Problem3

Generate data:

```
data = load('data/spambase.data');
X = data(:,1:57); Y = data(:,end);
[X Y] = reorderData(X,Y);
[Xt,Xv,Yt,Yv] = splitData(X,Y,.6);
[Xt,S] = rescale(Xt); Xv = rescale(Xv,S);
```

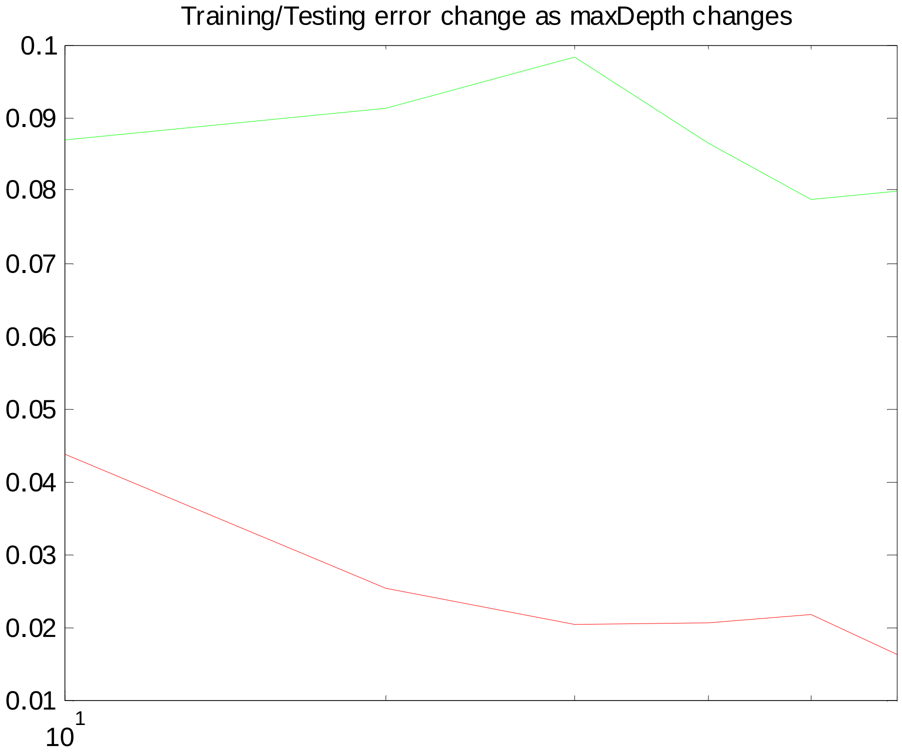
a.

```
[N,D] = size(data);
tc = treeClassify(Xt,Yt,10,10,0.01,round(D/2)); %Default
etr = mean(Yt ~= predict(tc,Xt))  %training data
ete = mean(Yv ~= predict(tc,Xv))  %validation data

etr =

    0.0424

ete =

    0.0788
```

nParent = 10 :   minimum # of data required to split a node
maxDepth = 10 : maximum depth of the decision tree
minScore: 0.01 : minimum value of the score improvement to split a node
nFeatures: round(D/2) : # of available features for splitting at each node

b.

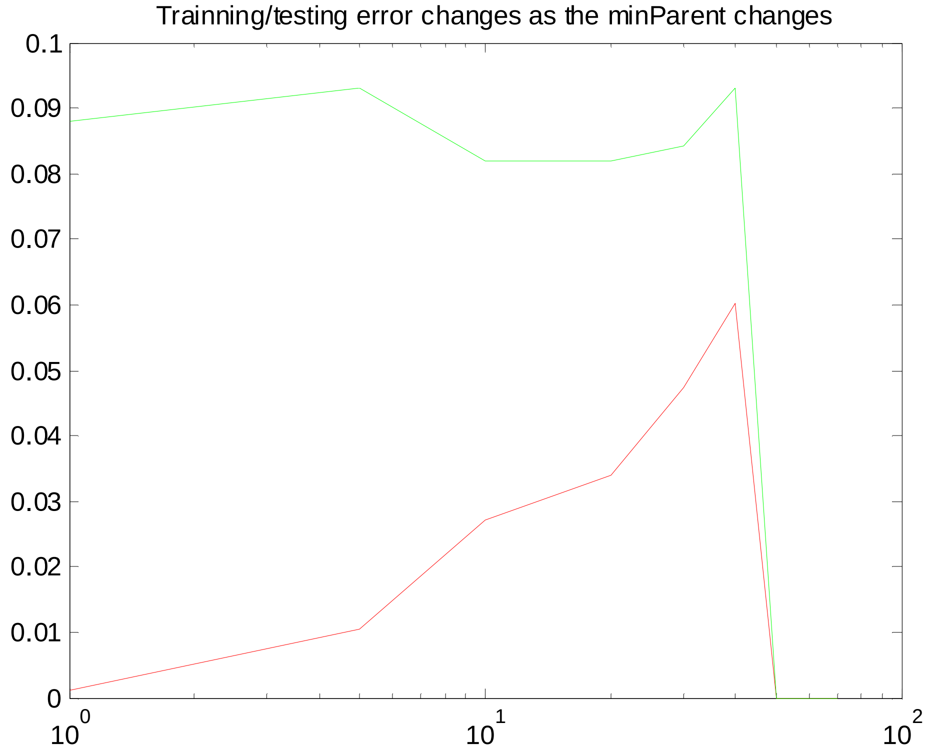
```
%maxDepth
K = [10,20,30,40,50,60];
etr = 0*K; ete = 0*K;
for i = 1:6,
   tc = treeClassify(Xt,Yt,10,K(i),0.01,round(D/2));
   %report the training and validation accuracy
   etr(i) = mean(Yt ~= predict(tc,Xt));  %training data
   ete(i) = mean(Yv ~= predict(tc,Xv));  %validation data
end
figure; semilogx(K,etr,'r-',K,ete,'g-');
```



