FES758b / S&DS363 / S&DS563
Multivariate Statistics
Homework #4 : Discriminant Analysis
Due : Friday, 3/8/2019 11:59 pm on CANVAS

HOMEWORK ASSIGNMENT

PLEASE turn in the following answers for YOUR DATASET! If Discriminant Analysis is not appropriate for your data, use one of the two loaner datasets described on the next page.

List names/emails of everyone who worked on this assignments.

NOTE – some questions below will not be appropriate if you have only two groups, or if you only have two possible discriminating variables. Modify assignment as appropriate!

1. Evaluate the assumptions implicit to Discriminant Analysis for your data – multivariate normality WITHIN each group (i.e. chi-square quantile plots) and similarity of covariances matrices (look at Box's M or just look at raw standard deviations/covariance matrices). Comment on what you find. Comment on whether you think transformations might help your data to meet the assumptions of DA. If you think they might, make some transformations and find out! You might also want to make a matrix plot (or a pairs plot) to get a sense of what your data looks like two variables at a time (use different symbols for each group).

2. Perform stepwise discriminant analysis on your data. Comment on which model seems the best. Use quadratic discriminant analysis if appropriate. If you end up with only one significant discriminating variable, you might want to just force a second variable in the model (i.e. add a technically 'non-significant' discriminator).

3. Comment on whether there is statistical evidence that the multivariate group means are different (i.e. Wilks Lambda test).

4. How many discriminant functions are significant? What is the relative discriminating power of each function?

5. Use classification, both regular and leave-one-out (or cross-validation) to evaluate the discriminating ability of your functions.

6. Provide some evidence as to which of your original variables are the 'best' discriminators amongst your groups (look at standardized discriminant coefficients).

7. Make score plots for the first two or three DA function scores (be sure to use different symbols/colors for each group). Comment on what you see.

8. Bonus (and optional)– try kernel smoothing or k-nearest neighbors and get the admiration of your professor and TA (and some extra credit)! You'll have to use SAS or R for this.

LOANER DATASETS (choose one)
(if Discriminant Analysis is not appropriate for your data)

1) alzheim.xls and alzheim.sas (actually a program that makes the datafile). Data and sas file are from Michael Friendly at York University. This file contains data from study by Robert Hart concerning ability of memory tests to distinguish between patients with mild Alzheimer's disease, depression, and controls. The grouping variable is called GROUP. The other variables are

```
subj    = 'Subject ID'
trecall = 'Total recall'
hirecall= 'HI Imagery Recall'
lirecall= 'LO Imagery Recall'
tunrem  = 'Unreminded Memory'
hiunrem = 'HI Imagery Unreminded Memory'
liunrem = 'LO Imagery Unreminded Memory'
store   = 'Storage'
recog   = 'Recognition Memory';
```

NOTE that trecall is just the sum of hirecall and lirecall (i.e. you can't use all three of these variables!)

2) The file RemoteSensing.csv contains data from satellite imagery. The goal is to use the amount of light reflectance in four bandwidths to predict crop. The variables are as follows:

- Crop : Either corn, soybeans, cotton, or clover. This the grouping variable (i.e. you are trying to discriminate amongst these crops).
- X1 : amount of light reflectance in the first of four bandwidths
- X2 : amount of light reflectance in the second of four bandwidths
- X3 : etc.
- X4 : etc.

SAMPLE DATA SET

The example below is JUST FOR YOUR PRACTICE. NOTHING TO TURN IN HERE!

IRIS.csv contains data analyzed initially by R.A. Fisher (grandfather of statistics). In consists of four different measurements taken on three different species of iris, 50 replications per species. The variables are species (group variable), petal length and width, and sepal length and width.

