

# Applied Machine Learning

## Linear Models (2)

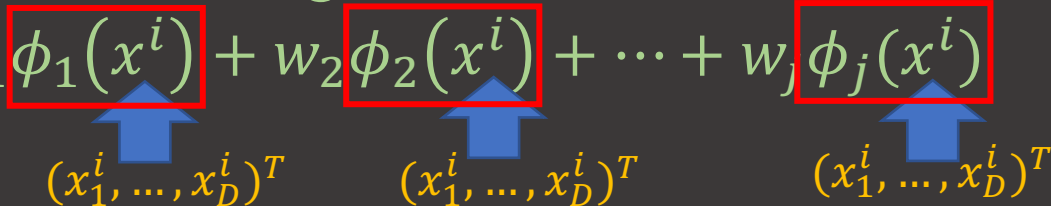
Computer Science, Fall 2022

Instructor: Xuhong Zhang

# Recap from last lecture: Regression

- Given data  $X = \{x_1, \dots, x_n\}$ , where  $x_i \in \mathbb{R}^d$
- Corresponding labels  $y = \{y_1, \dots, y_n\}$ , where  $y_i \in \mathbb{R}$ 
  - $X$  is high-dimensional input
  - $Y$  is single-numeric value

# Recap from last lecture: Regression

- $y(x^i, w) = w_0 + w_1 x_1^i + \cdots + w_D x_D^i$  where  $x^i = (x_1^i, \dots, x_D^i)^T$ 
  - it is a linear function of the parameters  $w_0, \dots, w_D$ .
  - It is also a linear function of the input variables  $x^i$ .
  - $y(x^i, w)$  is a single numerical value.
- $y(x^i, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x^i)$ 
  - $y(x^i, w)$  is a single numerical value.
  - $M$  is a fixed number, an integer.
  - $\sum_{j=1}^{M-1} w_j \phi_j(x^i) = w_1 \phi_1(x^i) + w_2 \phi_2(x^i) + \cdots + w_j \phi_j(x^i)$ 