From the above picture, we should pick maxDepth = 50. Because it gives us a lowest test error and gives us a low training error.

c.

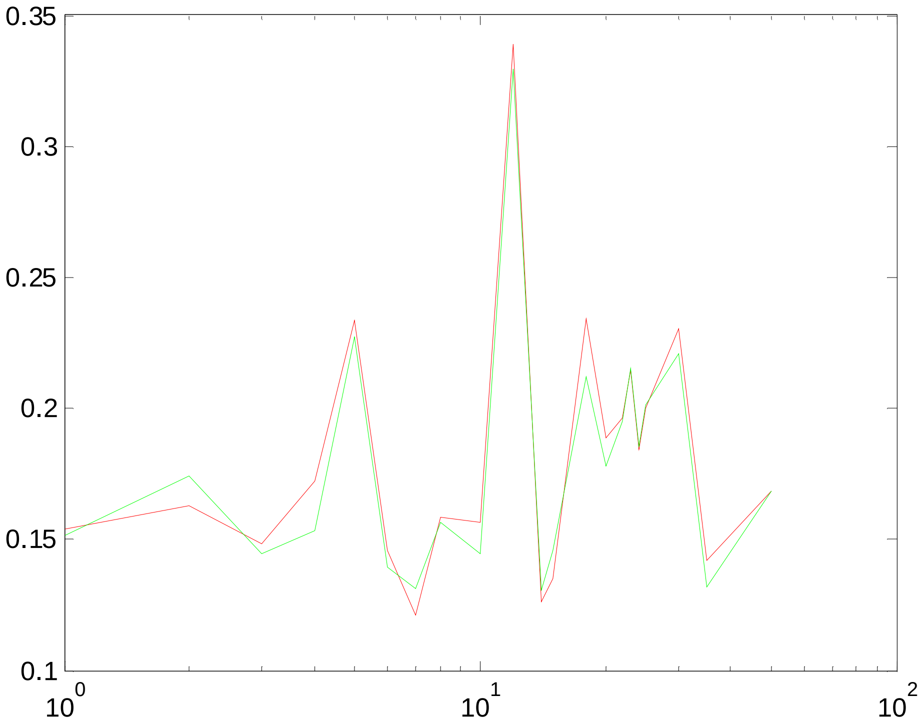
```
%nParent
K = [10,20,30,40,50,60];
etr = 0*K; ete = 0*K;
for i = 1:6,
   tc = treeClassify(Xt,Yt,K(i),100000,0.01,round(D/2));
   %report the training and validation accuracy
   etr(i) = mean(Yt ~= predict(tc,Xt));  %training data
   ete(i) = mean(Yv ~= predict(tc,Xv));  %validation data
end
figure; semilogx(K,etr,'r-',K,ete,'g-');
```



The above picture shows that we should pick the minParent in the range 10 to 25. Because it gives us a lowest test error and gives us a low training error.

d.