SAS SOLUTION

1. Evaluate the assumptions implicit to Discriminant Analysis for your data – multivariate normality WITHIN each group (i.e. chi-square quantile plots) and similarity of covariances matrices (look at Box's M or just look at raw standard deviations). Comment on what you find. Comment on whether you think transformations might help your data to meet the assumptions of DA. If you think they might, make some transformations and find out!

```
*Get data from online;
filename foo url 'http://reuningscherer.net/stat660/data/iris.csv';
PROC IMPORT OUT= WORK.iris
            DATAFILE= foo
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

*check multivariate normality of the residuals;
*INCLUDE THE MACRO;
filename foo2 url "http://reuningscherer.net/stat660/sas/multnorm.sas.txt";
%include foo2;


* make subgroup of the data;
data setosa; set iris; if var8='Setosa';  run;
%multnorm(data=setosa, var=sepallen sepalwid petallen petalwid)

data versicolor; set iris; if var8='Versic';  run;
%multnorm(data=versicolor, var=sepallen sepalwid petallen petalwid)

data virginica; set iris; if var8='Virgin';  run;
%multnorm(data=virginica, var=sepallen sepalwid petallen petalwid)
```

Chi-square quantile plots look fine for each group.

2. Perform stepwise discriminant analysis on your data (unless using R, in which case, fit a few different models for different sets of variables). Comment on which model seems the best. Use quadratic discriminant analysis if appropriate.

```
proc stepdisc data= iris;
class var8;
var sepallen sepalwid petallen petalwid;
run;

* make plots to evaluate equality of covariance matrices;
Proc GPlot DATA=iris;
 PLOT sepallen*sepalwid = var8;
 PLOT petallen*petalwid = var8s;
 Symbol1 V=Dot    H=1 I=None C=Red;
 Symbol2 V=Star   H=1 I=None C=Blue;
 Symbol3 V=Square   H=1 I=None C=Green;
Run;

*run quadratic DA and get output to answer following questions;
proc discrim data=iris pool=no out=outiris crossvalidate manova canonical;
class var8;
var sepallen sepalwid petallen petalwid;
run;
```

The SAS output shows that all four variables are added in this process. So, model with all four predictors is the best.

Plots (not shown here) suggest that there may be inequality of covariance matrices, so try quadratic discriminant analysis. Also, when DA is performed, log determinants are 5.3, 7.5, and 9.5, suggests inequality of covariance matrices.

3. Comment on whether there is statistical evidence that the multivariate group means are different.

The MANOVA option in proc discrim gives this output. Wilks lambda is significant so we conclude multivariate means are not the same.

Multivariate Statistics and F Approximations

S=2 M=0.5 N=71

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02343863	199.15	8	288	<.0001
Pillai's Trace	1.19189883	53.47	8	290	<.0001
Hotelling-Lawley Trace	32.47732024	582.20	8	203.4	<.0001
Roy's Greatest Root	32.19192920	1166.96	4	145	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

4. How many discriminant functions are significant?

The CANONICAL option in proc discrim provides this output. There are 3 groups and 4 variables so min(3-1,4)=2 discriminant functions. Output shows that both are significant.

Canonical Discriminant Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.984821	0.984508	0.002468	0.969872
2	0.471197	0.461445	0.063734	0.222027

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	32.1919	31.9065	0.9912	0.9912	0.02343863	199.15	8	288	<.0001
2	0.2854		0.0088	1.0000	0.77797337	13.79	3	145	<.0001

5. Use classification, both regular and leave-one-out to evaluate the discriminating ability of your functions.

Classification Summary for Calibration Data: MYLIB.IRIS
Resubstitution Summary using Quadratic Discriminant Function
Number of Observations and Percent Classified into Species1

From Species1	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	48	2	50
	0.00	96.00	4.00	100.00
Virginica	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
	33.33	32.67	34.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species1

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species1

From

Species1	Setosa	Versicolor	Virginica	Total
Setosa	50	0	0	50
	100.00	0.00	0.00	100.00
Versicolor	0	47	3	50
	0.00	94.00	6.00	100.00
Virginica	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	48	52	150
	33.33	32.00	34.67	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Species1

	Setosa	Versicolor	Virginica	Total
Rate	0.0000	0.0600	0.0200	0.0267
Priors	0.3333	0.3333	0.3333	

Classification makes 3 errors for resubstitution and 54 errors for crossvalidation – so error rate is 2-3%, very good!