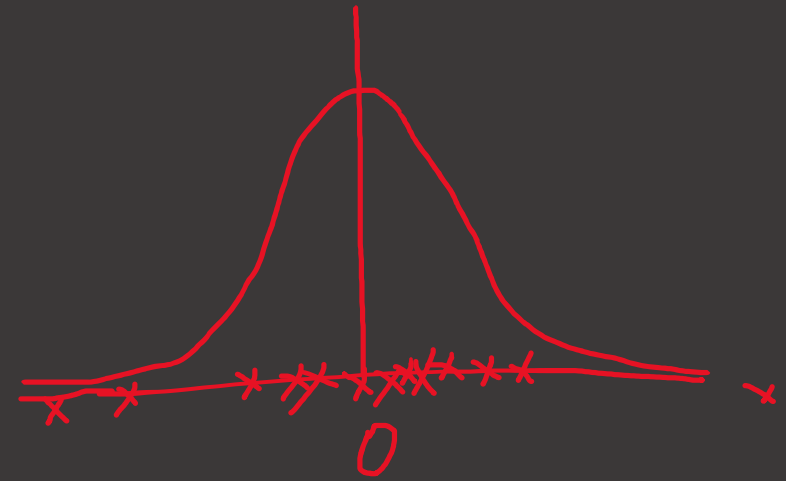
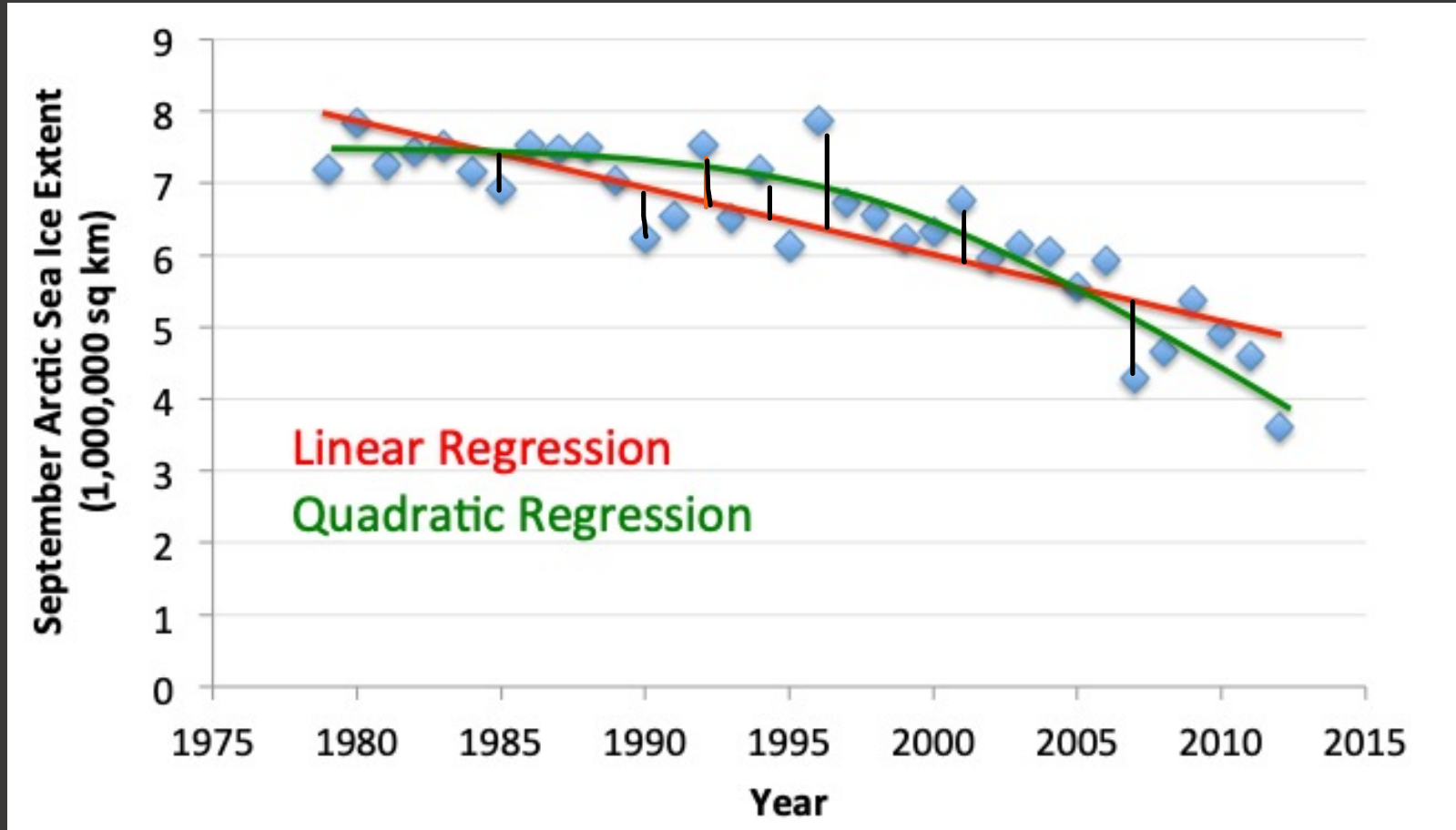
# Recap from last lecture: Regression

- Solving for  $w$  we obtain

$$w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t$$

$$\Phi = \begin{pmatrix} \boxed{\phi_0(\mathbf{x}^1)} & \cdots & \boxed{\phi_{M-1}(\mathbf{x}^1)} \\ \vdots & \ddots & \vdots \\ \boxed{\phi_0(\mathbf{x}^N)} & \cdots & \boxed{\phi_{M-1}(\mathbf{x}^N)} \end{pmatrix}$$

# Recap from last lecture: Regression



# Linear Models for Classification

- The goal in classification is to take an input vector  $\mathbf{x}$  and to assign it to one of the  $K$  discrete classes  $C_k$  where  $k = 1, \dots, K$ .
- In the most common scenario, the classes are taken to be disjoint, so that each input is assigned to one and only one class.
- The input space is thereby divided into decision regions whose boundaries are called decision boundaries or decision surfaces.

# Linear Models for Classification

- Here the decision surfaces are linear functions of the input vector  $\mathbf{x}$  and hence are defined by  $(D - 1)$  dimensional hyperplanes within the  $D$  dimensional input space.
- Datasets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable.
- There are different ways of using target values to represent class labels.
  - Binary representation, a single target variable  $t \in \{0,1\}$
  - $K > 2$ , use a 1-of- $K$  coding scheme, e.g.  $t = (0,1,0,0,0)^T$

# Linear Models for Classification

- For linear regression model, we have the form  $y(x) = w^T x + w_0$ , so that  $y$  is a real number.
- For classification problem, however, we wish to predict discrete class labels. To achieve this, we consider a generalization of this model in which we transform the linear function of  $w$  using a nonlinear function  $f(\cdot)$ .

$$y(x) = f(w^T x + w_0)$$

$f(\cdot)$  is activation  
function/link function

The model is called a  
generalized linear model.

