# CS722/822: Machine Learning

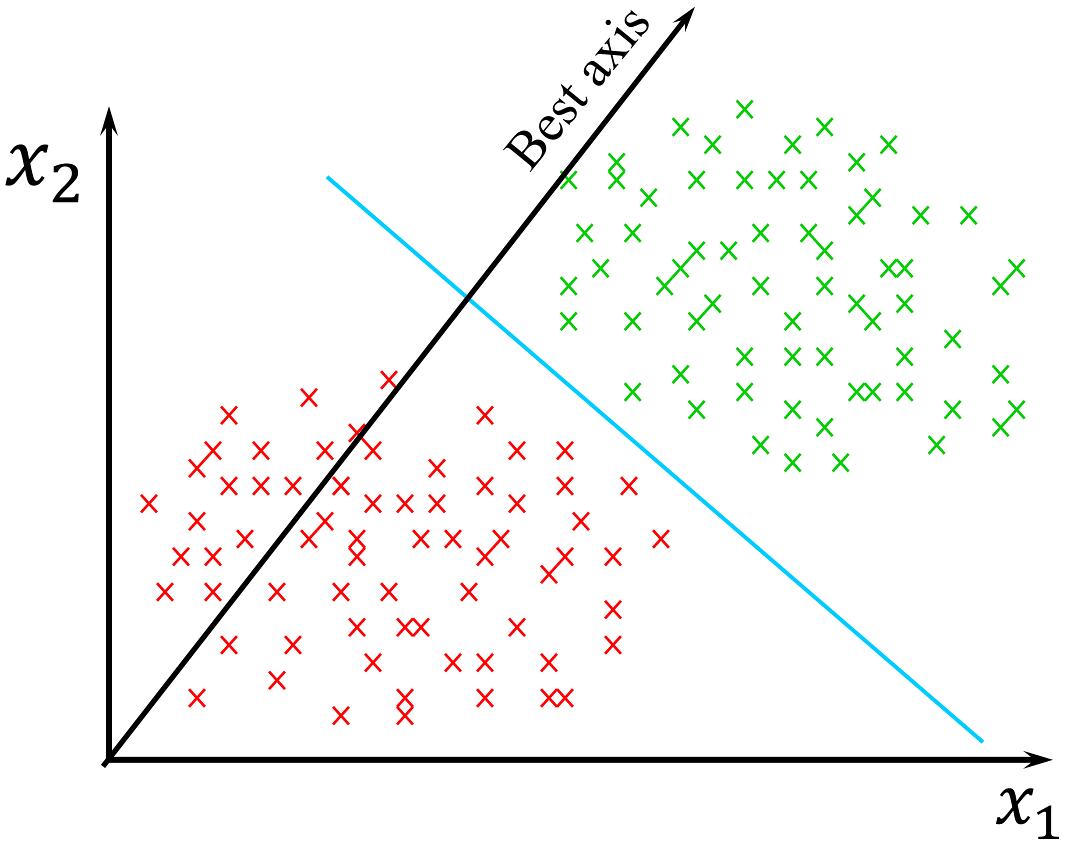
Instructor: Jiangwen Sun
Computer Science Department

