Linear Discriminant Analysis

AW

Lecture Overview

1. Fisher Linear Discriminant
2. Application of LDA
3. Hard Cases

Projection Maps

- Let $\overline{w} \in \mathbb{R}^d$ be a unit vector, i.e. $\|\overline{w}\| = \sqrt{\overline{w}^T\overline{w}} = \sqrt{\overline{w} \bullet \overline{w}} = 1$
- For a vector $\overline{x} \in \mathbb{R}^d$ projection of $\overline{x}$ onto $\overline{w}$ is $proj_{\overline{w}}\overline{x} = \frac{\overline{w}\bullet\overline{x}}{\overline{w}\bullet\overline{w}}\overline{w} = (\overline{w} \bullet \overline{x})\overline{w}$
- Denote $a = \overline{w} \bullet \overline{x}$ coordinate (offset) of x along $\overline{w}$. The coordinate map $proj_{\overline{w}} : \mathbb{R}^d \to \mathbb{R} :: x \to a$ is a linear transformation that maps original d-dimensional space to a 1-dimensional space (along $\overline{w}$)

Denote

- X a subset d dimensional space, i.e. a dataset with d real-valued features, and $Y = \{0, 1\}$ class labels
- $\mathscr{D}$ is (unknown but fixed) distribution on $X \times Y$
- $S = \{(\overline{x}_1, y_1), \ldots, (\overline{x}_m, y_m)\}$ an i.i.d. w.r.t. D sample from $X \times Y$ (training set). Let also $S_j = \{(\overline{x}, y) \in S \mid y = j\}$ be subsample of class $j \in \{0, 1\}$

The set $\{a_1, \ldots, a_m\}$ of coordinates of vectors $\{\overline{x}_1, \ldots, \overline{x}_m\}$ is the image of this set under projection map of training set S

Example of Projection

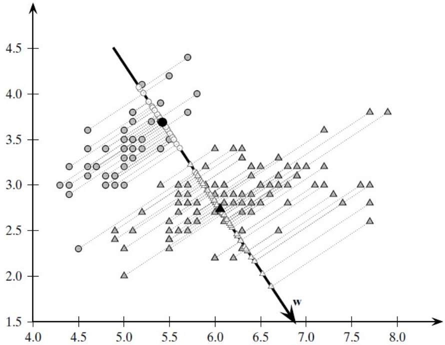


Figure: Projection of Iris data onto a vector $\overline{w}$

2-dimensional Iris dataset (sepal length and sepal width as the attributes); iris-setosa is class c_1 (circles), other two Iris types are class c_2 (triangles). There are $n_1 = 50$ points in c_1 and $n_2 = 100$ points in c_2. One possible vector $\overline{w}$ is shown.

Mean and Scatter in Projection Space

- Let n_i be a number of data points in class i, i.e. $|S_i|$. Then mean of class i under projection along $\overline{w}$ is

$$\begin{aligned} \left\|\mu^i_{\overline{w}}\right\| &= \frac{1}{n_i}\sum_{\overline{x}\in S_i}\|proj_{\overline{w}}\overline{x}\| = \frac{1}{n_i}\sum_{\overline{x}\in S_i} a_x \\ &= \frac{1}{n_i}\sum_{\overline{x}\in S_i}\overline{w}\bullet\overline{x} \\ &= \overline{w}\bullet\left(\frac{1}{n_i}\sum_{\overline{x}\in S_i}\overline{x}\right) = \left\|proj_{\overline{w}}\overline{\mu}^i\right\| \end{aligned}$$

  where $\overline{\mu}^i$ is a multivariate mean of class i

- Scatter of a sample S is unnormalized variance, i.e. $s = (n-1)\sigma^2$ (or $n\sigma^2$ when mean is known). Scatter of class i under projection along $\overline{w}$ is

