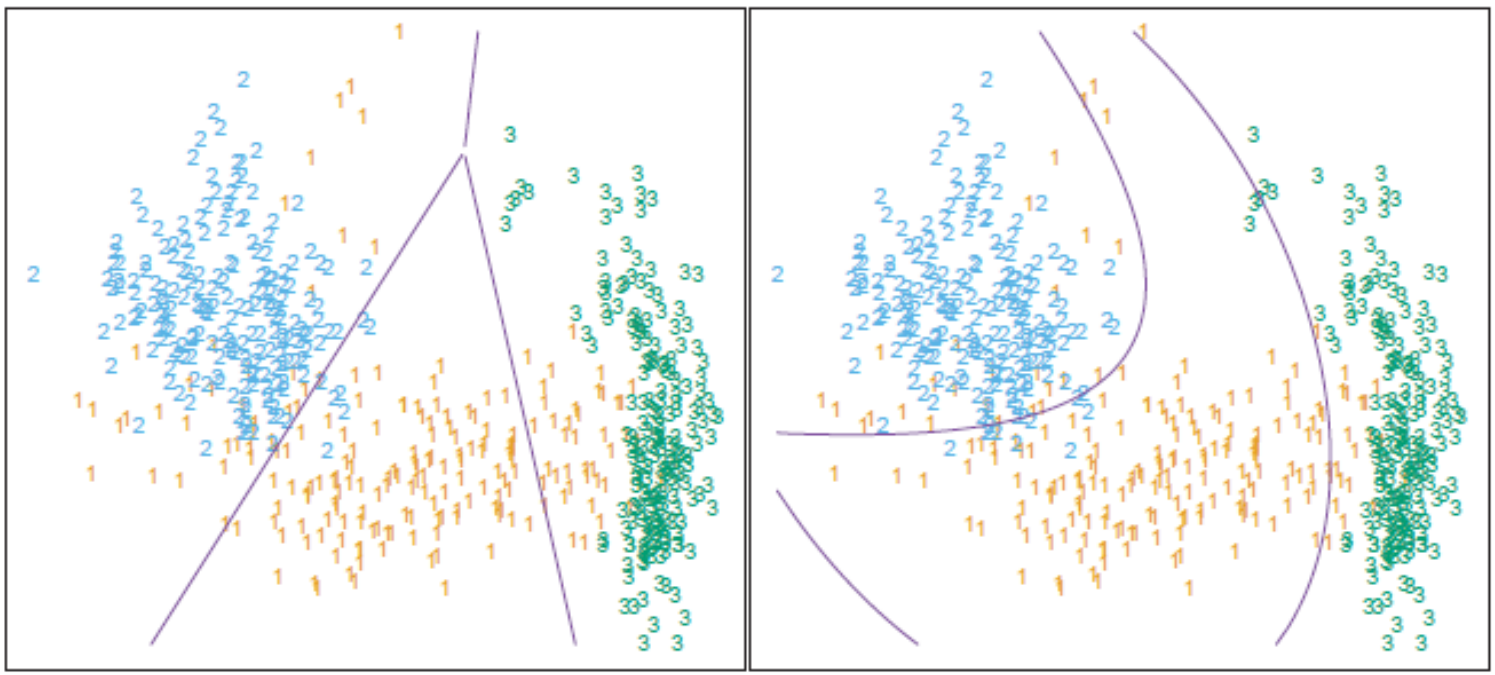
STATS762 Regression for Data Science

Linear methods for classification

May 21, 2019

Classification problem

Predictor Y takes values in a discrete set $\mathcal{K}$ and we divide the input space into a collection of regions labelled according to the classification.



Samples from three distributions and boundaries.

Classification problem

Predictor Y takes values in a discrete set $\mathcal{G}$ and we divide the input space into a collection of regions labelled according to the classification.
Three classification methods in this class;

- Logistic regression/Multinomial logistic regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

Revision - Logistic regression

Given p-predictor variables $x_1, ..., x_p$, the class of response variable y is either 0 or 1 (two classes). The log-odd probability of classes is modelled by

$$\log \frac{P(y = 1|x_1, ..., x_p)}{P(y = 0|x_1, ..., x_p)} = \beta_0 + \beta_1 x_1, + ... + \beta_p x_p \, .$$

This is equivalent to GLM with a Binomial distribution with logit link function.
The probability of $y = 1$ is

$$P(y = 1|x_1, ..., x_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1, + ... + \beta_p x_p)}}$$

The class prediction for $x_1', ..., x_p'$ is 1 with $P(y'|x_1', ..., x_p')$. Otherwise the class is predicted to be 0.

The decision boundary is the set of points satisfying

$$\beta_0 + \beta_1 x_1, + ... + \beta_p x_p = 0 \, .$$

Low birth weight

Birth.csv contains 189 births and low=1 indicates that the birth weight is low.

```
> Birth=read.csv(file='Birthwt.csv',header=TRUE)
> Birth$low=as.factor(Birth$low)
> head(Birth)
         age1         age2         age3       lwt1        lwt2         lwt3 white black smoke
1 -0.05833434  0.011046300  0.029561818  0.12446282 -0.02133871 -0.130731102     0     1     0
2  0.13436561  0.055245529 -0.096907046  0.06006722 -0.06922831 -0.033348413     0     0     0
3 -0.04457005 -0.009415469  0.045088774 -0.05918388  0.03746349  0.004618178     1     0     1
4 -0.03080577 -0.026243567  0.052489640 -0.05202881  0.02390664  0.019034579     1     0     1
5 -0.07209862  0.035141739  0.004821882 -0.05441384  0.02832410  0.014571538     1     0     1
6 -0.03080577 -0.026243567  0.052489640 -0.01386846 -0.03296942  0.049559472     0     0     0
  ptl1 ptl2m ht ui ftv1 ftv2 ftv3m low
1    0     0  0  1    0    0     0   0
2    0     0  0  0    0    0     1   0
3    0     0  0  0    1    0     0   0
4    0     0  0  1    0    1     0   0
5    0     0  0  1    0    0     0   0
6    0     0  0  0    0    0     0   0
```

Low birth weight

```
Birth=read.csv(file='Birthwt.csv',header=TRUE)
Birth$low=as.factor(Birth$low)
Birth.logistic <- multinom(low ~ ., data = Birth)
coef(Birth.logistic)
```

```
(Intercept)        age1        age2        age3        lwt1
-1.6335668 -13.2366667 -21.7955955 -16.2522595  -7.0484965
lwt2        lwt3       white       black       smoke
-2.4012762  -4.7102126  -0.7322719   0.5423045   0.8732237
ptl1       ptl2m          ht          ui        ftv1
1.6780989  -0.3168657   2.1082575   0.8099013  -0.3932662
ftv2       ftv3m
-0.1630313   0.7328330
```

Low birth weight

```
bp <- predict(Birth.logistic,newdata=Birth)
table(bp,Birth$low)
```

The confusion matrix is

	y=0	y=1
y'=0	114	32
y'=1	16	27

Total 141 points are correctly predicted and 48 are incorrectly predicted.
i.e., 74.6% of datapoints are corrected predicted.
i.e. 16 of 0's are incorrectly predicted by 1.