6. Provide some evidence as to which of your original variables are the 'best' discriminators amongst your groups.

The DISCRIM Procedure
Canonical Discriminant Analysis

Total-Sample Standardized Canonical Coefficients

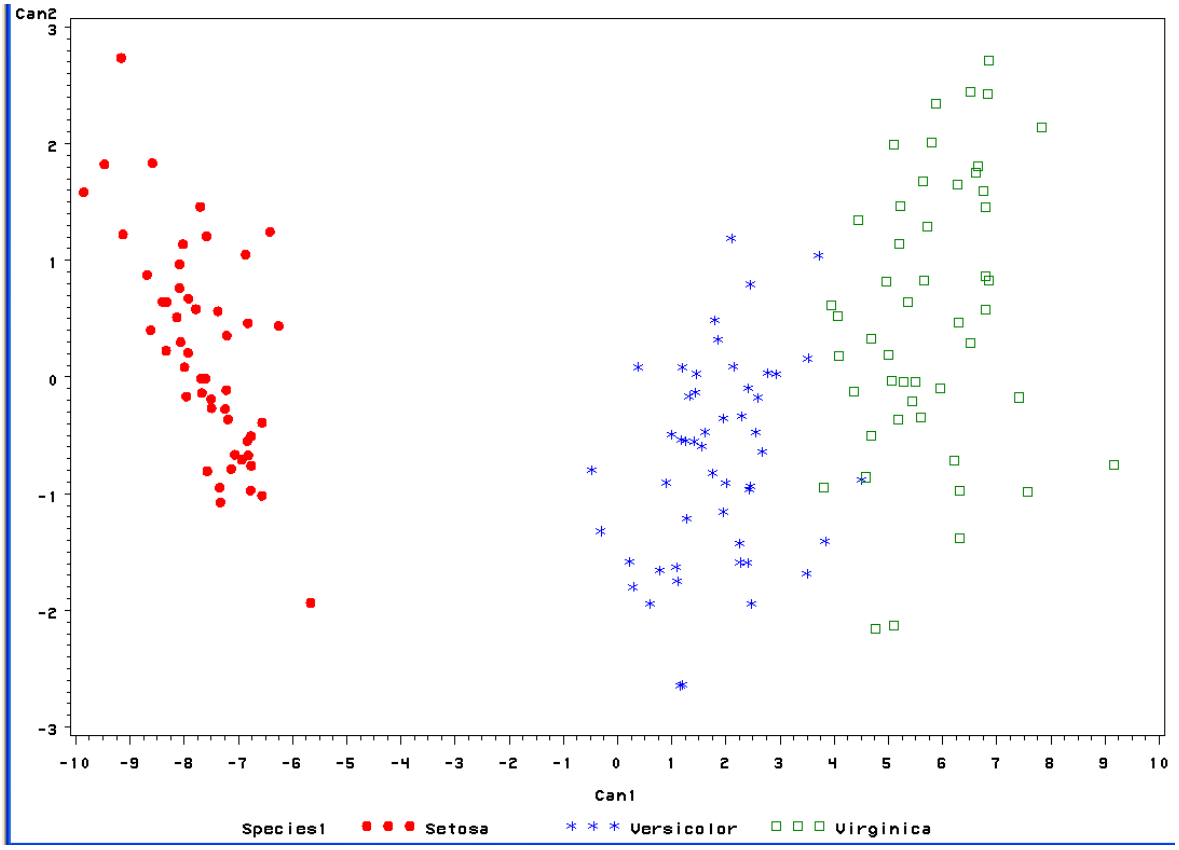
Variable	Label	Can1	Can2
sepallen	sepallen	-0.686779533	0.019958173
sepalwid	sepalwid	-0.668825075	0.943441829
petallen	petallen	3.885795047	-1.645118866
petalwid	petalwid	2.142238715	2.164135931

Standardized coefficients are shown above. The largest coefficient for DA function one is for petal length, followed by petal width. Petal variables still have the largest coefficients for the second DA function – SO, petal characteristics are the best discriminators.