$$\begin{aligned} \mathsf{s}^i_{\overline{w}} &= \sum_{\overline{x}\in S_i}\left(\|proj_{\overline{w}}\overline{x}\| - \left\|\mu^i_{\overline{w}}\right\|\right)^2 \\ &= \sum_{\overline{x}\in S_i}\left(\|a_x\overline{w}\| - \left\|proj_{\overline{w}}\overline{\mu}^i\right\|\right)^2 \\ &= \sum_{\overline{x}\in S_i}\left(\overline{w}\bullet\overline{x} - \overline{w}\bullet\overline{\mu}^i\right)^2 \\ &= \sum_{\overline{x}\in S_i}\left(\overline{w}\bullet\left(\overline{x}-\overline{\mu}^i\right)\right)\left(\left(\overline{x}-\overline{\mu}^i\right)\bullet\overline{w}\right) \\ &= \left(\overline{w}\bullet\left(\sum_{\overline{x}\in S_i}\left(\overline{x}-\overline{\mu}^i\right)\left(\overline{x}-\overline{\mu}^i\right)^T\right)\right)\bullet\overline{w} \end{aligned}$$

Scatter computation

Let X be a set of vectors $\overline{x}_1, \dots, \overline{x}_k$ from $\mathbb{R}^d$. Denote X matrix formed by these vectors, i.e. $X = [\overline{x}_1, \dots, \overline{x}_k]$. We call $\mathcal{SC}_X = \sum_{i=1}^{k} \left(\overline{x}_i - \overline{\mu}\right)\left(\overline{x}_i - \overline{\mu}\right)^T$ scatter matrix of X.

Notice that $\mathcal{SC}_X = \sum_{i=1}^{k} \left(\overline{x}_i - \overline{\mu}\right)\left(\overline{x}_i - \overline{\mu}\right)^T = X_c^T X_c$ where $X_c = [\overline{x}_1^c, \dots, \overline{x}_k^c]$ is the matrix of centered vectors. It is computed as $X_c = XC_k$ and $C_k = I_k - \frac{1}{k}\overline{1}_k\overline{1}_k^T$ is centering matrix of k vectors. Here I is identity matrix and $\overline{1}$ is 1-vector that has 1 in all positions.

So $\mathcal{SC}_X = X_c^T X_c = (n-1)\Sigma_X$ where Σ_X is covariance matrix of X

So $\mathrm{s}_{\overline{w}}^i = \left(\overline{w} \bullet \left(\sum_{\overline{x} \in S_i} \left(\overline{x} - \overline{\mu}^i\right)\left(\overline{x} - \overline{\mu}^i\right)^T\right)\right)\overline{w} = \overline{w}^T SC_{S_i}\overline{w} = (n-1)\overline{w}\Sigma_{S_i}\overline{w}$

Note that for a non-zero vector $\overline{z} \in \mathbb{R}^d$ we have $\overline{z}^T \mathcal{SC}_X \overline{z} = \overline{z}^T \left(X_c^T X_c\right)\right)\overline{z} = \|X_c \overline{z}\| \geq 0$ i.e. $\mathcal{SC}_X$ is positive semidefinite, so all of its eigenvalues are nonnegative.

Class Separation in Projection Space

Reasonable ways to separate classes is:

- Maximize separation by maximizing the difference between the projections of means, i.e. find such $\overline{w}$ that guarantees max $|\mu_{\overline{w}}{}^{1} - \mu_{\overline{w}}{}^{2}|$;
- Minimize scatter $\mathrm{s}_{\overline{w}}{}^{i}$ in each class. Why?

Class Separation in Projection Space

Reasonable ways to separate classes is:

- Maximize separation by maximizing the difference between the projections of means, i.e. find such $\overline{w}$ that guarantees max $\left|\mu_{\overline{w}}{}^{1} - \mu_{\overline{w}}{}^{2}\right|$;
- Minimize scatter $\mathrm{s}_{\overline{w}}{}^{i}$ in each class. Why?
    - Large variance causes intermixing of points of the two classes because of the spread of the points so no good separation. Thus separation is good if means are separated and the variance of the projected points for each class is not large

Given that the number of datapoints in training set fixed and cannot be changed minimizing variance is the same as minimizing scatter.

We need to choose $\overline{w}$ to optimize both (possibly conflicting) criteria!