Iris data

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa

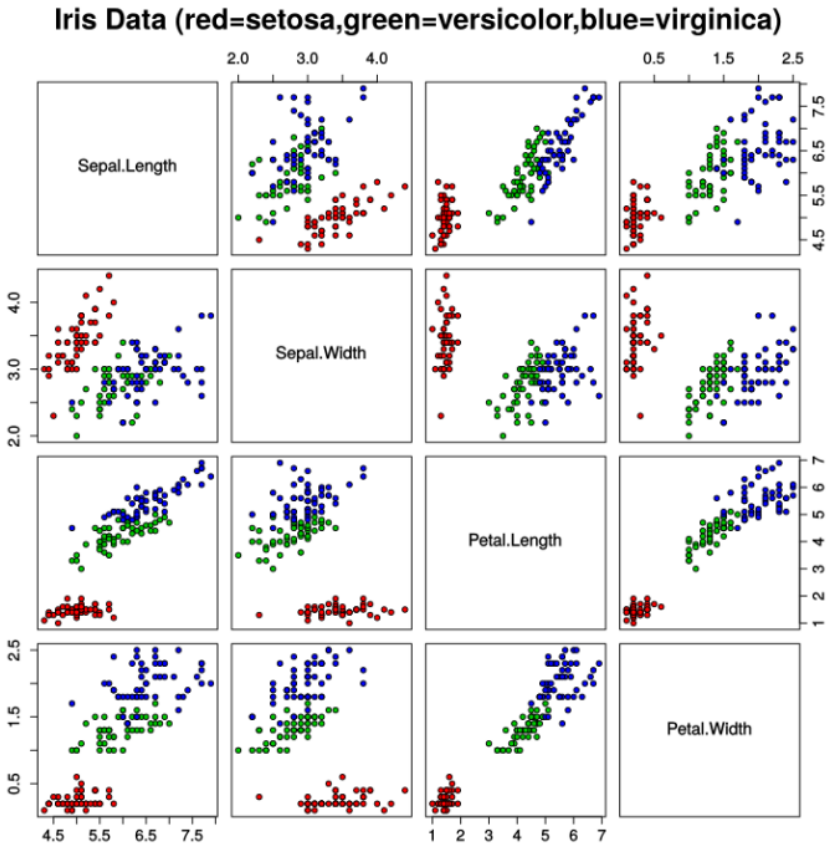
Iris Versicolor

Iris Setosa

Iris Virginica

Iris data

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.



Can we predict the species given petal and sepal information? But there more than 2 outcomes!

Multinomial logistic regression

When the response variable Y has more than two possible outcomes, the multinomial logistic regression is used to predict the probability of the different possible outcomes of a categorically distributed dependent variable.

The idea is that logistic regressions are modelled to log-odd ratios with respect to the reference class.

There is no probability distribution assumption for X. Distribution free!

Multinomial logistic regression

There are K possible outcomes (K classes) and the K-class is the reference class.
Let's represent $K - 1$ log-odds using the reference outcome.

$$\log \frac{Pr(Y = 1|X)}{Pr(Y = K|X)} = \beta_{1,0} + \beta_1^T X$$

$$\log \frac{Pr(Y = 2|X)}{Pr(Y = K|X)} = \beta_{2,0} + \beta_2^T X$$

$$\vdots$$

$$\log \frac{Pr(Y = K - 1|X)}{Pr(Y = K|X)} = \beta_{K-1,0} + \beta_{K-1}^T X$$

Note that the reference class (common denominator) can be any class.

Multinomial logistic regression

A probability for a class k, $k = 1, \ldots, K-1$ is

$$Pr(Y = k|X) = \frac{\exp(\beta_{k,0} + \beta_k^T y)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l,0} + \beta_l^T y)}$$

and a probability for a class K is

$$Pr(Y = K|X) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\beta_{k,0} + \beta_k^T X)}.$$

Note that $\sum_{l=1}^{K} Pr(Y = l|X) = 1$.

Unknown parameters $\beta_{1,0}, \beta_1, \ldots, \beta_{K-1,0}, \beta_{K-1}$ are estimated using the Maximum likelihood estimation.

Implementation in R

Multinomial logistic regression implementation in R

```
library(nnet)
multinom(formula, data, subset)
```

formula Formula expression as for regression models, of the form response predictors.

data Data of interest

subset An index vector specifying the cases to be used in the training sample.

The function also computes the residual deviance (-2log.likelihood) and AIC.

Usual functions such as coef and predict can be used to extract information.

Iris data

Multinomial logistic model is trained using 120 samples (train data) and the model will be validated for the remaining data (test data). log-odd ratios for the train data are modelled by

$$\begin{aligned}
\log \frac{P(Y=\textit{versicolor})}{P(Y=\textit{setosa})} &= 77.02 - 22.94\textit{Sepal.Length} - 53.70\textit{Sepal.Width} \\
&\quad + 101.42\textit{Petal.Length} + 26.34\textit{Petal.Width} \\
\log \frac{P(Y=\textit{virginica})}{P(Y=\textit{setosa})} &= -153.95 - 59.50\textit{Sepal.Length} - 96.15\textit{Sepal.Width} \\
&\quad + 191.89\textit{Petal.Length} + 107.42\textit{Petal.Width}
\end{aligned}$$

i.e., A one-unit increases in Sepal.Length results a decreases in the log odds of *versicolor* vs *setosa* at the rate of 22.94.

i.e., A one-unit increases in Petal.Length results an increase in the log odds of *virginica* vs *setosa* at the rate of 191.89.

Iris data

Species probability ratios are modelled by

$$\begin{aligned}
\frac{P(Y=versicolor)}{P(Y=setosa)} &= \exp(77.02 - 22.94 Sepal.Length - 53.70 Sepal.Width \\
&\quad + 101.42 Petal.Length + 26.34 Petal.Width) \\
\frac{P(Y=virginica)}{P(Y=setosa)} &= \exp(-153.95 - 59.50 Sepal.Length - 96.15 Sepal.Width \\
&\quad + 191.89 Petal.Length + 107.42 Petal.Width)
\end{aligned}$$

i.e., Species probability ratio for *versicolor* over *setosa* decreases by the power of 22.94 with one-unit increase with Sepal.Length.

i.e., Species probability ratio for *virginica* over *setosa* increases by the power of 191.89 with one-unit increase with Petal.Length.

Iris data

Class probability estimate follows;

Obs	Predict	P(versicolor)	P(virginica)	P(setosa)
setosa	setosa	3.500292e-21	5.347927e-194	1.000000e+00
⋮	⋮	⋮	⋮	
versicolor	versicolor	1.000000e+00	3.129722e-37	7.794636e-113
⋮	⋮	⋮	⋮	
virginica	virginica	2.698059e-82	1.000000e+00	1.275528e-308
⋮	⋮	⋮	⋮	

These are three samples out of 150.

Iris data

Species are predicted for the test data (30 samples) according to the class probability estimate and the confusion matrix is

	Obs=setosa	Obs=versicolor	Obs=virginica
Predict=setosa	10	0	0
Predict=versicolor	0	10	0
Predict=virginica	0	1	9

- 1 out of 30 iris is incorrectly predicted.
- 96.67% of 30 samples in the test data are correctly predicted.

Note: A confusion matrix is a table that is often used to describe the performance of a classification model.

Multinomial logistic regression

- Adding or deleting alternative outcome categories does not affect the odds among the remaining outcomes.
- Diagnostics and model fit
- Multinomial regression uses the MLE to find multiple equations and it requires a large sample size.
- Distribution free classification.

Distribution based classification

Let $f_k(x)$ is the class-conditional density of $Y = k$ and let π_k be the prior probability of class k with $\sum_{k=1}^{K} \pi_k = 1$.
A simple Bayes rule gives

$$Pr(Y = k|X) = \frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

Based on this, many techniques are proposed.

- linear and quadratic discriminant analysis using Gaussian densities
- mixture model (distribution based clustering)
- nonparametric density estimate for each class density f_k
- Naive Bayes

LDA/QDA

For both LDA and QDA, it is assumed that the class-conditional density $P(Y = k|X)$ is a Gaussian distribution.

Each class, X is modelled by a class-specific Gaussian distribution.

Classes are identified by class-specific distributions.