7. Make a score plot for the first two or thee DA functions. Comment on what you see.

```
PROC GPLOT DATA=OUTIRIS;
  PLOT CAN2*CAN1=SPECIES1;
```

```
RUN;
```



Discrimination is mostly in the direction of DA function 1, but there is some separation of Versicolor from Virginica/Setosa in the direction of DA function 2.

8. Bonus (and optional)– try kernel smoothing or k-nearest neighbors and get the admiration of your professor and TA! You'll have to use SAS or R for this.

Examples of this are given online in irisdiscrim.sas

R SOLUTION

See examples up on Canvas!

SPSS Steps for Discriminant Analysis
Iris data used as example

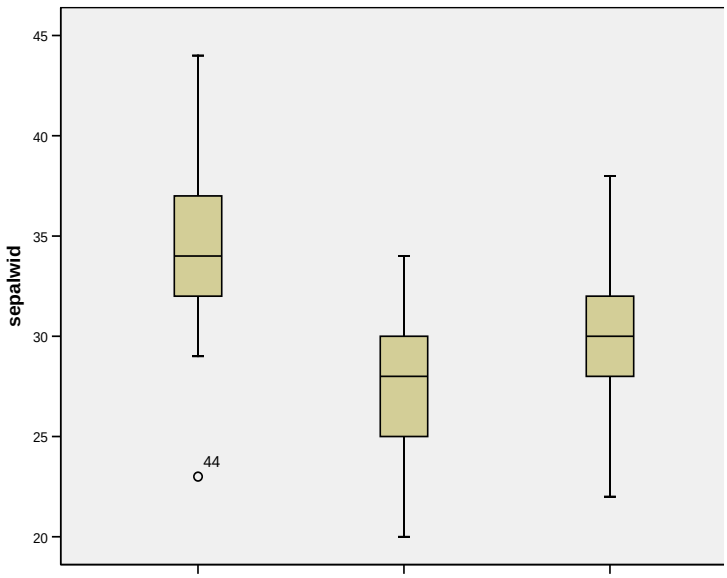
1. Evaluate the assumptions implicit to Discriminant Analysis for your data – multivariate normality WITHIN each group (i.e. chi-square quantile plots) and similarity of covariance matrices (look at Box's M or just look at raw standard deviations). Comment on what you find. Comment on whether you think transformations might help your data to meet the assumptions of DA. If you think they might, make some transformations and find out!

<u>To evaluate univariate normality for each group</u>

Boxplots:

- Use ***Graph→Boxplot***.
- There are a number of options that will work. In the dialogue box, choose ***Simple*** and click on ***summaries of groups of cases***. Then list the variable of interest (here we have 4 variables of interest: sepal length, sepal width, petal length, and petal width) and the category axis (species number or species).