Fisher LDA Optimization Objective

To choose projection vector $\overline{w}$ that minimizes the sum of projected scatters of classes AND on maximizes difference between projected means of classes using a single maximization criterion, we can maximize the ratio of objectives. This single objective is called Fisher LDA objective:

$$\max_{\overline{w}\in\mathbb{R}^d} J(\overline{w}) = \frac{\left(\|\mu_{\overline{w}}{}^1\| - \|\mu_{\overline{w}}{}^2\|\right)^2}{\mathtt{s}_{\overline{w}}{}^1 + \mathtt{s}_{\overline{w}}{}^2}$$

Notice that

$$\begin{aligned}
\left(\|\mu_{\overline{w}}{}^1\| - \|\mu_{\overline{w}}{}^2\|\right)^2 &= \left(\overline{w} \bullet (\mu^1 - \mu^2)\right)^2 \\
&= (\overline{w} \bullet (\mu^1 - \mu^2)) \cdot ((\mu^1 - \mu^2) \bullet \overline{w}) \\
&= \overline{w}^T \left((\mu^1 - \mu^2)(\mu^1 - \mu^2)^T\right) \overline{w}
\end{aligned}$$

Let $S_B = (\mu^1 - \mu^2)(\mu^1 - \mu^2)^T$. It is called Between Classes Scatter Matrix. Then $\left(\|\mu_{\overline{w}}{}^1\| - \|\mu_{\overline{w}}{}^2\|\right)^2 = \overline{w}^T S_B \overline{w}$

Fisher Optimization Problem

We have shown that $\mathtt{s}_{\overline{w}}^{i} = \overline{w}^{T} SC_{S_i} \overline{w}$, so

$$\begin{aligned} \mathtt{s}_{\overline{w}}^{1} + \mathtt{s}_{\overline{w}}^{2} &= \overline{w}^{T} SC_{S_1} \overline{w} + \overline{w}^{T} SC_{S_2} \overline{w} \\ &= \overline{w}^{T} \left(SC_{S_1} + SC_{S_2} \right) \overline{w} \\ &= \overline{w}^{T} S_W \overline{w} \end{aligned}$$

where $S_W = SC_{S_1} + SC_{S_2}$ is Within Classes Scatter Matrix. Since both SC_{S_1} and SC_{S_2} are positive semidefinite S_W is also positive semi-definite.
We can now write Fisher LDA optimization problem as

$$\max_{\overline{w} \in \mathbb{R}^d} J(\overline{w}) = \max_{\overline{w} \in \mathbb{R}^d} \frac{\overline{w}^{T} S_B \overline{w}}{\overline{w}^{T} S_W \overline{w}}$$

Digression: Covariance Matrix is Semi-definite

A $n \times n$ matrix M is positive-definite (positive-semidefinite) if for every vector $vecy \in \mathbb{R}^n$ holds $\vec{y}^T M \vec{y} > 0$ (resp $\vec{y}^T M \vec{y} \geq 0$.

For a sample S of n random vectors$\{\vec{x}_i \in \mathbb{R}^k | i = 1, \ldots, n\}$, the mean vector is

$$\vec{\mu} = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i,$$

the sample covariance matrix is

$$Q = \frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu}).$$

So, for a nonzero vector $\vec{y} \in \mathbb{R}^k$, we have

$$\begin{array}{rcl} \vec{y}^T Q \vec{y} & = & \vec{y}^T \left(\frac{1}{n} \sum_{i=1}^{n} (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu}) \right) \vec{y} \\ & = & \frac{1}{n} \sum_{i=1}^{n} \vec{y}^T (\vec{x}_i - \vec{\mu})(\vec{x}_i - \vec{\mu}) \vec{y} \\ & = & \frac{1}{n} \sum_{i=1}^{n} \left((\vec{x}_i - \vec{\mu})^T \vec{y} \right)^2 \geq 0 \end{array}$$

Therefore, Q is always positive semi-definite, and so is scatter.

Digression: Invertibility of Positive-definite Matrix

A $n \times n$ matrix M is positive-definite if for every vector $vecy \in \mathbb{R}^n$ holds $\vec{y}^T M \vec{y} > 0$.

