

Statistics 503 Exam 1 SOLUTION

Friday, March 11 2005 Open notes

1. (3pts) The following table has numerous problems.

Variable	Correlation
palmitic	-0.1256826
palmitoleic	-0.2779403
stearic	0.07973278
oleic	0.8553294
linoleic	-0.8661034
linolenic	-0.2033158
arachidic	-0.5997424

Table 1: Correlations between predicted values from an LDA classifier and each variable used.

Give three ways to improve the table.

I Round the numbers to 2-3 decimal places.

II Sort the rows from highest to lowest correlation.

III Expand the caption to explain these are correlations of what?

2. (2pts) Which of the following two sets of variance-covariance matrices would be considered to be heterogeneous?

$$A. \mathbf{S}_1 = \begin{bmatrix} 1000 & 10 \\ 10 & 1000 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 500 & 10 \\ 10 & 500 \end{bmatrix}$$

A is an example of heterogeneous variance-covariance as used in the LDA literature. \mathbf{S}_1 and \mathbf{S}_2 are different. This difference between groups is what affects LDA.

$$B. \mathbf{S}_1 = \begin{bmatrix} 1000 & 10 \\ 10 & 500 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 1000 & 10 \\ 10 & 500 \end{bmatrix}$$

B is an example of homogeneous variance-covariance. The two groups have the same variance and covariance for each variable, Although the variances for each variable are different.

3. (2pts) The following regression model was fit to the movies data, and this conclusion was proffered:

Conclusion: Because the model deviance is so low relative to the null deviance, the users ratings are almost perfectly predicted by a combination of the average critics rating, the movie genre and the MPAA rating.

Do you believe it? Explain your answer.

Although the model deviance is much smaller than the null deviance none of the predictor variables have significant contributions to the model. This suggests it is overspecified.

```
> summary(glm(users.ratings.gpa~d.movies.users[,2]+critics.av+genre+mpaa))
```

Call:

```
glm(formula = users.ratings.gpa ~ d.movies.users[, 2] + critics.av +  
     genre + mpaa)
```

Deviance Residuals:

1	2	3	4	5	6	7
0.377343	-0.040107	0.188356	0.154576	-0.154576	0.000000	-0.006653
8	9	10	11	12	13	14
0.165106	-0.324236	0.000000	0.324236	-0.565698	0.046760	-0.165106

Coefficients:

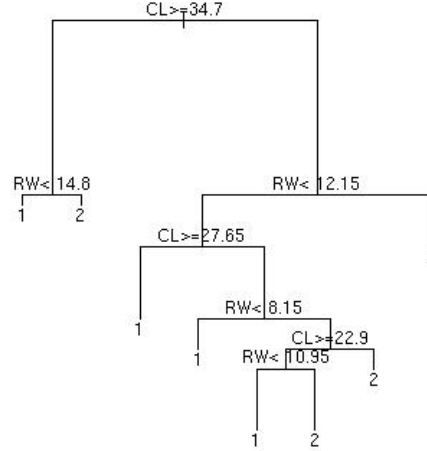
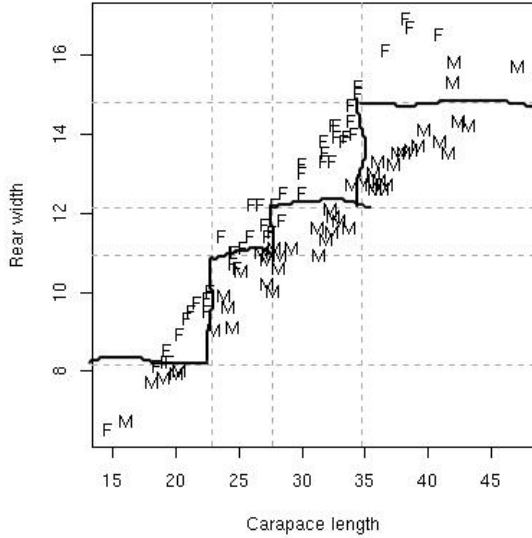
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.583e-01	1.330e+00	0.721	0.503
Number.of.Reviewers	-2.330e-05	2.467e-05	-0.944	0.388
critics.av	5.130e-01	3.326e-01	1.543	0.184
genreChild	1.333e+00	8.450e-01	1.577	0.176
genreComedy	6.408e-01	5.442e-01	1.178	0.292
genreDrama	-3.209e-02	3.921e-01	-0.082	0.938
mpaaPG	-4.319e-01	6.840e-01	-0.631	0.555
mpaaPG-13	6.014e-01	7.283e-01	0.826	0.447
mpaaR	5.355e-01	8.816e-01	0.607	0.570