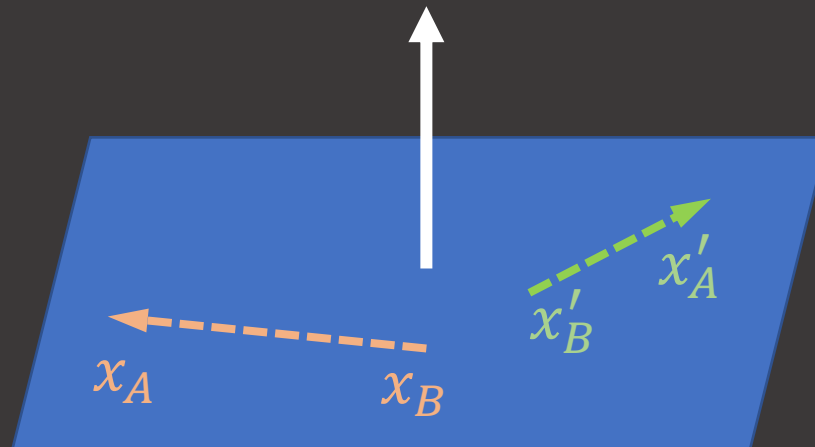


# Discriminant Functions

- A discriminant is a function that takes an input vector  $x$  and assigns it to one of  $K$  classes, denoted  $C_k$ .
- Suppose we focus on two classes, the simplest representation of a linear discriminant function is  $y(x) = w^T x + w_0$  and
  - Assign  $x$  to class  $C_1$  if  $y(x) \geq 0$  and to class  $C_2$  otherwise.
  - The corresponding decision boundary is therefore defined by the relation  $y(x) = 0$ , which corresponds to a  $(D - 1)$  –dimensional hyperplane within the  $D$ -dimensional input space.

# Properties

- Now consider two points  $x_A$  and  $x_B$  both of which lie on the decision surface.
- Because  $y(x_A) = y(x_B) = 0$ , we have  $w^T(x_A - x_B) = 0$ , and hence the vector  $w$  is orthogonal to every vector lying within the decision surface.



# Fisher's linear discriminant

- One way to view a linear classification model is in terms of dimensionality reduction.
- Consider the case of two classes, and suppose we take the  $D$ -dimensional input vector  $\mathbf{x}$  and project it down to one dimension using

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$$

- If we place a threshold on  $y$  and classify  $y \geq -w_0$  as class  $\mathcal{C}_1$  and otherwise class  $\mathcal{C}_2$ , we can get a standard linear classifier.

# Fisher's linear discriminant

- In general, the projection onto one dimension leads to a considerable loss of information, and classes that are well separated in the original  $D$ -dimensional space may become strongly overlapping in one dimension.
- **The KEY:** However, we can select a projection that maximizes the class separation by adjusting the components of the weight vector  $w$ .

# Fisher's linear discriminant

- Now suppose for the two-classes problem, we have  $N_1$  points of class  $C_1$  and  $N_2$  points of class  $C_2$ . So the mean vectors of the two classes are given by

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n, \text{ and } \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

- The simplest measure of the separation of the classes, when projected onto  $w$ , is the separation of the projected class means.