Hence 0 is not its eigen-value of M. If $M\vec{x} = 0\vec{x}$ for some non-zero $\vec{x}$, so for this vector $\vec{y}^T M \vec{y} = 0$, so which contradicts M to positive-definiteness of M

Equivalently, M is invertible, because if it is not then $M\vec{x} = 0$ has non-trivial solution.

Solving Fisher LDA

Helpful Vector Differentiation Identities:

Let matrix A be not a function of vector $\overline{x}$, while vector $\overline{u} = \overline{u(x)}$ and $\overline{v} = \overline{v(x)}$. Then

$$\frac{d}{d\overline{x}}\left(\overline{u}^T A \overline{v}\right) = \frac{d}{d\overline{x}}\left(\overline{u}^T A\right)\overline{v} + \frac{d}{d\overline{x}}\left(A\overline{v}\right)\overline{u} \qquad \frac{d\,\overline{x}}{d\overline{x}} = I = \frac{d\,\overline{x}^T}{d\overline{x}}$$

$$\frac{d}{d\overline{x}}\left(A\overline{v}\right) = \frac{d\,\overline{v}}{d\overline{x}}A^T \qquad \frac{d}{d\overline{x}}\left(\overline{v}^T A\right) = \frac{d\,\overline{v}^T}{d\overline{x}}A$$

Taking derivative of Fisher objective:

$$\frac{d}{d\overline{w}}\left(\frac{\overline{w}^T S_B \overline{w}}{\overline{w}^T S_W \overline{w}}\right) = \frac{(S_B^T\overline{w}+S_B\overline{w})\left(\overline{w}^T S_W \overline{w}\right)-(S_W^T\overline{w}+S_W\overline{w})\left(\overline{w}^T S_B \overline{w}\right)}{\left(\overline{w}^T S_W \overline{w}\right)^2} = 0$$

Using symmetry of S_B and S_W and assuming S_W is positive definite ($\overline{w}^T S_W \overline{w}$) it becomes

$$2S_B\overline{w}\left(\overline{w}^T S_W \overline{w}\right) = 2S_W\overline{w}\left(\overline{w}^T S_B \overline{w}\right)$$

Solving Fisher LDA

Taking derivative of Fisher objective:

$$\frac{d}{d\overline{w}}\left(\frac{\overline{w}^T S_B \overline{w}}{\overline{w}^T S_W \overline{w}}\right) = \frac{(S_B^T\overline{w}+S_B\overline{w})\left(\overline{w}^T S_W \overline{w}\right)-(S_W^T\overline{w}+S_W\overline{w})\left(\overline{w}^T S_B \overline{w}\right)}{\left(\overline{w}^T S_W \overline{w}\right)^2} = 0$$

Using symmetry of S_B and S_W and assuming S_W is positive definite ($\overline{w}^T S_W \overline{w}$) it becomes

$$2 S_B \overline{w}\left(\overline{w}^T S_W \overline{w}\right) = 2 S_W \overline{w}\left(\overline{w}^T S_B \overline{w}\right)$$

Or equivalently

$$S_B \overline{w} = S_W \overline{w}\left(\frac{\overline{w}^T S_B \overline{w}}{\overline{w}^T S_W \overline{w}}\right) = J(\overline{w}) S_W \overline{w}$$

Notice that $J(\overline{w})$ is a real number. By assumption S_W is positive-definite, so S_W^{-1} exists then we can multiply both sides by it. We get

$$S_W^{-1} S_B \overline{w} = S_W^{-1} J(\overline{w}) S_W \overline{w} = J(\overline{w}) S_W^{-1} S_W \overline{w} = J(\overline{w})\overline{w}$$

Solving Fisher LDA continued

We are going to address the case when S_W^{-1} does not exist later (this means it is not positive-definite and that 0 is its eigenvalue).

Denote for now $\lambda = J(\overline{w})$ and $A = S_W^{-1} S_B$. Then we have $A\overline{w} = \lambda\overline{w}$. So what does this well known equation tells you?

