for the following records: 131, 165, 892, 1057? Anything special about those values of these few records?

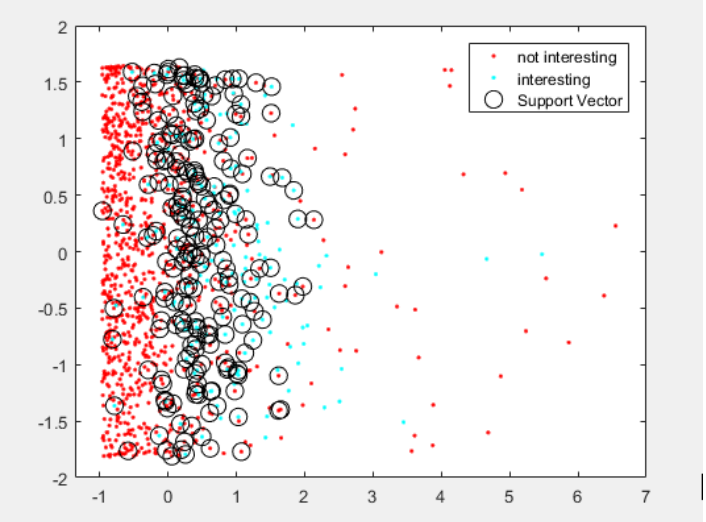
score(131, 1)
ans = 21.1794

score(165, 1)
ans = -9.3468

score(892, 1)
ans = -4.5855

score(1057, 1)
ans = 26.2964

Can you explain what is supposed to be special about these records? They don't seem to be support vectors.



```
%SEIS763 Machine Learning
%Garth Mortensen, mort0052@stthomas.edu
%Assignment 5

%% clear all
% clc
% close all

filename = 'C:\tmp\CellDNA.csv';
delimiter = ',';
formatSpec = '%f%f%f%f%f%f%f%f%f%f%f%f%f%f%[^\n\r]';
fileID = fopen(filename,'r');
dataArray = textscan(fileID, formatSpec, 'Delimiter', delimiter,
'ReturnOnError', false);
fclose(fileID);
VarName1 = dataArray{:, 1};
VarName2 = dataArray{:, 2};
VarName3 = dataArray{:, 3};
VarName4 = dataArray{:, 4};
VarName5 = dataArray{:, 5};
VarName6 = dataArray{:, 6};
VarName7 = dataArray{:, 7};
VarName8 = dataArray{:, 8};
VarName9 = dataArray{:, 9};
VarName10 = dataArray{:, 10};
VarName11 = dataArray{:, 11};
VarName12 = dataArray{:, 12};
VarName13 = dataArray{:, 13};
VarName14 = dataArray{:, 14};
clearvars filename delimiter formatSpec fileID dataArray ans;


%% Target------------------------------------------------------------
Y = VarName14;
% clear VarName14;

%Convert this dependent variable to binary values.
%Y = 0 means bacterium is not interesting to study
%We need binary vector
%If Y > 0, set equal to 1
Y(Y ~= 0) = 1;
%for mnrfit(), we need categorical response variable, not numeric
Y_cat = categorical(Y);

%% Matrix------------------------------------------------------------
XAll = [VarName1 VarName2 VarName3 VarName4 VarName5 VarName6 VarName7,...
    VarName8 VarName9 VarName10 VarName11 VarName12 VarName13];

% Standardize numerics----------------------------------------------
% Standardize predictors first using Z score
% various numeric measurements (i.e. size, center, etc) from thousands
% of bacterium under microscope. All the measurements are
% in different units.
XAll_scaled = zscore(XAll);
X_mean = mean(XAll_scaled); %check mean ~= 0
```

```
%% Support Vector Machine----------------------------------------
SVMModel = fitcsvm(XAll_scaled, Y_cat)

%% diagnosis
% 1. How many support vectors did you find?
countSupportVectors = sum(SVMModel.IsSupportVector == 1);
beta0 = SVMModel.Bias;
betas = SVMModel.Beta;

%ImportantRecords = [SVMModel.IsSupportVector SVMModel.alpha]

%% Diagnosis 2
% score = wtX + b. shows criticalness of point. if 0, right on decision
% boundary. If -1, right on bottom right.
% label = predicted label e.g. (-1, 1)
%SVM, slide 30
[label, score] = predict(SVMModel, XAll_scaled); %why does score give 2
columns?

%% 2. List top 3 records that have the smallest ** absolute** values from w T
• X + b calculation.
absolute = abs(score); %this returns 2 identical columns

%I need top 3, not top 1
%[M,I] = min(absolute)
%Minimums = topkrows(absolute, 3, 'ascend') %introduced in 2016b, i have
%2016a.

%here is a workaround
%https://www.mathworks.com/matlabcentral/answers/371490-what-am-i-doing-
wrong-with-the-function-topkrows
S = sortrows(absolute);
S(1:3,1:1)


%% 3. What are the "wT • X + b" values for the following records:
% 131, 165, 892, 1057?
score(131, 1)
score(165, 1)
score(892, 1)
score(1057, 1)

% Anything special about those values?
```