$$m_2 - m_1 = w^T (\mathbf{m}_2 - \mathbf{m}_1)$$

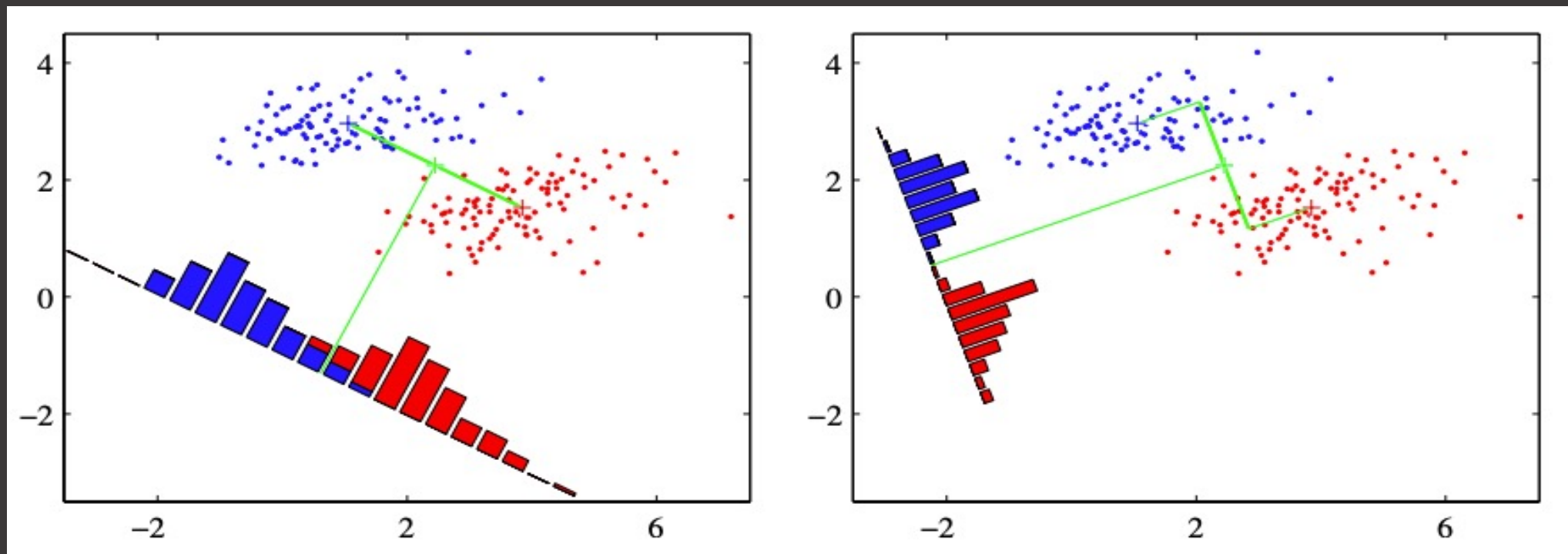
$$m_k = w^T \mathbf{m}_k$$

# Fisher's linear discriminant

- Problem 1: In order to enlarge the expression  $m_2 - m_1 = w^T(\mathbf{m}_2 - \mathbf{m}_1)$ , we can make the magnitude of  $w$  arbitrarily large.
- We could constrain  $w$  to have unit length, so that  $\sum_i w_i^2 = 1$ .

# Fisher's linear discriminant

## Problem 2:



There is considerable class overlap in the projected space on the left plot. The right plot shows the projection Based on the Fisher linear discriminant, showing the greatly improved class separation.

# Fisher's linear discriminant

- The idea proposed by Fisher is to maximize a function that will give a large separation between the projected class means while also giving a small variance within each class, thereby minimizing the class overlap.
- Suppose  $m_k = w^T \mathbf{m}_k$  is the mean of the projected data from class  $C_k$ . The within-class variance of the transformed data from class  $C_k$  is defined as

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2 \text{ where } y_n = w^T x_n$$



# Fisher's linear discriminant

- With  $s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$  for the within-class variance, we can define the total within-class variance for the whole dataset to be simply  $s_1^2 + s_2^2$ .
- The fisher criterion is defined to be the ratio of the between-class variance to the within-class variance and is given by

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \xrightarrow[\substack{y(x) = \mathbf{w}^T \mathbf{x} \\ m_k = \mathbf{w}^T \mathbf{m}_k \\ s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2}]{\hspace{1cm}} J(w) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

# Fisher's linear discriminant

- $J(w) = \frac{w^T S_B w}{w^T S_W w}$  where

- $S_B$  is the between-class covariance matrix and is given by

$$S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

- $S_W$  is the total within-class covariance matrix, given by

$$S_W = \sum_{n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

# Fisher's linear discriminant

- Now differentiating  $J(w) = \frac{w^T S_B w}{w^T S_w w}$  with respect to  $w$ , we can find that  $J(w)$  is maximized when

$$(w^T S_B w) S_w w = (w^T S_w w) S_B w$$

- Because  $S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$ , we see that  $S_B w$  is always in the direction of  $(\mathbf{m}_2 - \mathbf{m}_1)$ .
- In addition, we do not care about the magnitude of  $w$ , only its direction, and so we can drop the scalar factor  $(w^T S_B w)$  and  $(w^T S_w w)$ .
- Multiplying both sides by  $S_w^{-1}$ , we then get
$$w \propto S_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

# Fisher's linear discriminant

$$w \propto S_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

is known as Fisher's linear discriminant.