Solving Fisher LDA continued

We are going to address the case when S_W^{-1} does not exist later (this means it is not positive-definite and that 0 is its eigenvalue).

Denote for now $\lambda = J(\overline{w})$ and $A = S_W^{-1} S_B$. Then we have $A\overline{w} = \lambda\overline{w}$. So what does this well known equation tells you?
λ is eigenvalue of $A = S_W^{-1} S_B$. In addition

- Since we are looking to maximize $J(\overline{w})$ we are looking for a greatest eigenvalue of A
- $J(\overline{w}) = \frac{\overline{w}^T S_B \overline{w}}{\overline{w}^T S_W \overline{w}}$ so it must be real. Thus we are looking for greatest real eigenvalue of A and its eigenvector $\overline{w}$

If $S_W^{-1} S_B$ has real eigenvalue we have a required vector $\overline{w}$ and can construct the Fisher LDA classification algorithm.

We'll address the case when $S_W^{-1} S_B$ has no real eigenvalues later.

Learning Fisher LDA

- Separate training instances of class 0 and 1 into sets X_0 and X_1. Compute sizes m_0 and m_1 of these sets.
- Compute multidimensional means $\overline{\mu}_0$ and $\overline{\mu}_1$ of sets X_0 and X_0 respectively. Then compute $S_B = (\overline{\mu}_0 - \overline{\mu}_1)(\overline{\mu}_0 - \overline{\mu}_1)^T$
- For $i = 0, 1$ compute mean centered matrices $X_i^c = X_i \left(I_{m_i} - \frac{1}{m_i} \overline{1}_{m_i} \overline{1}_{m_i}^T \right)$
- For $i = 0, 1$ compute scatter matrices $SC_i = (X_i^c)^T (X_i^c)$ and compute $S_W = SC_1 + SC_2$
- If S_W is non-invertible stop and return failure
- Otherwise compute $A = S_W^{-1} S_B$
- Compute largest real eigenvalue λ of A if any. If no real eigenvalue exists then stop and return failure.
- For largest real eigenvalue λ of A compute its normalized eigenvector $\overline{w}$
- Compute $a_{\overline{w}}^i = \mathrm{ext}_{x_j \in C_i} \{ \overline{w} \bullet \overline{x}_j \}$ for $i \in \{0, 1\}$ where $\mathrm{ext} \in \{\max, \min\}$, $p = \frac{|a_0 - a_1|}{2}$ and $sep = p + \min_{i \in 0,1} a_i$

Classifying with Fisher LDA

Classifier:

Input: a new instance $\overline{x} \in \mathbb{R}^d$

- For an input $\overline{x}$ compute $a_x = \overline{x} \bullet \overline{w}$ where $\overline{w}$ is the result of learning algorithm
- If $a_x \leq sep$ then assign to $\overline{x}$ label j returned by learning algorithm. Otherwise assign to $\overline{x}$ label $1 - j$.

Notes on Learning Fisher LDA classifier:

- Note that the computation of separation points finding two closest projections that belong to opposite classes and finding midpoint between them obtaining maximum margin. If classes overlap then this way of separation fails. Soft separation by using midpoint between projections of means is also used sometimes.
- In terms of computational complexity, computing S_W takes $O(md^2)$ time, and computing the dominant eigenvalue-eigenvector pair takes $O(d^3)$ time in the worst case. Thus, the total time is $O(d^3 + md^2)$.

Lecture Overview

1 Fisher Linear Discriminant

2 Application of LDA

3 Hard Cases

Fisher LDA Classification Example

Given the training set below. Classify $\overline{x} = (2, 10)^T$

Datapoint feature	1	2	3	4	5	6	7	8	9	10
A_1	4	2	2	3	4	9	6	9	8	10
A_2	2	4	3	6	4	10	8	5	7	8