Linear Discriminant Analysis (LDA)

Suppose that f_k be a multivariate Gaussian

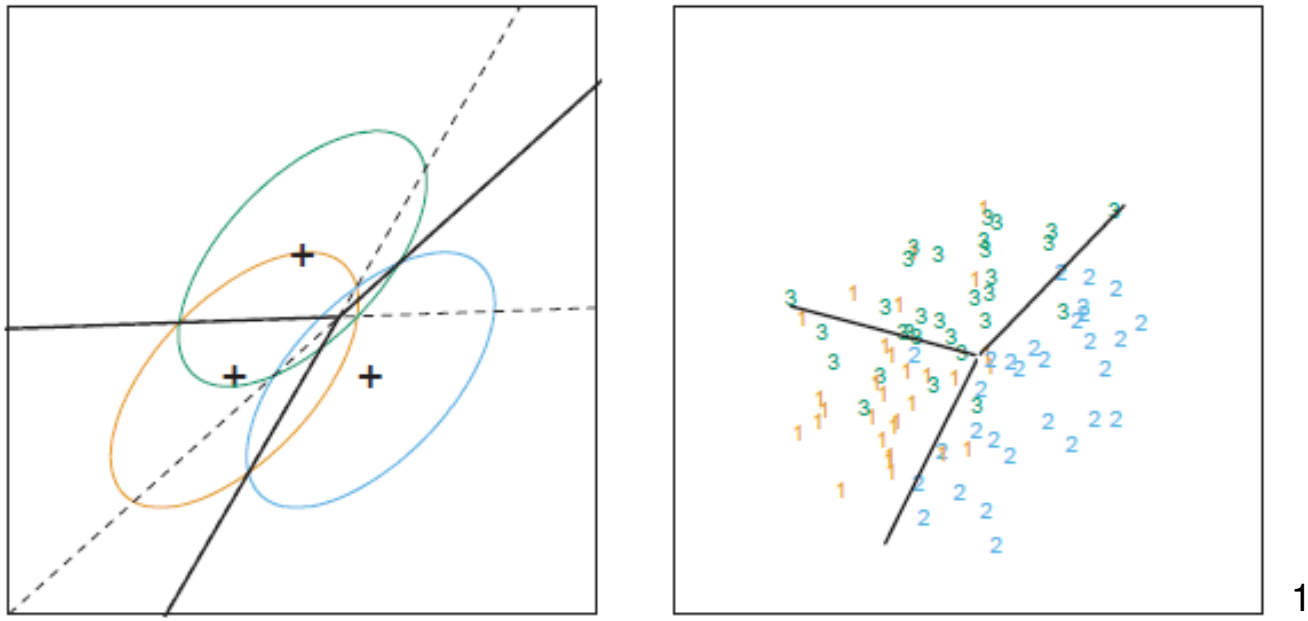
$$f_k = \frac{\exp(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))}{(2\pi)^{p/2}|\Sigma_k|^{1/2}}.$$

The LDA is a special case that all classes have a common covariance matrix $\Sigma = \Sigma_1 = \cdots = \Sigma_K$.
The comparison between two class k and j is

$$\begin{aligned}\log\frac{Pr(Y=k|X)}{Pr(Y=j|X)} &= \log\frac{\pi_k}{\pi_j} + \log\frac{f_k(x)}{f_j(x)} \\ &= \log\frac{\pi_k}{\pi_j} - \frac{1}{2}(\mu_k+\mu_j)^T\Sigma^{-1}(\mu_k-\mu_j) + y^T\Sigma^{-1}(\mu_k-\mu_j).\end{aligned}$$

- Implies the decision boundary between k and j.
- Divides $\mathbb{R}^p$ into K-hyperplanes to determine K-classes.

Linear Discriminant Analysis (LDA)



Three Gaussian distributions with the common covariance matrix and different means - (Left) Contours of 95% of the probability and (Right) 30 class samples. The black lines represent the Bayes decision boundaries.

[1] The Elements of Statistical Learning. Hastie, T., Tibshirani, R., Friedman, J. Spring Series in Statistics

Linear Discriminant Analysis (LDA)

The *linear discriminant function* is

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

and the boundary decision is a linear function of x. The class is predicted by the k-value maximizing $\delta_k(x)$; $y = \mathrm{argmax}_k \delta_k(x)$.

Key parameters are estimated using the train data;

- $\hat{\pi}_k = n_k / n$, a proportion of sample from the k-class.
- $\hat{\mu}_k = \sum_{y_i = k} x_i / n_k$, a sample mean
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{y_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (n - K)$, a sample covariance matrix

Linear Discriminant Analysis (LDA)

If the class k is favoured over j for X,

$$\log \frac{Pr(k|X)}{Pr(j|X)} = \log \frac{\pi_k}{\pi_j} - \frac{1}{2}(\mu_k + \mu_j)^T \Sigma^{-1}(\mu_k - \mu_j) + x^T \Sigma^{-1}(\mu_k - \mu_j) > 0$$

and this is equivalent to

$$x^T \Sigma^{-1}(\mu_k - \mu_j) > -\log \frac{\pi_k}{\pi_j} + \frac{1}{2}(\mu_k + \mu_j)^T \Sigma^{-1}(\mu_k - \mu_j)\,.$$

The left function can be seen as a linear transformation of Y with the line direction $v = \Sigma^{-1}(\mu_k - \mu_j)$.
The boundary between k and j groups is points satisfying

$$\log \frac{\pi_k}{\pi_j} + \frac{1}{2}(\mu_k + \mu_j)^T \Sigma^{-1}(\mu_k - \mu_j) + x^T \Sigma^{-1}(\mu_k - \mu_j) = 0\,.$$

LDA implementation in R

```
library(MASS)
lda(formula, data, prior, subset)
```

formula Formula expression as for regression models.

data Data of interest.

subset An index vector specifying the cases to be used in the training sample.

prior Prior probability of class membership. If unspecified, the class proportions for the training set are used.

Iris data

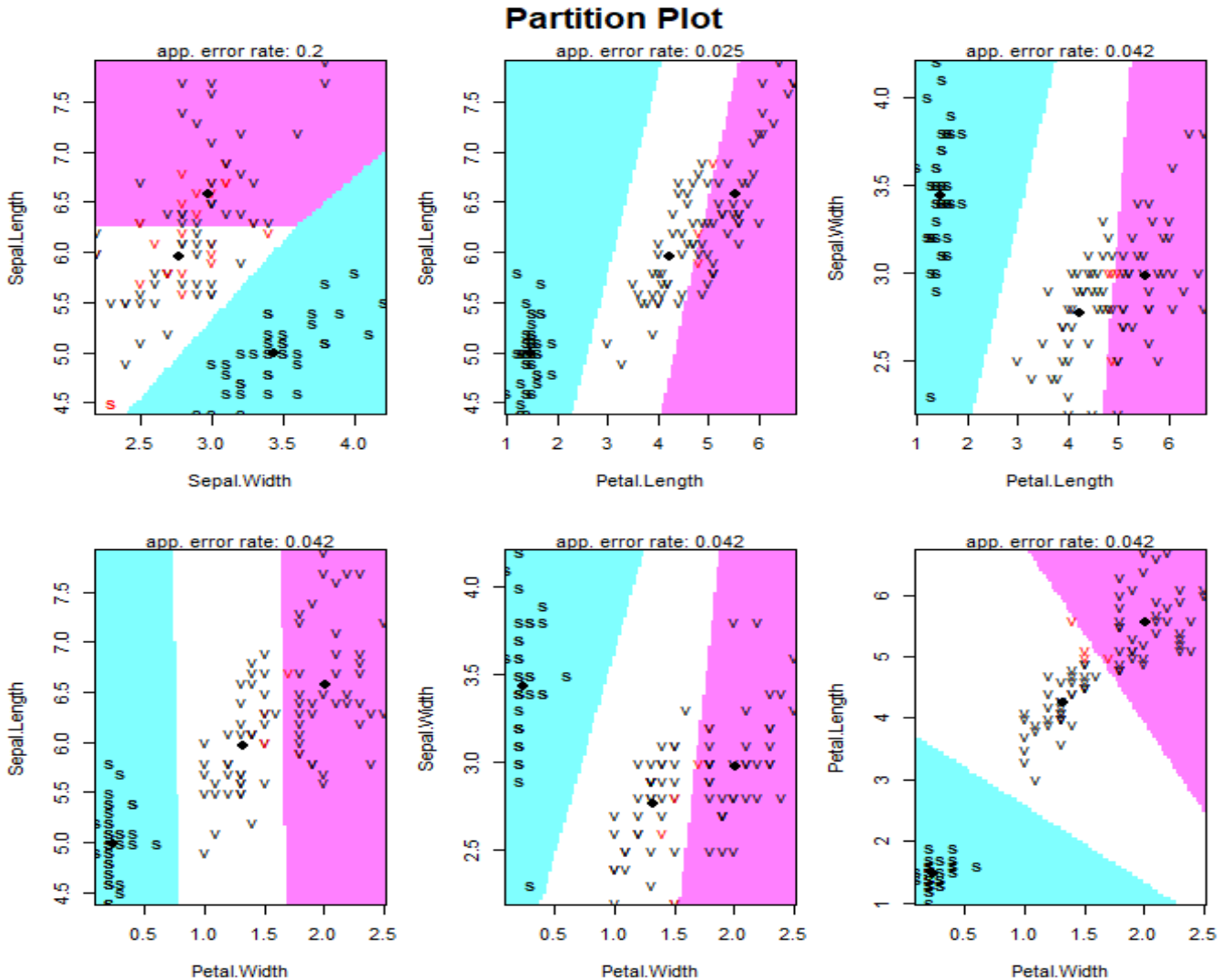
Some outputs of lda-simulation;

```
Prior probabilities of groups:
setosa versicolor  virginica
0.3333333  0.3250000  0.3416667

Group means: Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa          4.992500    3.437500     1.470000    0.240000
versicolor      5.961538    2.766667     4.246154    1.315385
virginica       6.578049    2.978049     5.558537    2.014634
```

- Prior probabilities are proportions of classes in the training data.
- The separation between group centers are shown.