(Dispersion parameter for gaussian family taken to be 0.1628568)

Null deviance: 3.46825 on 13 degrees of freedom
Residual deviance: 0.81428 on 5 degrees of freedom
AIC: 19.907

Number of Fisher Scoring iterations: 2

4. (2pts) These are results from building a tree for data on crabs. Draw the boundary between the two groups.



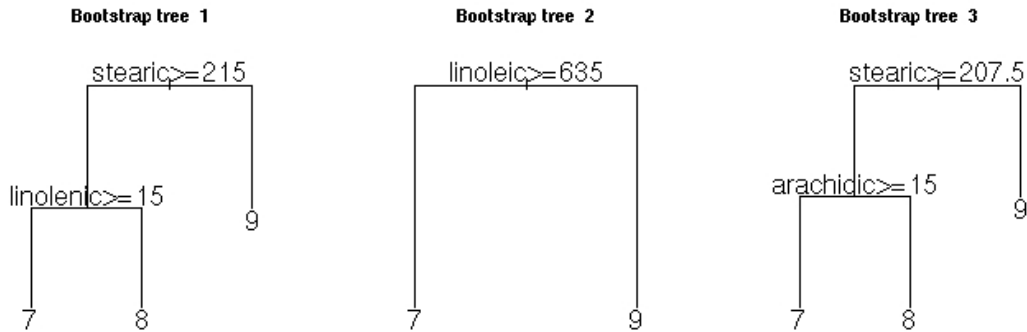
n= 100

node), split, n, loss, yval, (yprob)
* denotes terminal node

- 1) root 100 50 1 (0.50000000 0.50000000)
- 2) $CL \geq 34.7$ 25 4 1 (0.84000000 0.16000000)
- 4) $RW < 14.8$ 18 0 1 (1.00000000 0.00000000) *
- 5) $RW \geq 14.8$ 7 3 2 (0.42857143 0.57142857) *
- 3) $CL < 34.7$ 75 29 2 (0.38666667 0.61333333)
- 6) $RW < 12.15$ 51 23 1 (0.54901961 0.45098039)
- 12) $CL \geq 27.65$ 14 1 1 (0.92857143 0.07142857) *
- 13) $CL < 27.65$ 37 15 2 (0.40540541 0.59459459)
- 26) $RW < 8.15$ 7 2 1 (0.71428571 0.28571429) *
- 27) $RW \geq 8.15$ 30 10 2 (0.33333333 0.66666667)
- 54) $CL \geq 22.9$ 20 10 1 (0.50000000 0.50000000)
- 108) $RW < 10.95$ 10 2 1 (0.80000000 0.20000000) *
- 109) $RW \geq 10.95$ 10 2 2 (0.20000000 0.80000000) *
- 55) $CL < 22.9$ 10 0 2 (0.00000000 1.00000000) *
- 7) $RW \geq 12.15$ 24 1 2 (0.04166667 0.95833333) *

5. (3pts) Three trees are built from bootstrap samples of the northern olive oils data (predictors are palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic, and there are three areas, 7, 8, 9). Predict the class for this case for both trees. If the voting scheme used to combine the predictions is majority rule, which class would it be predicted to be? (Recall that if the split statement is true go to the left node.)

$$\mathbf{x}_o = (1020 \ 100 \ 220 \ 7530 \ 1030 \ 0 \ 0 \ 3)'$$



8,7,8, which would give a majority rule prediction of 8.