$$\begin{aligned}
X_0 &= \{\{4,2,2,3,4\}\{2,4,3,6,4\}\} \\
X_1 &= \{\{9,6,9,8,10\},\{10,8,5,7,8\}\} \\
\overline{\mu}_0 &= \left(3, \tfrac{19}{5}\right) \qquad \overline{\mu}_1 \;=\; \left(\tfrac{42}{5}, \tfrac{38}{5}\right) \\
X_0^c &= \{\{1,-1,-1,0,1\},\{-9/5,1/5,-4/5,11/5,1/5\}\} \\
X_1^c &= \{\{3/5,-12/5,3/5,-2/5,8/5\},\{12/5,2/5,-13/5,-3/5,2/5\}\} \\
SC_0 &= \{\{4,-1\},\{-1,44/5\}\} \\
SC_1 &= \{\{46/5,-1/5\},\{-1/5,66/5\}\} \\
S_W &= SC_0 + SC_1 \;=\; \{\{66/5,-6/5\},\{-6/5,22\}\} \\
S_B &= \{\{729/25,513/25\},\{513/25,361/25\}\}
\end{aligned}$$

Fisher LDA Classification Example

Given the training set below. Classify $\overline{x} = (2, 10)^T$

Datapoint feature	1	2	3	4	5	6	7	8	9	10
A_1	4	2	2	3	4	9	6	9	8	10
A_2	2	4	3	6	4	10	8	5	7	8

$$S_W^{-1} S_B = \tfrac{1}{3010}\{\{6939, 4883\}, \{3186, 2242\}\}$$

$\lambda = \frac{9181}{3010} = 3.050$ is maximum eigenvalue

$\overline{w} = \begin{pmatrix} \frac{257}{118} \\ 1 \end{pmatrix}$ is its eigenvector. Normalized $\overline{w} = \begin{pmatrix} 0.909 \\ 0.417 \end{pmatrix}$

$$\mu_{\overline{w}}^{0} = \overline{w} \bullet \overline{\mu}_0 = 4.3166 \qquad \overline{\mu}_{\overline{w}}^{1} = \overline{w} \bullet \overline{\mu}_1 = 10.8048$$

$$sep = 7.5607 \qquad j = 0$$

Fisher LDA Classification Example

Given the training set below. Classify $\overline{x} = (2, 10)^T$

Datapoint feature	1	2	3	4	5	6	7	8	9	10
A_1	4	2	2	3	4	9	6	9	8	10
A_2	2	4	3	6	4	10	8	5	7	8

$$\overline{w} = \begin{pmatrix} \frac{257}{118} \\ 1 \end{pmatrix} \text{ is its eigenvector. Normalized } \overline{w} = \begin{pmatrix} 0.909 \\ 0.417 \end{pmatrix}$$

$$\mu_{\overline{w}}^{0} = \overline{w} \bullet \overline{\mu}_0 = 4.3166 \qquad \overline{\mu}_{\overline{w}}^{1} = \overline{w} \bullet \overline{\mu}_1 = 10.8048$$

$$sep = 7.5607 \qquad j = 0$$

$$\overline{x} \bullet \overline{w} = \begin{pmatrix} 2 \\ 10 \end{pmatrix} \bullet \begin{pmatrix} 0.909 \\ 0.417 \end{pmatrix} = 5.978 < 7.5607$$

$\begin{pmatrix} 2 \\ 10 \end{pmatrix}$ is labeled 0

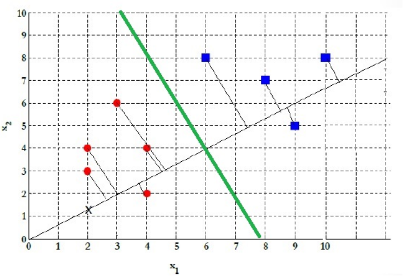
Fisher LDA Classification Example - cont.

Class X_0:

$$\begin{aligned}&Proj_{\overline{w}}:\\&(4,2)\rightarrow 4.469\overline{w};\\&(2,4)\rightarrow 3.486\overline{w};\\&(2,3)\rightarrow 3.069\overline{w};\\&(3,6)\rightarrow 5.229\overline{w};\\&(4,4)\rightarrow 5.304\overline{w}\end{aligned}$$

Class X_1:

$$\begin{aligned}&Proj_{\overline{w}}:\\&(9,10)\rightarrow 12.351\overline{w};\\&(6,8)\rightarrow 8.79\overline{w};\\&(9,5)\rightarrow 10.266\overline{w};\\&(8,7)\rightarrow 10.191\overline{w};\\&(10,8)\rightarrow 12.426\overline{w}\end{aligned}$$



$sep = 7.5607$ correctly separates classes.