Iris data

Boundaries of the three groups for all pairs of covariates.



Iris data

Species prediction; y, species observation and y', species prediction

	y=setosa	y=versicolor	y=virginica
y'=setosa	8	0	0
y'=versicolor	2	9	2
y'=virginica	0	2	7

- 6 samples in the test data are incorrectly predicted.
- 80% of samples in the test data are correctly predicted.

The equal covariance matrix (equal variation within classes) assumption seems to be too strict and is not very useful for this classification problem. Perhaps relaxing this assumption may be helpful. QDA!

Quadratic Discriminant Analysis (QDA)

- No assumption of a common covariance matrix. The covariance of each of the classes is different.
- The quadratic discriminant function is

$$\delta_k(x) = \log \pi_k - \frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k|\,.$$

- The class is predicted by maximizing $\delta_k(x)$; $y = \text{argmax}_k \delta_k(x)$.
  i.e., If the class k is favoured over j for X

$$\log\frac{Pr(k|X)}{Pr(j|X)} = \log\frac{\pi_k}{\pi_j} + \log\frac{\sqrt{|\Sigma_j|}}{\sqrt{|\Sigma_k|}} + \log\frac{e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}}{e^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)}} > 0$$

- The class boundary decision becomes a quadratic function of y.

Iris data

Some outputs of qda-simulation; Same train data information.

```
Prior probabilities of groups:
setosa versicolor  virginica
0.3333333  0.3250000  0.3416667

Group means: Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa          4.992500    3.437500     1.470000    0.240000
versicolor      5.961538    2.766667     4.246154    1.315385
virginica       6.578049    2.978049     5.558537    2.014634
```

- Prior probabilities are proportions of classes in the training data.
- The separation between group centers are shown.

Iris data

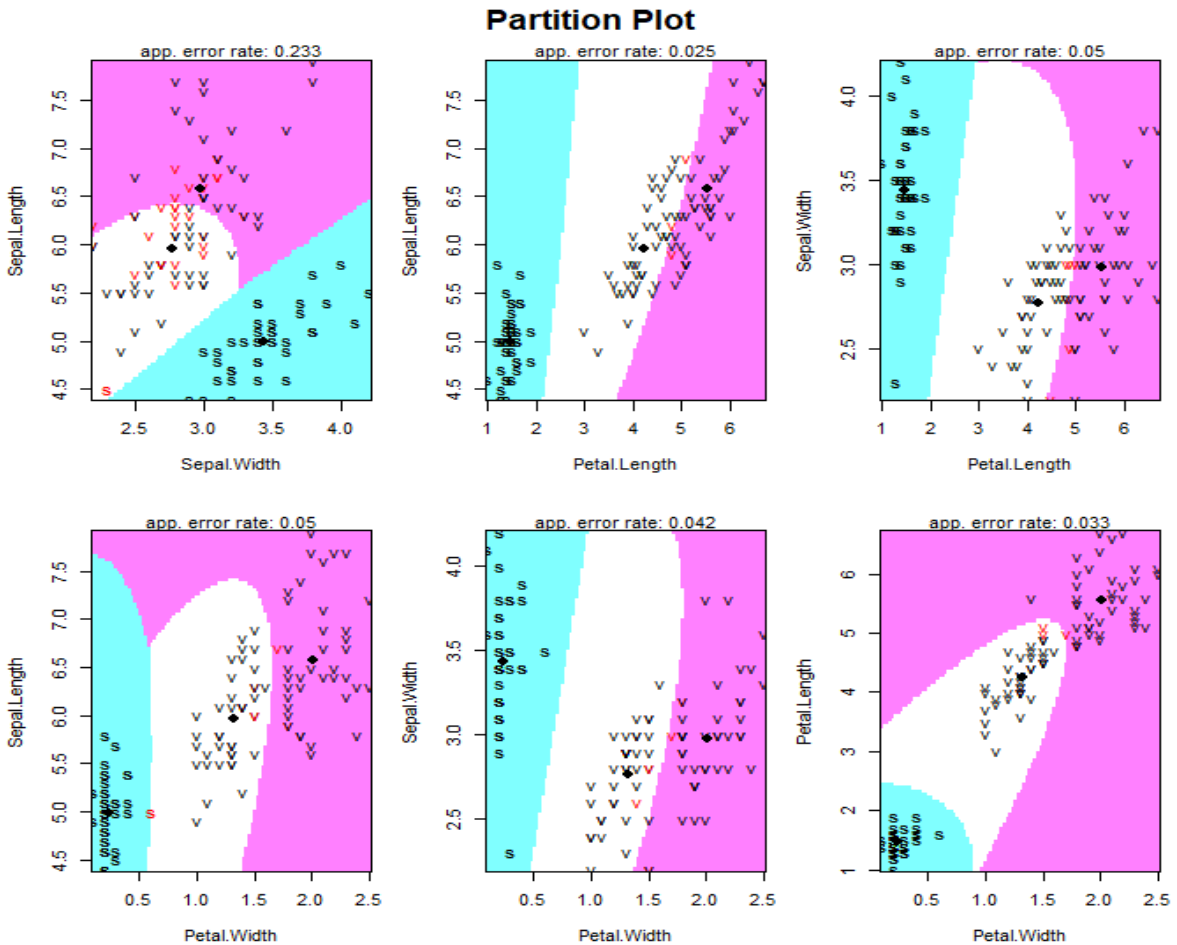
Some outputs of qda-simulation; Scaling matrix S

```
setosa             1         2           3           4
Sepal.Length -2.91563  3.054115  0.72524940   0.5127399
Sepal.Width   0.00000 -4.071636  0.09833735  -0.4011328
Petal.Length  0.00000  0.000000 -5.90284433   1.7818261
Petal.Width   0.00000  0.000000  0.00000000 -11.4199898

versicolor         1         2           3           4
Sepal.Length 2.017889  1.168536 -2.303647 -0.5926211
Sepal.Width  0.000000 -4.178214 -1.211024  1.8475690
Petal.Length 0.000000  0.000000  3.610818  2.6767707
Petal.Width  0.000000  0.000000  0.000000 -8.6430163

virginica          1          2           3           4
Sepal.Length -1.628103 -0.9163113 -2.6025135  0.3057130
Sepal.Width   0.000000  3.4480101 -0.4018851 -2.0800743
Petal.Length  0.000000  0.0000000  3.6658642 -0.2948522
Petal.Width   0.000000  0.0000000  0.0000000  4.5469465
```

Covariance is obtained by $(S^T S)^{-1}$.