6. (2pts) Write down the classification rule for the following output from R.

```
Call:
lda(tmp.dat[, 2:4], tmp.dat[, 1], prior = c(0.5, 0.5))
```

```
Prior probabilities of groups:
  2  3
0.5 0.5
```

```
Group means:
      oleic linoleic arachidic
2 7268.020 1196.5306  73.17347
3 7793.053  727.0331  37.57616
```

```
Coefficients of linear discriminants:
              LD1
oleic      -0.002838103
linoleic   -0.011048486
arachidic  -0.035906792
```

Assign the new observation \mathbf{x}_o to group 3 if

$$(-0.002838103 \quad -0.011048486 \quad -0.035906792)(\mathbf{x}_o - \frac{1}{2} \begin{bmatrix} 7268.020 + 7793.053 \\ 1196.5306 + 727.0331 \\ 73.17347 + 33.987 \end{bmatrix}) > 0 \quad (1)$$

else assign to group 2.

where 1 reduces to $-0.002838103 \times \text{oleic} - 0.011048486 \times \text{linoleic} - 0.035906792 \times \text{arachidic} + 30.6884 > 0$.

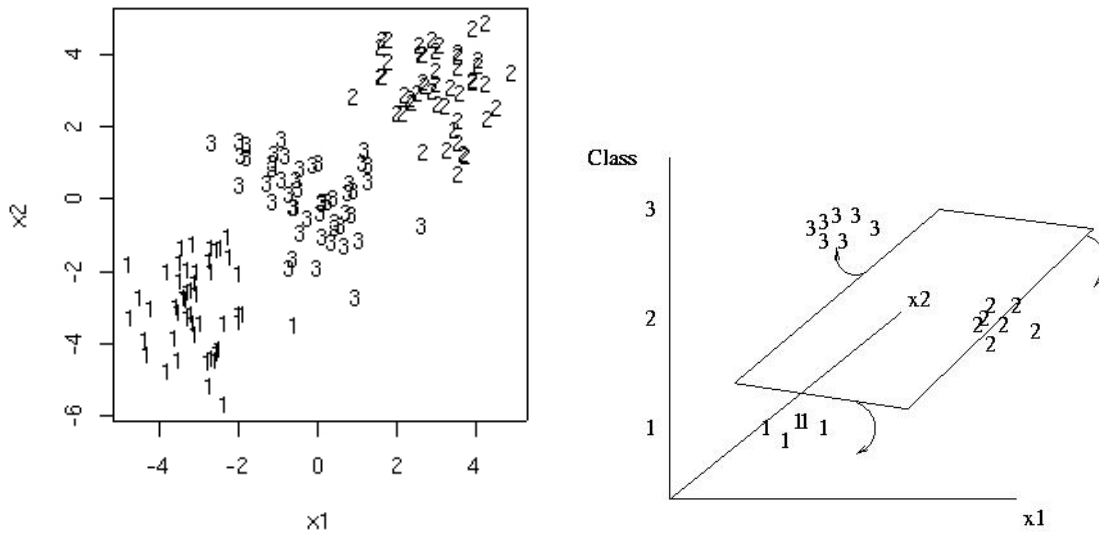
7. (3pts) Read this data description.

Which of the following would you consider to be the major questions to answer using this data? Group the questions into three categories: important (I), less important (L), cannot be answered by this data (N).

A graduate student at ISU used a web crawler to download data on movies stored at www.imdb.com. The information retrieved includes data on 20557 movies that have appeared from 1893 until 2005. The variables collected were budget, length, average user rating, number of users rating, MPAA rating and genre.

- (a) I How have movie budgets changed over time?
- (b) I What are the top 100 movies of the past century according to average user rating?
- (c) L How does average user rating change by genre?
- (d) N Do women rate romance movies higher on average than men?
- (e) L Are documentaries rated on average more highly than action movies?
- (f) I Do the movies that have more users rating them have higher ratings on average?
- (g) I Is the budget related to the average user rating?
- (h) L Are more recent movies more highly rated on average?
- (i) L Do the shorter movies have smaller budgets?
- (j) I Is movie length related to average user rating?
- (k) N Do action movies take longer to produce than documentaries?

8. (3pts) The following plot shows a data set where there are two predictors and three classes. We use a simple regression model to predict the class. What problem will be observed with the result? How does this relate to a feedforward neural network classifier?



A regression model will try to fit a 2D plane to this data, and group 3 will be masked. The class labels are in an awkward order. The plane would need to be bent for it to fit the group 3 points. As a nested regression model, a neural network model will bend the plane to fit the classes, although it will produce some strange artifacts in the prediction regions.