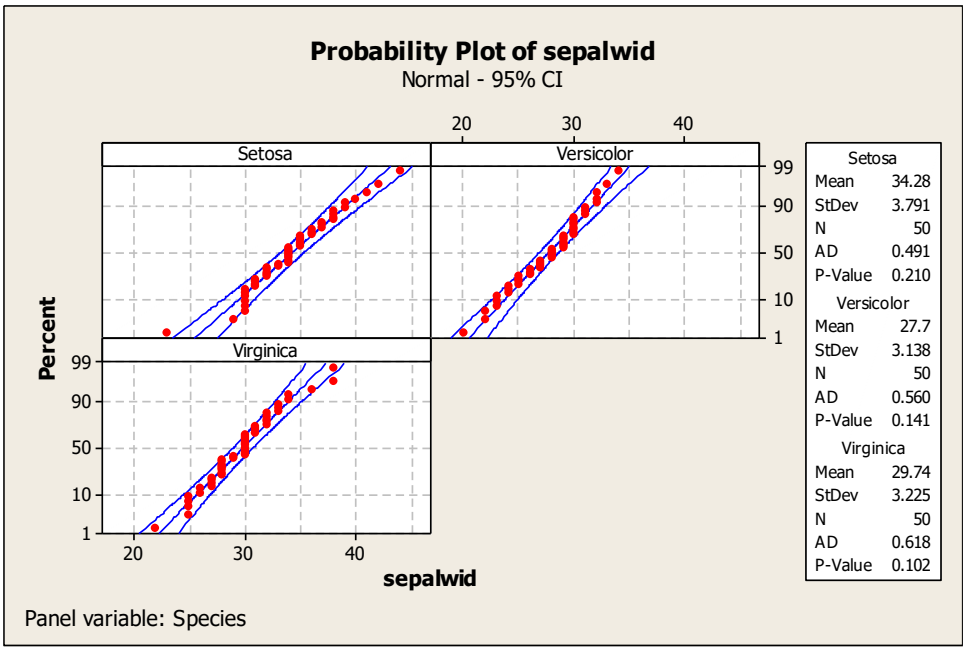
Output for the variable sepal width for each of the 3 groups:



Normal Quantile Plots:

- Unfortunately there is no easy way to make separate normal quantile plots for each species or grouping in SPSS.
- Go to Minitab (or one of the other programs) for this.
- In MINITAB Use ***Graph→Probability Plot***. Chose ***simple*** and select the variables that you want to graph. Click on the button ***multiple graphs*** and select the ***by variables*** tab. When there, select the option that you prefer (there are two presented) and enter your grouping variable (in this case species). I chose the "by variables with groups in separate panels" option.

Output for the variable sepal width for all 3 species:



<u>To make any necessary transformations</u>

Transformations:

- Use ***Transform→Compute***. Select the Target Variable (a new variable you created) and the appropriate numeric expression (e.g., log, natural log, etc.).

<u>To evaluate multivariate normality: Chi-square Quantile Plots</u>

- Like Normal quantile plots, unfortunately it is not possible to create chi-square quantile plots by group in SPSS.
- Use one of the other programs.

<u>To evaluate similarity of covariance matrices</u>

- Run Discriminant Analysis. Use ***Analyze→Classify→Discriminant***. Enter independent variables and the grouping variable (must be numeric rather than text—here species groups are categorized numerically from 1-3).
- You can select either method (stepwise or independents together) for the purposes of evaluating the covariance matrices

- Click on the button ***Statistics*** and select ***Mean*** and ***Box's M***
- Look at the GROUP STATISTICS information to evaluate whether the standard deviations for each group are roughly equal.
- See if the LOG DETERMINANTS are about the same.
- Look at BOX'S TEST OF EQUALITY OF COVARIANCE MATRICES.

Relevant output is shown in class notes and IRIS example.

2. Perform stepwise discriminant analysis on your data (unless using R, in which case, fit a few different models for different sets of variables). Comment on which model seems the best. Use quadratic discriminant analysis if appropriate.

Perform stepwise discriminant analysis

- Run Discriminant Analysis. Use ***Analyze→Classify→Discriminant***. Enter independent variables and the grouping variable (must be numeric rather than text—here species groups are categorized numerically as 1-3).
- Select ***Stepwise***.
- Based on your results from step 1, select either linear or quadratic discriminant analysis.
    - For linear: Leave as is (default is within-groups covariance matrix).
    - For quadratic (unequal covariance matrices): Click on the ***Classify*** button and select ***Separate groups*** under the Use Covariance Matrix Section.
- To specify threshold values for entry, removal of model terms. Click on the ***Method*** button and change the values under ***Criteria***.