Iris data

Boundaries of the three species for all pairs of covariates;



Iris data

Species membership probabilities for test data;

	prediction	setosa	versicolor	virginica
10	1	1.000000e+00	1.939852e-20	3.074246e-42
13	1	1.000000e+00	1.156013e-19	7.636634e-42
14	1	1.000000e+00	8.221906e-20	5.391686e-42
16	1	1.000000e+00	1.073771e-40	1.572583e-63
17	1	1.000000e+00	2.634520e-32	4.772562e-56
22	1	1.000000e+00	9.096346e-26	2.522076e-47
24	1	1.000000e+00	2.449560e-16	3.147541e-36
26	1	1.000000e+00	2.388778e-17	1.058343e-38
37	1	1.000000e+00	3.365059e-29	7.275027e-55
46	1	1.000000e+00	1.036567e-17	3.051268e-39
51	2	1.873434e-93	9.999888e-01	1.117047e-05
52	2	1.063951e-85	9.999125e-01	8.746071e-05
56	2	2.547396e-79	9.952434e-01	4.756580e-03
61	2	3.532178e-42	9.999349e-01	6.510487e-05
62	2	5.872831e-75	9.996815e-01	3.184808e-04
84	3	1.448678e-117	8.278758e-02	9.172124e-01
85	2	1.765355e-84	9.502255e-01	4.977445e-02
86	2	9.249440e-87	9.984985e-01	1.501541e-03
88	2	3.475780e-84	9.986637e-01	1.336314e-03
94	2	4.644362e-35	9.999995e-01	5.488211e-07
95	2	3.405821e-69	9.995440e-01	4.559744e-04
107	3	1.910310e-97	8.392983e-04	9.991607e-01
111	3	2.217926e-135	1.141574e-02	9.885843e-01
113	3	8.516511e-164	1.614888e-04	9.998385e-01
119	3	9.516144e-269	1.134035e-09	1.000000e+00
121	3	1.457729e-185	1.879731e-06	9.999981e-01
124	3	2.051255e-119	4.590726e-02	9.540927e-01
130	3	9.112623e-158	2.328233e-02	9.767177e-01
141	3	2.337582e-188	1.006037e-08	1.000000e+00
145	3	2.033599e-199	1.393443e-09	1.000000e+00

Iris data

Species prediction for the test data (30 points); y, species observation and y' species prediction

	y=setosa	y=versicolor	y=virginica
y'=setosa	10	0	0
y'=versicolor	0	10	0
y'=virginica	0	1	9

- Only one sample is incorrectly predicted.
- 96.7% of samples in the test data are correctly predicted.

By nature, species-specific variations are different and the flexibility of QDA yields a better classification model for the Iris data.

LDA and QDA

x is classified to Class k over Class j when $\delta_k(x) > \delta_j(x)$ and the boundary between them is found by x's that $\delta_k(x) = \delta_j(x)$.

- Linear discriminant function :

$$\delta_k(x) = y^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

- Quadratic discriminant function :

$$\delta_k(x) = \log \pi_k - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\Sigma_k|$$

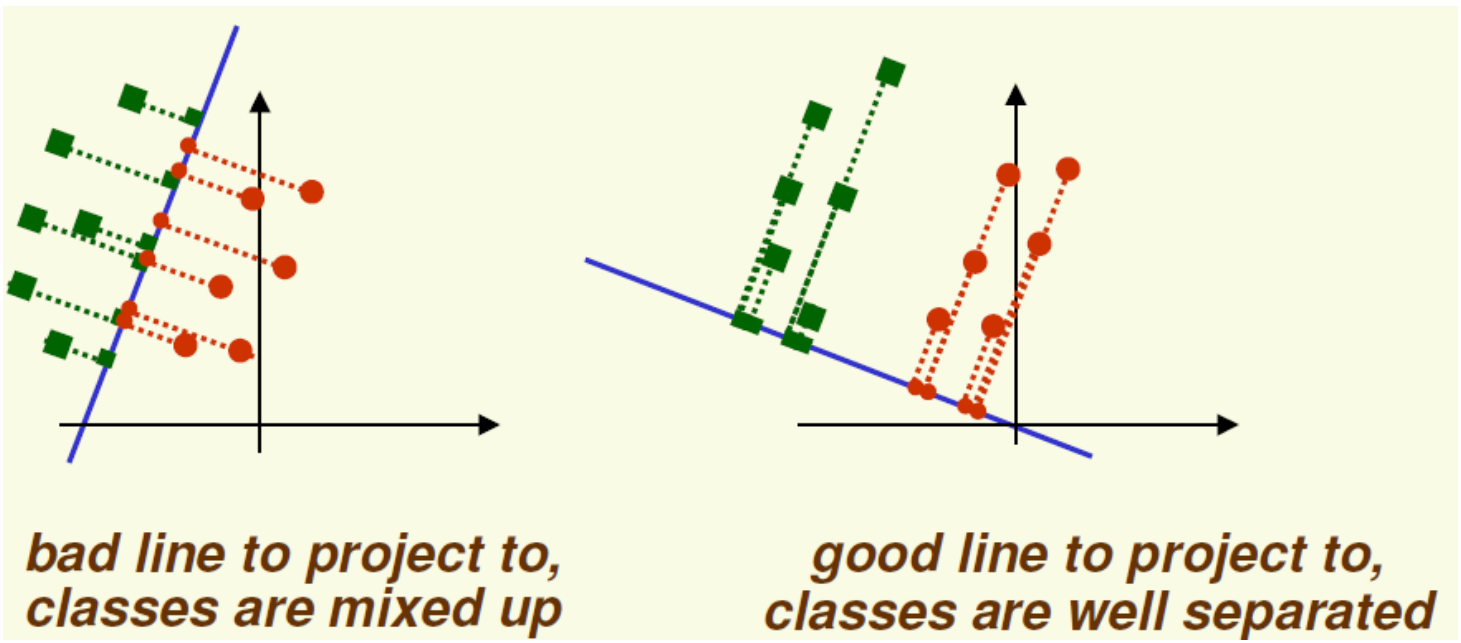
LDA and QDA

- Gaussian model assumption
- Similar except the covariance assumption;
  Common covariance (LDA) and different class-specific covariance (QDA)
- Simple decision boundaries;
  linear (LDA) and quadratic equation (QDA)
- Stable estimate via the Gaussian models

Discriminant function

In classical statistics books, the discriminant means linear combinations of predictors.

Discriminant Function Analysis finds the linear combinations of variables that maximize the separation (discrimination) of groups.



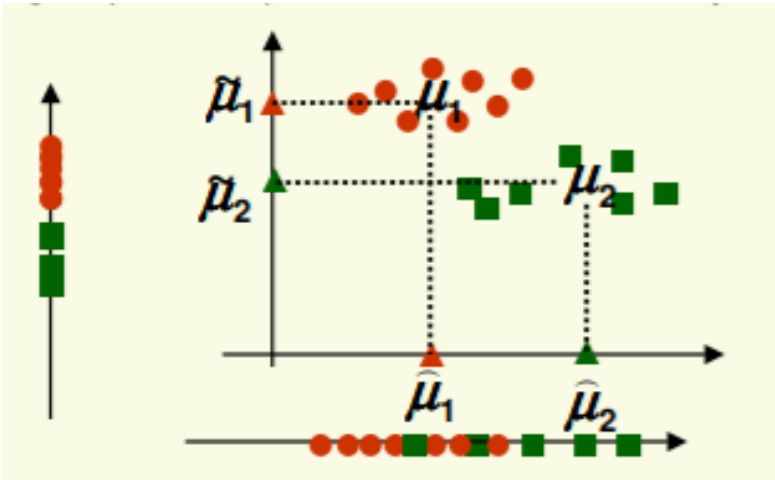
Discriminant function

- The line direction by unit vector is v. Scalar $v^T x$ is the projection of x into a one dimensional subspace.
- For K-class problems, there are $K - 1$ number of linear discriminants.
- $K - 1$ linear discriminants are ordered by variance explanation.

Discriminant function

We want projected points are very well separated between classes.

i.e., Big separation between classes and small variation within classes.



The vertical axes is a better line than the horizontal line for v to separate classes.

We want the line direction v such that

$$|\mu_1 - \mu_2| \ll |v^T \mu_1 - v^T \mu_2|$$

Discriminant function

Class separation should be normalize by taking an account of variance.

Fisher linear discriminant is to project on line in the direction v maximizing

$$J(v) = \frac{\text{varation between classes}}{\text{variation within classes}} = \frac{v^T S_B v}{v^T S_W v}$$

Here

$$S_B = \sum_{k=1}^{K} (\mu_k - \bar{x})(\mu_k - \bar{x})^T$$

where $\bar{x}$ is an overall mean and

$$S_W = \sum_{k=1}^{K} \sum_{x_i \in Class\ k} (x_i - \mu_k)(x_i - \mu_k)^T$$

Fisher's linear discriminant

There are two classes ($K = 2$) [2].