LDA in R

```
library(MASS);library(mlbench);library(sets)
data(BreastCancer)
x<-sample(1:nrow(BreastCancer),nrow(BreastCancer)/3,F)
BrCaLDA<-lda(Class ~ .,BreastCancer[-x,-1])
plot(BrCaLDA, panel = panel.lda, cex = 0.7,
         dimen = 1, abbrev = FALSE,type="density")
table(BreastCancer[-x,]$Class,
        predict(BrCaLDA,BreastCancer[-x,2:10])$class)
acc.BrCaLDATr<-sum(BreastCancer[-x,]$Class=
    =predict(BrCaLDA,BreastCancer[-x,2:10])$class,
   na.rm=TRUE)/ nrow(BreastCancer[-x,]);acc.BrCaLDATr
#warnings()
pred <- predict(BrCaLDA,BreastCancer[x,2:10])
table(BreastCancer[x,]$Class,pred$class)
acc.BrCaLDA <- sum(pred$class=
           =BreastCancer[x,]$Class,na.rm=TRUE)/
          nrow(BreastCancer[x,]);acc.BrCaLDA
```

Lecture Overview

1. Fisher Linear Discriminant
2. Application of LDA
3. Hard Cases

$S_W^{-1}S_B$ Lacks Real Eigenvalues

S_B is symmetric by construction, so we can find C such that $C^2 = S_B$ as follows:

- S_B is symmetric so orthogonally diagonalizable, i.e. $S_B = U\Lambda U^T$ where λ is diagonal matrix of eigenvalues and U is orthogonal matrix of normalized eigenvectors.
- We obtain $\Lambda^{\frac{1}{2}}$ by taking square roots of entries of Λ, so $\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}} = \Lambda$
- For orthogonal matrices $U_1 = U^T$, so taking $C = U\Lambda^{\frac{1}{2}}U^T$ we have $C^2 = (U\Lambda^{\frac{1}{2}}U^T)(U\Lambda^{\frac{1}{2}}U^T) = U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T = C$

C is obviously invertible, take $C^{-1} = U\left(\Lambda^{\frac{1}{2}}\right)^{-1}U^T$.

We need to solve $S_W^{-1}S_B\overline{w} = \lambda\overline{w}$. Take $\overline{w} = C^{-1}\overline{v}$ for some unknown $\overline{v}$. Since $S_B = CC$ we need $\overline{v}$ such that $S_W^{-1}S_B\overline{w} = S_W^{-1}(CC)C^{-1}\overline{v} = S_W^{-1}C\overline{v} = \lambda C^{-1}\overline{v}$. In other words, $\lambda\overline{v} = CS_W^{-1}C\overline{v}$. Here C is symmetric by construction (diagonal matrix is symmetric + MGM^T is symmetric whenever G is). So $C = C^T$ which means $CS_W^{-1}C = CS_W^{-1}C^T$ is symmetric because S_W is. So $CS_W^{-1}C$ has d real eigenvalues, which is what we needed.

S_W is Singular

What if the matrix $S_W = SC_0 + SC_1$ is singular?

- There are 2 possible cases:
  1. $ColS_B \subseteq ColS_W$
  2. $ColS_B \nsubseteq ColS_W$
- In case 1 the Moore-Penrose pseudo-inverse matrix S_W^+ can be used in place of S_W^{-1} for the solution and it gives correct answer. Note that S_B is a rank one matrix, so to check which case it is, we only need to check for one column of S_B if it is in $ColS_W$.
- In case 2 the story gets complicated and LDA may not have a solution for given features. Then to apply LDA change of coordinates (PCA) is in order possibly with reduction of dimensionality

Reading

ZM sections 1.3, 20.1, 20.2, 20.3