Relevant output is shown in class notes and IRIS example.

3. Comment on whether there is statistical evidence that the multivariate group means are different.
4. How many discriminant functions are significant?

- Nothing special required here. See your Discriminant Output.
- You may also want to look at correlation of the possible discriminating variables. Use ***Analyze→Correlate→Bivariate Correlations***. Enter your variables of interest.

Some output from Iris data:

Correlations

		sepallen	sepalwid	petallen	petalwid
sepallen	Pearson Correlation	1	-.118	.872**	.818**
	Sig. (2-tailed)		.152	.000	.000
	N	150	150	150	150
sepalwid	Pearson Correlation	-.118	1	-.428**	-.366**
	Sig. (2-tailed)	.152		.000	.000
	N	150	150	150	150
petallen	Pearson Correlation	.872**	-.428**	1	.963**
	Sig. (2-tailed)	.000	.000		.000
	N	150	150	150	150
petalwid	Pearson Correlation	.818**	-.366**	.963**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	150	150	150	150

**. Correlation is significant at the 0.01 level (2-tailed).

- For each discriminant function, you get a test of whether or not all groups have the same mean discriminant function scores. To see which group means are not equal. On the main DA dialogue box, click on the ***Save*** button and check the ***discriminant scores box***. The scores for each discriminant function will be saved as new columns in your data table. Then Use ***Analyze→One Way ANOVA*** to evaluate differences in pairs of means. Click the ***post hoc*** button and select an appropriate technique (e.g., Tukey or Scheffe's comparisons).

Output from the Iris example:

Multiple Comparisons

Tukey HSD

Dependent Variable	(I) speciesnum	(J) speciesnum	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Discriminant Scores from Function 1 for Analysis 1	1	2	-9.4326494*	.20000000	.000	-9.9061881	-8.9591107
		3	-13.390150*	.20000000	.000	-13.8636890	-12.9166117
	2	1	9.43264942*	.20000000	.000	8.9591107	9.9061881
		3	-3.9575009*	.20000000	.000	-4.4310396	-3.4839623
	3	1	13.390150*	.20000000	.000	12.9166117	13.8636890
		2	3.95750095*	.20000000	.000	3.4839623	4.4310396
Discriminant Scores from Function 2 for Analysis 1	1	2	.94303264*	.20000000	.000	.4694940	1.4165713
		3	-.29763359	.20000000	.300	-.7711723	.1759051
	2	1	-.94303264*	.20000000	.000	-1.4165713	-.4694940
		3	-1.2406662*	.20000000	.000	-1.7142049	-.7671276
	3	1	.29763359	.20000000	.300	-.1759051	.7711723
		2	1.24066623*	.20000000	.000	.7671276	1.7142049

*. The mean difference is significant at the .05 level.

5. Use classification, both regular and leave-one-out to evaluate the discriminating ability of your functions.

- To access these options, on the main DA dialogue box, click on the ***Classify*** button and go to the **Display** section.
- To get a summary of classification results, check the ***summary table box*** (this will show you the "regular" results).
- Leave-one out is only available for Linear DA (within-groups covariance matrix must be selected in the Use Covariance Matrix Section). To select leave-one-out, check the ***leave-one-out classification box***.

Some output from the leave-one-out method:

Classification Results[b,c]

			Predicted Group Membership			
		speciesnum	1	2	3	Total
Original	Count	1	50	0	0	50
		2	0	48	2	50
		3	0	1	49	50
	%	1	100.0	.0	.0	100.0
		2	.0	96.0	4.0	100.0
		3	.0	2.0	98.0	100.0
Cross-validated[a]	Count	1	50	0	0	50
		2	0	48	2	50
		3	0	1	49	50
	%	1	100.0	.0	.0	100.0
		2	.0	96.0	4.0	100.0
		3	.0	2.0	98.0	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 98.0% of original grouped cases correctly classified.

c. 98.0% of cross-validated grouped cases correctly classified.