We often use $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and the objective function becomes

$$J(v) = \frac{v^T S_B v}{v^T S_W v} = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}.$$

where $\tilde{\mu}_1 = v^T \mu_1$ and $\tilde{S}_1^2 = \sum_{v^T x_i \in Class1} (v^T x_i - v^T \mu_1)^2$.

$J(v)$ is maximized when $v = (S_1 + S_2)^{-1}(\mu_1 - \mu_2)$. Here S_1 is a covariance matrix for Class 1; $S_1 = \sum_{x_i \in Class1} (x_i - \mu_1)(x_i - \mu_1)^T$.

i.e., This is equivalent to QDA for $K = 2$ problems.

i.e., When $S_1 = S_2 = S$, this is equivalent to LDA for $K = 2$ problems.

---

[2] Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics. 7 (2): 179188

Example

10 samples from two classes;

class	x1	x2
1	-0.24	-0.11
1	-0.75	-0.09
1	-1.32	-1.67
1	1.05	1.16
1	0.90	0.61
2	1.28	1.12
2	2.12	2.28
2	0.79	2.20
2	1.94	1.86
2	1.37	2.11

Class centers are $\mu_1 = [-0.072, -0.020]$ and $\mu_2 = [1.500, 1.914]$.

$$S_1 = \begin{bmatrix} 4.24908 & 4.0581 \\ 4.05810 & 4.5248 \end{bmatrix} \; S_2 = \begin{bmatrix} 4.85108 & 1.90346 \\ 1.90346 & 4.05884 \end{bmatrix}$$

The optimal linear direction is $v = [-0.68, -0.73]$.

Iris data

For three species, there are two normalized linear discriminant (LD1 and LD2). Coefficients of linear discriminants follow;

```
Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.8354281  0.4352594
Sepal.Width   1.6758932 -2.6651653
Petal.Length -2.3351504  0.5216537
Petal.Width  -2.8235954 -2.4495497

Proportion of trace:  LD1    LD2
                    0.9909 0.0091
```

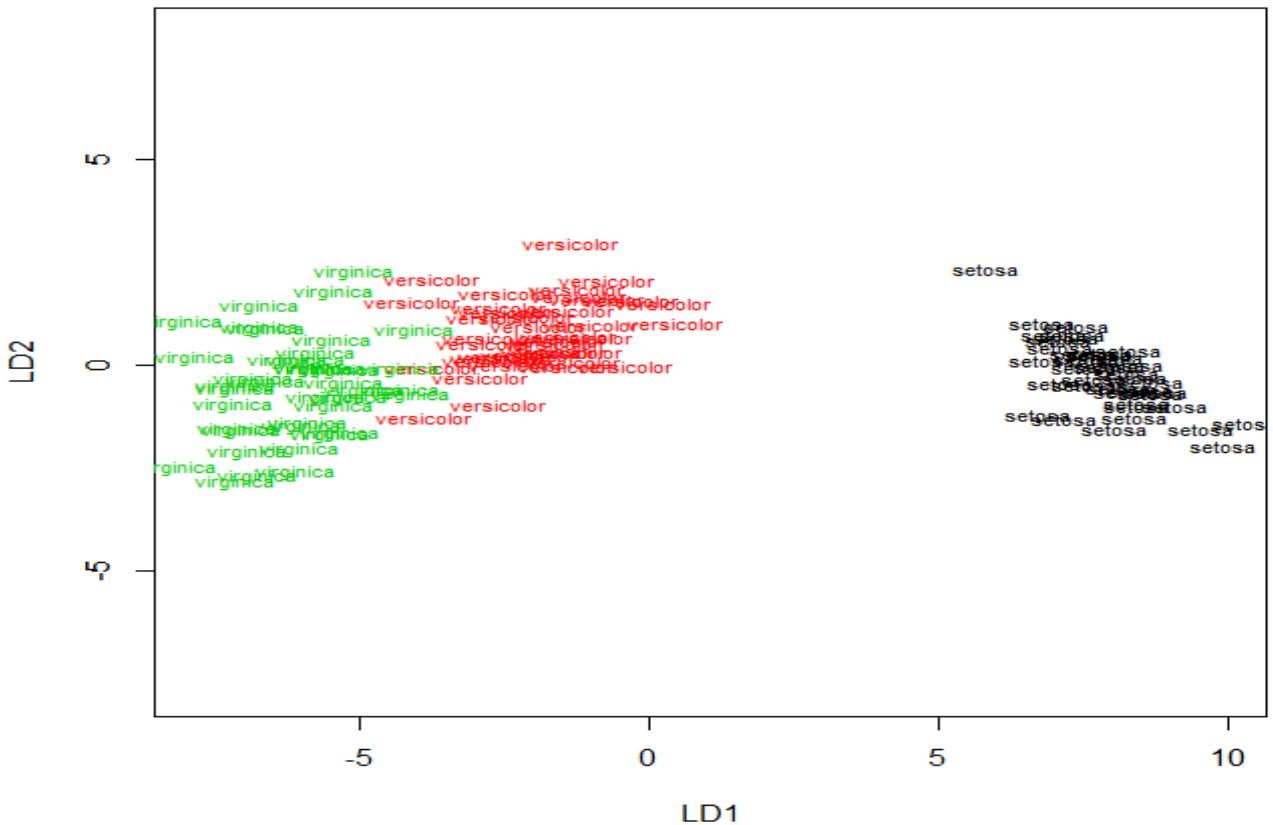
For example the first Fisher linear discriminant (LD1) is

$$0.84*\textbf{Sepal.Length}+1.68*\textbf{Sepal.Width}-2.33*\textbf{Petal.Length}-2.82*\textbf{Petal.Width}$$

LD1 explains more than 99% of the between-group variance in the iris dataset. All four variables are useful to identify species.

Iris data

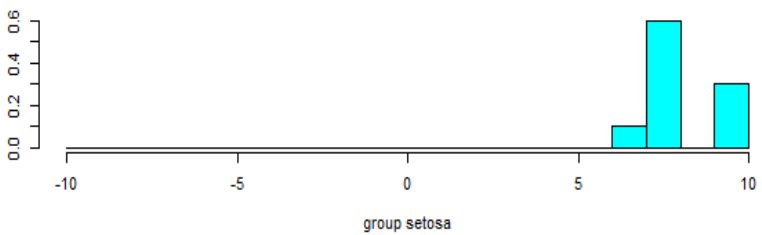
Classification on the normalized linear discriminant functions.

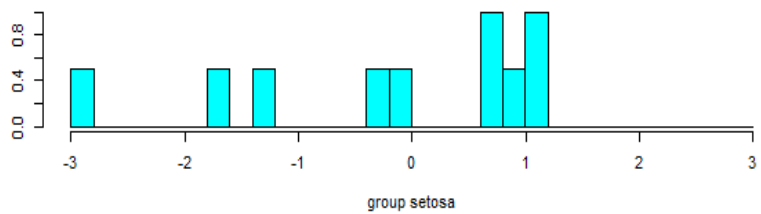


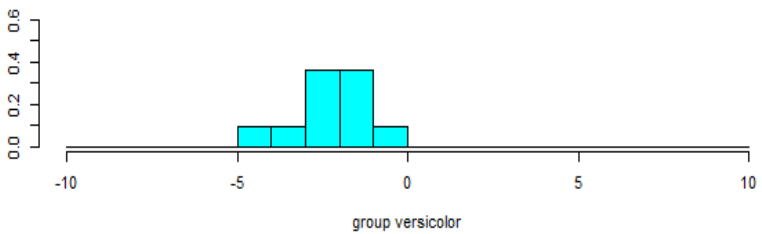
UD1 projects points from different groups further away - easier to separate projected points from different groups.