- Strictly it is not a discriminant but rather a specific choice of direction for projection of the data down to one dimension.
- However, the projected data can subsequently be used to construct a discriminant, by choosing a threshold  $y_0$  so that we classify a new point as belonging to  $\mathcal{C}_1$  if  $y(x) \geq y_0$  and classify it as belonging to  $\mathcal{C}_2$  otherwise.

# Fisher's discriminant for multiple classes

- Now let's generalize the Fisher's discriminant to  $K > 2$  classes and we shall assume that the dimensionality  $D$  of the input space is greater than the number of  $K$  of classes.
- We introduce  $D' > 1$  linear 'features'  $y_k = \mathbf{w}_k^T \mathbf{x}$ , where  $k = 1, \dots, D'$ .
- The weight vectors  $\{\mathbf{w}_k\}$  can be considered to be the columns of a matrix  $\mathbf{W}$ , so that
$$\mathbf{y} = \mathbf{W}^T \mathbf{x}$$
- We are not including any bias for  $\mathbf{y}$ .

# Fisher's discriminant for multiple classes

- Similar to the within class covariance matrix to the case of 2 classes, for  $K$  classes, we have

$$S_w = \sum_{k=1}^K S_k$$

where

$$S_k = \sum_{n \in C_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$$
$$\mathbf{m}_k = \frac{1}{N_k} \sum_{n \in C_k} \mathbf{x}_k$$

# Fisher's discriminant for multiple classes

- To generalize the between-class covariance, let's first consider the total covariance matrix

$$S_T = \sum_{n=1}^N (\mathbf{x}_n - \mathbf{m})(\mathbf{x}_n - \mathbf{m})^T$$

where  $\mathbf{m}$  is the mean of the total data set

$$\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n = \frac{1}{N} \sum_{k=1}^K N_k \mathbf{m}_k$$

- Now the key point is, the total covariance matrix can be decomposed into the sum of the with-in class covariance matrix, plus an additional matrix  $S_B$ , which is a measure of the between-class covariance

$$S_T = S_W + S_B$$

# Fisher's discriminant for multiple classes

$$S_T = S_W + S_B$$

where

$$S_B = \sum_{k=1}^K N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

- Note that these covariance matrices have been defined in the original  $\mathbf{x}$ -space. We can now define similar matrices in the projected  $D'$ -dimensional  $y$ -space

$$s'_W = \sum_{k=1}^K \sum_{n \in C_k} (y_n - \mu_k)(y_n - \mu_k)^T \text{ and } s'_B = \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

$$\text{where } \mu_k = \frac{1}{N_k} \sum_{n \in C_k} y_n, \text{ and } \mu = \frac{1}{N} \sum_{k=1}^K N_k \mu_k$$



# Fisher's discriminant for multiple classes

- Again we wish to construct a scalar that is large when the between-class covariance is large and when the within-class covariance is small.
- There are different choices of criterion. One example is given by

$$J(W) = \text{Tr}\{S_W'^{-1} S_B'\}$$

- This criterion can then be rewritten as an explicit function of the projection matrix  $W$  in the form

$$J(w) = \text{Tr}\{(WS_W W^T)^{-1} (WS_B W^T)\}$$

# Fisher's discriminant for multiple classes

- The weight values are determined by those eigenvectors of  $s_W'^{-1}s_B'$  that correspond to the  $D'$  largest eigenvalues.  
See Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition (Second ed.). Academic Press.

# Fisher's discriminant for multiple classes

- One more thing needs attention is that

$$S_B = \sum_{k=1}^K N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

is composed of the sum of  $K$  matrices, each of which is an outer product of two vectors and therefore of rank 1.

- In addition, only  $(K - 1)$  of these matrices are independent as a result of the constraint  $\mathbf{m} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n = \frac{1}{N} \sum_{k=1}^K N_k \mathbf{m}_k$ .
  - Thus,  $S_B$  has rank at most to  $(K - 1)$  and so there are at most  $(K - 1)$  nonzero eigenvalues.
  - This shows that the projection onto the  $(K - 1)$  –dimensional subspace spanned by the eigenvectors of  $S_B$  does not alter the value of  $J(W)$ , and so we are therefore unable to find more than  $(K - 1)$  linear ‘features’ by this means.