6. Provide some evidence as to which of your original variables are the 'best' discriminators amongst your groups.

- Nothing special required here.
- Standardized coefficients are provided as part of the regular Discriminant output. If you want to see unstandardized coefficients on the main dialogue box, click the ***Statistics button*** and then check the appropriate box.

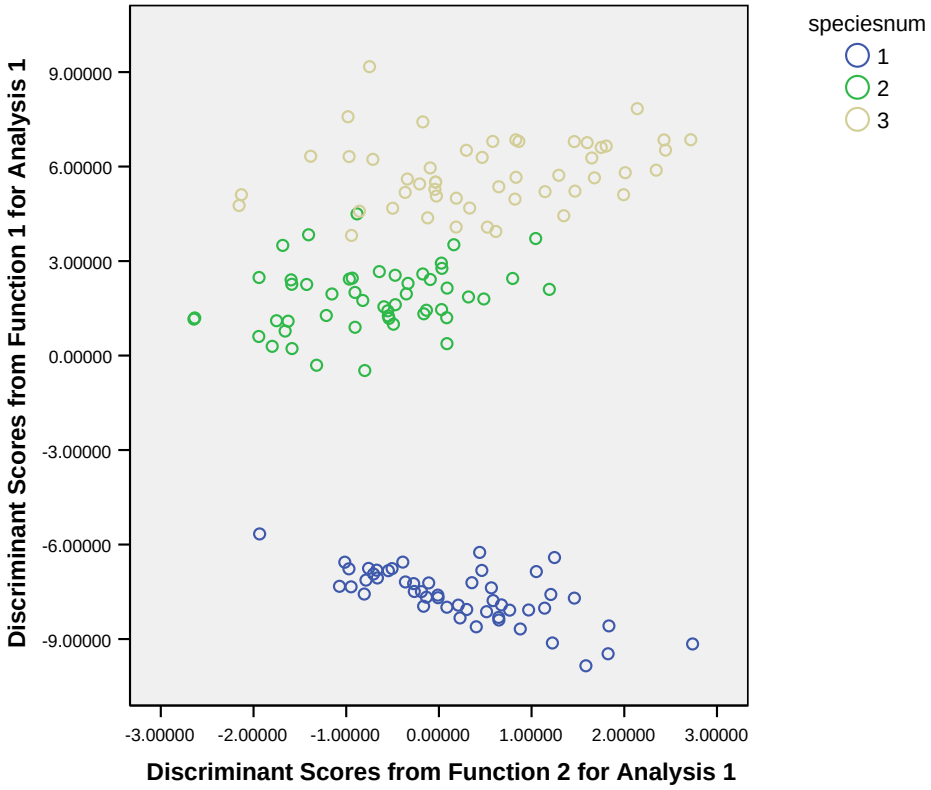
tandardized Canonical Discriminant Function Coefficients

	Function	
	1	2
sepallen	-.427	.012
sepalwid	-.521	.735
petallen	.947	-.401
petalwid	.575	.581

7. Make a score plot for the first two or thee DA functions. Comment on what you see.

- From the main DA dialogue box, click on the ***Save*** button and check the ***Discriminant scores*** box. The scores for each discriminant function will be saved as new columns in your data table.
- Use ***Graphs→Scatterplot*** to graph the saved scores.

Example output:



8. Bonus (and optional)– try kernel smoothing or k-nearest neighbors and get the admiration of your professor and TA! You'll have to use SAS or R for this.

- Sorry, no SPSS here.

OTHER NOTES

Prior Probabilities

- In SPSS the only option is to assign prior probabilities based on group size. To access these options, on the main DA dialogue box, click on the ***Classify*** button. Select the ***computer from group sizes*** box under the Prior Probabilities section.

Misclassification Costs

- Hard to do in SPSS.