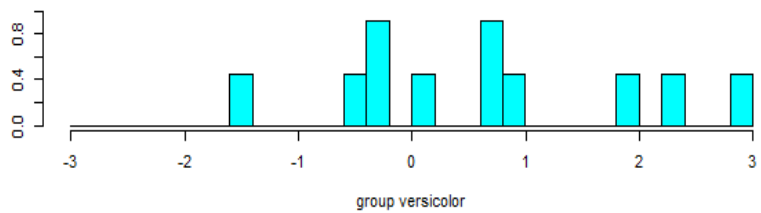
Iris data

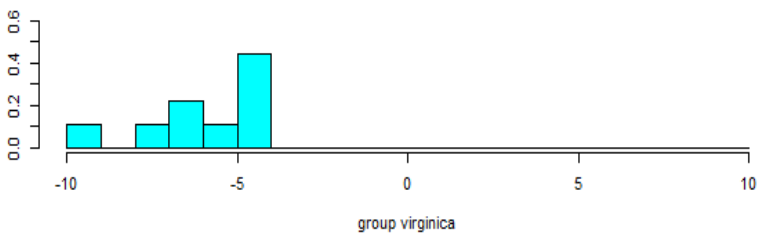
Projected points of 30 samples using LD1 (Left) and LD2 (Right).

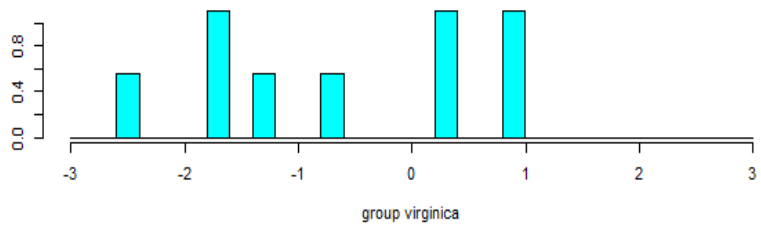












Projected points using LD1 are better separated between species.

LDA/QDA or Multinomial logistic regression

For the LDA, the log-posterior odds between class k and K is a linear function of x

$$\begin{aligned}\log \frac{Pr(k|X)}{Pr(K|X)} &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k + \mu_K)^T \Sigma^{-1}(\mu_k + \mu_K) + x^T \Sigma^{-1}(\mu_k + \mu_K) \\ &= \alpha_0 + \alpha_k^T x\end{aligned}$$

The linear logistic model is

$$\log \frac{Pr(k|X)}{Pr(K|X)} = \beta_{0,k} + \beta_k^T x$$

They have the same form and the difference is in the way the linear coefficients are estimated.

For both, the class probability for $k = 1, \ldots, K-1$ has the log-linear form

$$Pr(k|X) = \frac{e^{\beta_{o,k} + \beta_k^T x}}{1 + \sum_{j=1}^{K-1} e^{\beta_{o,j} + \beta_j^T x}}$$

For the logistic regression model, $Pr(k|X)$ is an arbitrary distribution and it is a Gaussian distribution for the LDA/QDA.

LDA/QDA or Multinomial logistic regression

For the QDA, the log-posterior odds between class k and K is a quadratic function of x

$$\begin{aligned}\log \frac{Pr(Y=k|X)}{Pr(Y=K|X)} &= \log \frac{\pi_k}{\pi_K} + \log \frac{\sqrt{|\Sigma_K|}}{\sqrt{|\Sigma_k|}} + \log \frac{e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}}{e^{-\frac{1}{2}(x-\mu_K)^T \Sigma_j^{-1}(x-\mu_K)}} \\ &= \gamma_0 + \gamma_1^T x + x^T \gamma_2 x\end{aligned}$$

This allows a flexible boundary form of log-odd when it needs.

For a logistic regression model, technical challenges have been reported when a log-odd is unstable. i.e., a denominator $Pr(Y = K|X)$ is very small.

No strong preference between LDA/QDA and multinomial logistic regression. It rather depends how you want to model the log-odd.

LDA/QDA or Multinomial logistic regression

The class prediction for LDA/QDA combines the prior knowledge on class probability and relative model fit to class-specific densities f's

$$Pr(Y = k|X) = \frac{\pi(k)f(X|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi(j)f(X|\mu_j, \Sigma_j)}.$$

i.e., $\Sigma_1 = ... = \Sigma_K$ when LDA.

If you have useful prior knowledge about class probability, you can add this information to predict class using LDA/QDA.

Revisit - Low birth weight

Confusion matrices comparison; Multinomial logistic regression (Left), LDA (Middle) and QDA (Right)

ML	y=0	y=1
y'=0	114	32
y'=1	16	27

LDA	y=0	y=1
y'=0	100	21
y'=1	30	38

QDA	y=0	y=1
y'=0	63	7
y'=1	67	52

- The overall prediction rate (both 0 and 1 outcomes) using the multinomial logistic regression is the highest; 0.68 (ML) > 0.67 (LDA) >0.56 (QDA).
- When the low birth identification is an interest, the QLD yields the false negative rate; 0.11 (QLD)<0.36 (LDA) < 0.54 (ML).
- If the low birth identification is an interest, the ML yields the lowest false positive rate; 0.12 (ML)<0.23 (LDA) < 0.52 (QLD).

More about classification

In this class, classification problem has been tackled by modelling log-odd using a linear regression.

Some popular approaches;

- Finite mixture model
- k-nearest neighbours algorithm
- Decision trees

Forensic Glass Fragments

Broken or shattered glass found at a crime scene is an important piece of forensic evidence.

Forensic Glass Fragments

214 samples of six glass types are collected in forensic work. For each fragment, the refractive index and 8 chemical compounds in percentage by weight of oxides are estimated.

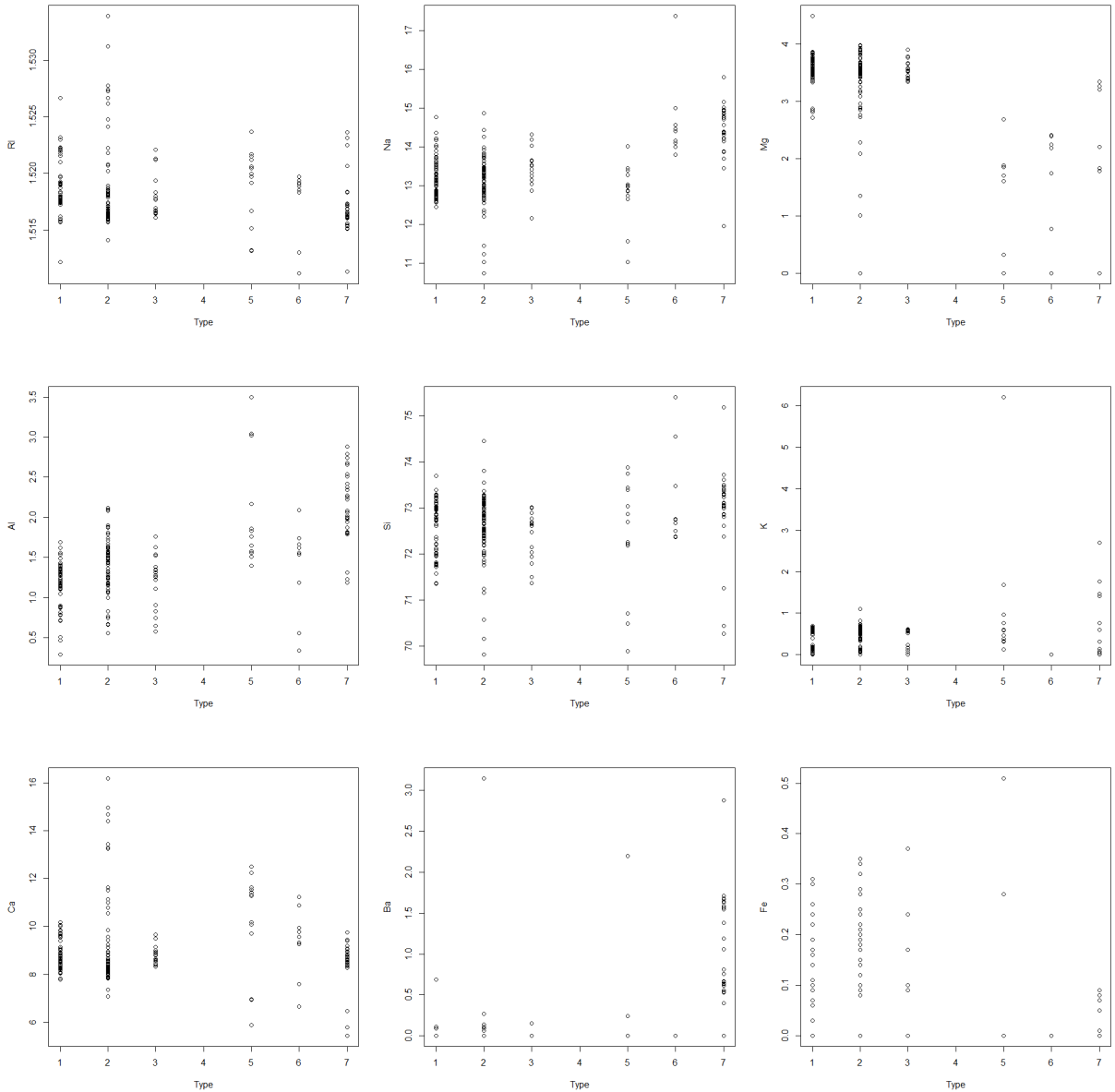
- RI - Refractive index
- Na - sodium.
- Mg - manganese.
- Al - aluminium.
- Si - silicon.
- K - potassium.
- Ca - calcium.
- Ba - barium.
- Fe - iron.