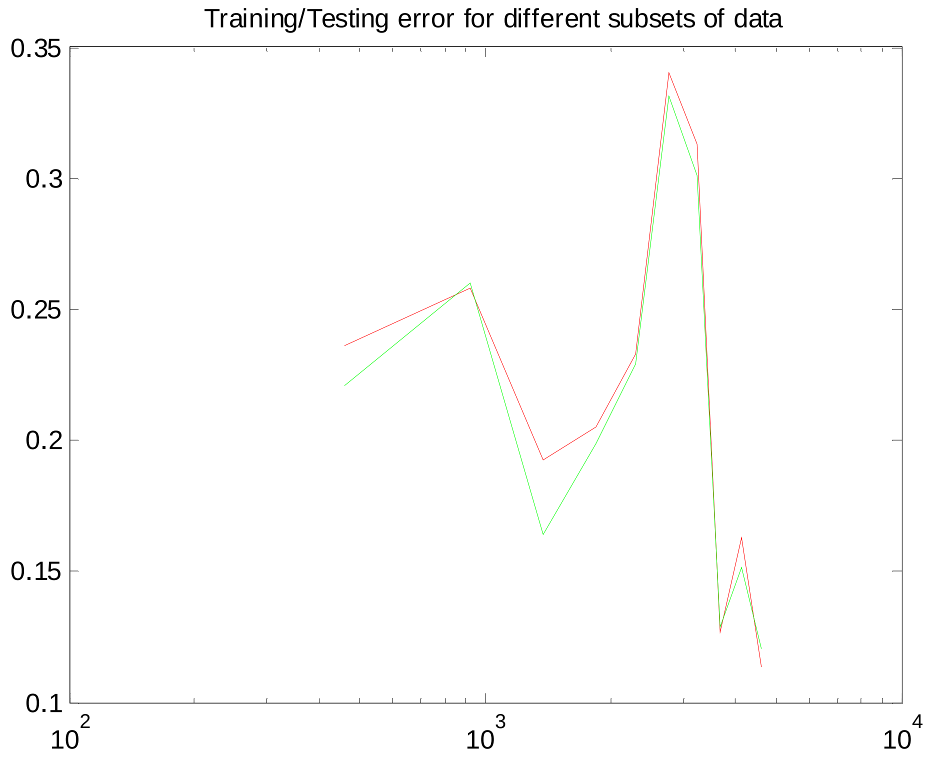
```
[N,D] = size(X);

Nbag =  [1 2 3 4 5 6 7 8 10 12 14 15 18 20 22 23 24 25 30 35 50];
etr = 0*Nbag; ete = 0*Nbag;
for j = 1:length(Nbag)
Classifiers = cell(1,Nbag(j));
for i = 1:Nbag(j)
   [Xboot Yboot] = bootstrapData(X,Y,N);%0.1*N
   Classifiers{i} = treeClassify(Xboot,Yboot);
end;

%Test data Xtest

[Ntest,D] = size(Xt);
predictYt = zeros(Ntest,Nbag(j));
[Nvalid,D] = size(Xv);
predictYv = zeros(Nvalid,Nbag(j));
for i = 1:Nbag(j),
   %report the training and validation accuracy
   predictYt(:,i) = predict(Classifiers{i},Xt);
   predictYv(:,i) = predict(Classifiers{i},Xv);
end;
predictYt_f = (mean(predictYt,2) > 0.5);
predictYv_f = (mean(predictYv,2) > 0.5);
etr(j) = mean(Yt ~= predict(Classifiers{i},Xt));  %training data
ete(j) = mean(Yv ~= predict(Classifiers{i},Xv));  %validation data
end

figure; semilogx(Nbag,etr,'r-',Nbag,ete,'g-');
```



The figure shows that the training error and test error change unregularly as the number of bags changes. So if we want to find the best number for bags, we need to learn it from specific data set. For this problem, we should pick up 6~8, 13~15,30~40 as the number of bags.

e.

```
[N,D] = size(X);
Nbag = 25;
Classifiers = cell(1,Nbag);
K = [0.1*N 0.2*N 0.3*N 0.4*N 0.5*N 0.6*N 0.7*N 0.8*N 0.9*N N];
etr = 0*length(K); ete = 0*length(K);
for k = 1:length(K)
for i = 1:Nbag
   [Xboot Yboot] = bootstrapData(X,Y,K(k));%0.1*N
   Classifiers{i} = treeClassify(Xboot,Yboot);
end;

%Test data Xtest
[Ntest,D] = size(Xt);
predictYt = zeros(Ntest,Nbag);
[Nvalid,D] = size(Xv);
predictYv = zeros(Nvalid,Nbag);
for i = 1:Nbag,
   %report the training and validation accuracy
   predictYt(:,i) = predict(Classifiers{i},Xt);
   predictYv(:,i) = predict(Classifiers{i},Xv);
end;
predictYt_f = (mean(predictYt,2) > 0.5);
predictYv_f = (mean(predictYv,2) > 0.5);
etr(k) = mean(Yt ~= predict(Classifiers{i},Xt));  %training data
ete(k) = mean(Yv ~= predict(Classifiers{i},Xv));  %validation data
end

figure; semilogx(K,etr,'r-',K,ete,'g-');
```



The figure shows that the training error and test error change unregularly as the number of the resample subsets of data changes. But as the subset of data comes close to the while original dataset, the error decreases a lot. So if we want to find the best number for resample dataset, we need to learn it from specific data set. For this problem, we can pick up 1300~1500, 3800~4700 as the number of resample dataset.