Forensic Glass Fragments

Six glass types follow;

1. window float glass (70 samples)
2. window non-float glass (76 samples)
3. vehicle window glass (17 samples)
5. containers (13 samples)
6. tableware (9 samples)
7. vehicle headlamps (29 samples)

Forensic Glass Fragments



Forensic Glass Fragments

(i) Multinomial regression model

Taking Type 1 as a reference level, intercept and coefficient estimates follow;

	(Intercept)	RI	Na	Mg	Al
2	114.01139	210.99092	-3.5715880	-6.14888398	-0.0777839
3	46.69565	-61.97027	1.6471464	-0.01788714	2.5121161
5	19.54782	14.22700	-0.4893655	-3.69586811	10.1611011
6	-14.59763	-21.52840	10.7663636	-7.48120815	34.9748591
7	-33.83528	22.99089	2.4341715	-5.00880431	6.2849258

	Si	K	Ca	Ba	Fe
2	-4.4509190	-3.70543961	-4.6895169	-5.757871	2.2610525
3	0.2207149	-0.67459086	0.6082768	-2.208131	1.5301451
5	-0.5204113	0.62817476	-0.4292740	-3.450644	-0.6424633
6	-0.9212133	-197.82120395	-4.7069924	-149.906448	-407.9088594
7	-0.1495441	-0.06454676	-2.2076868	-2.475847	-15.9357312

Residual Deviance: 299.4877

AIC: 399.4877

Forensic Glass Fragments

(i) Multinomial regression model
Glass type prediction; y actual type and y' type prediction;

	$y = 1$	$y = 2$	$y = 3$	$y = 5$	$y = 6$	$y = 7$
$y' = 1$	52	19	10	0	0	0
$y' = 2$	18	54	7	3	0	2
$y' = 3$	0	0	0	0	0	0
$y' = 5$	0	1	0	9	0	0
$y' = 6$	0	0	0	0	9	0
$y' = 7$	0	2	0	1	0	27

Forensic Glass Fragments

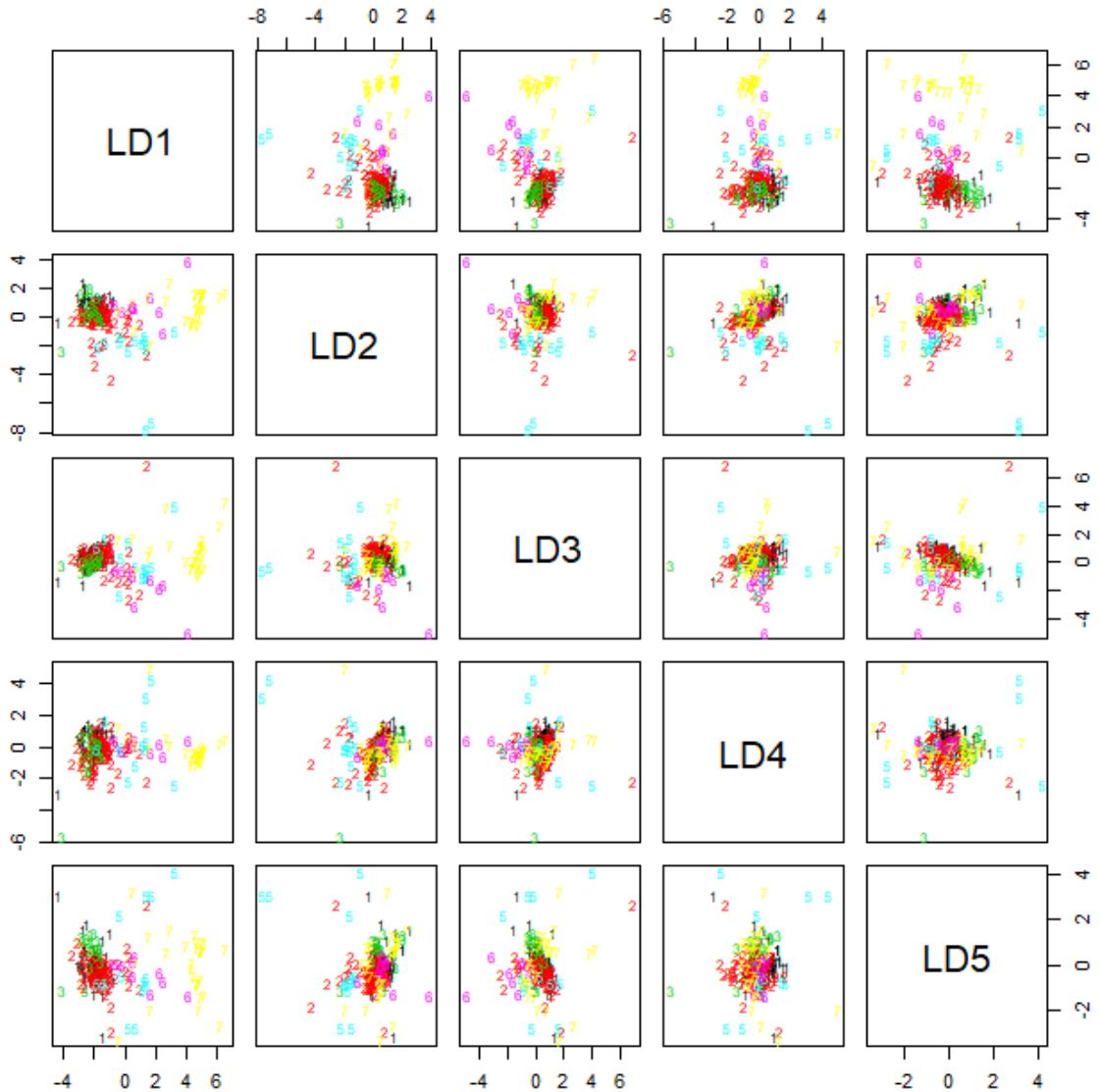
(ii) LDA
Glass type prediction; y actual type and y' type prediction;

	$y=1$	$y=2$	$y=3$	$y=5$	$y=6$	$y=7$
$y'=1$	52	17	11	0	1	1
$y'=2$	15	54	6	5	1	2
$y'=3$	3	0	0	0	0	0
$y'=5$	0	3	0	7	0	1
$y'=6$	0	2	0	0	6	0
$y'=7$	0	0	0	1	0	25

hmm.... this looks similar to the prediction table when the multinomial logistic regression.

Forensic Glass Fragments

(iii) Discriminant function - Types on projected lines



Forensic Glass Fragments

(iiii) QDA
Glass type prediction; y actual type and y' type prediction;

	$y = 1$	$y = 2$	$y = 3$	$y = 5$	$y = 7$
$y' = 1$	63	44	0	0	0
$y' = 2$	5	29	0	3	0
$y' = 3$	2	2	17	0	0
$y' = 5$	0	0	0	10	0
$y' = 7$	0	1	0	0	29

Except Type 2, the classification prediction is improved.

Reference

- The Elements of Statistical Learning. Hastie, T., Tibshirani, R., Friedman, J. Spring Series in Statistics.
  `http://web.stanford.edu/~hastie/ElemStatLearn/`
- Extending the linear model with R. Faraway, J. J. (2006) CHAPMAN & HALL/CRC.