# Linear Discriminant Analysis
# (Fisher's Linear Discriminant)

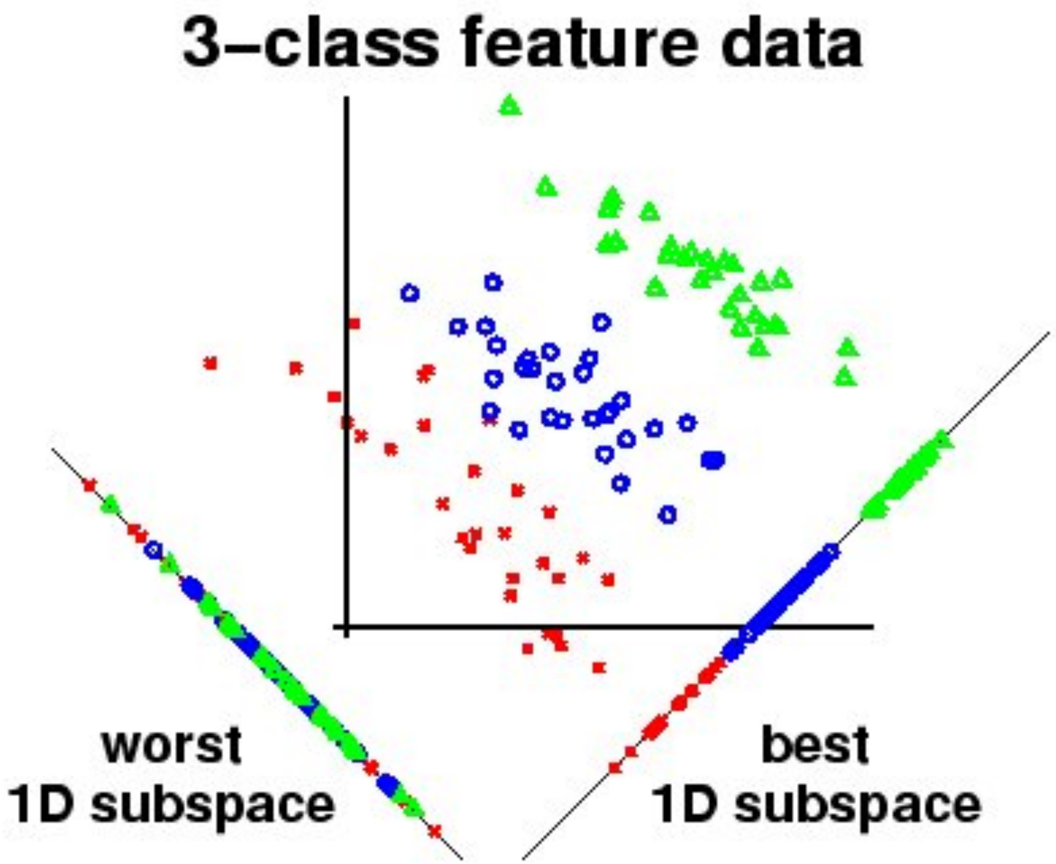
# Classification

- Training data is given
  - Each object is associated with a class label $y \in \{1, 2, \dots, K\}$ and a feature vector of $d$ measurements: $\mathrm{x} = (x_1, \dots, x_d)$.
- Build a predictive model: $f(\mathrm{x}) \rightarrow y$ from the training data.
- Unseen objects are to be classified as belonging to one of a number of predefined classes $\{1, 2, \dots, K\}$.
- Linear Discriminant Analysis / Fisher's linear discriminant

# Two classes



# Three classes



# Classifiers

- Given a training set

$$L = \{(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \ldots, (\mathrm{x}_n, y_n)\}$$

- Classifier $f$ built from $L$:

$$f\colon \mathrm{x} \to y \in \{1,2, \ldots, K\}$$

- **Bayes classifier** base on conditional densities $p(k|\mathrm{x})$,

$$f(\mathrm{x}) = \mathrm{argmax}_k p(k|\mathrm{x})$$

  This is a *maximum a posterior*, and $p(k|\mathrm{x})$ is a posterior density

# The Rules of Probability

- **Sum Rule**

$$p(X) = \sum_Y p(X,Y)$$

- **Product Rule**

$$p(X,Y) = p(Y|X)p(X)$$
$$= p(X|Y)p(Y)$$

- **Bayes' Rule**

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$p(X) = \sum_Y p(X|Y)p(Y)$ is irrelevant to $Y = k$

$$p(Y = k|X = \mathrm{x}) \propto p(X = \mathrm{x}|Y = k)p(Y = k)$$

posterior ∝ likelihood × prior

# Maximum a posterior

- $p(k|\mathrm{x}) = p(\mathrm{x}|k)p(k)/p(\mathrm{x})$

- Find a class label $k$ so that

$$\mathrm{argmax}_k\, p(k|\mathrm{x}) = \underset{k}{\mathrm{argmax}}\, p(\mathrm{x}|k)p(k)$$

- Naïve Bayes assumes independence among all features given the class:

$$p(\mathrm{x}|k) = p(x_1|k)p(x_2|k)\cdots p(x_d|k)$$

Very strong assumption

# Multivariate normal dist for each class

Assume multivariate Gaussian (normal) class densities $(\mathrm{x}|k) \sim N(\mu_k, \Sigma_k)$:

$$p(\mathrm{x}|k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}\left((\mathrm{x}-\mu_k)^T \Sigma_k^{-1} (\mathrm{x}-\mu_k)\right)\right)$$

**Maximizing posterior** is equivalent to **maximizing $\mathbf{p(x|k)p(k)}$**, and equivalent to **maximizing the logarithm of $\mathbf{p(x|k)p(k)}$**

$$\log(p(\mathrm{x}|k)p(k)) = -\frac{1}{2}\left((\mathrm{x}-\mu_k)^T \Sigma_k^{-1} (\mathrm{x}-\mu_k)\right) - \log\left(\sqrt{(2\pi)^d |\Sigma_k|}\right) + \log(p(k))$$

$$f(\mathrm{x}) = \underset{k}{\operatorname{argmin}}\left\{(\mathrm{x}-\mu_k)^T \Sigma_k^{-1} (\mathrm{x}-\mu_k) + \log|\Sigma_k| - 2\log(p(k))\right\}$$

# Two-class case

Two class labels: $k_1$ and $k_2$

If $p(\mathrm{x}|k_1)p(k_1) > p(\mathrm{x}|k_2)p(k_2)$ $\Rightarrow$ $f(\mathrm{x}) = k_1$

otherwise $\Rightarrow$ $f(\mathrm{x}) = k_2$

Equivalently, $\frac{p(\mathrm{x}|k_1)p(k_1)}{p(\mathrm{x}|k_2)p(k_2)} > 1$ $\Rightarrow$ $\frac{p(\mathrm{x}|k_1)}{p(\mathrm{x}|k_2)} > \frac{p(k_2)}{p(k_1)}$

$\Rightarrow$ $\log\frac{p(\mathrm{x}|k_1)}{p(\mathrm{x}|k_2)} > \log\frac{p(k_2)}{p(k_1)}$

$$(\mathrm{x}-\mu_1)^T\Sigma_1^{-1}(\mathrm{x}-\mu_1) + \log|\Sigma_1| - (\mathrm{x}-\mu_2)^T\Sigma_2^{-1}(\mathrm{x}-\mu_2) + \log|\Sigma_2| < T$$

$T$ Is a constant

# Guassian discriminant rule

- For multivariate Gaussian (normal) class densities, i.e.,

$$\{\mathrm{x} | (y = k)\} \sim N(\mu_k, \Sigma_k),$$

  the classification rule (predictive function or model) is

$$f(\mathrm{x}) = \underset{k}{\operatorname{argmin}} \left\{ (\mathrm{x} - \mu_k)^T \Sigma_k^{-1} (\mathrm{x} - \mu_k) + \log|\Sigma_k| - 2\log(p(k)) \right\}$$

- In two-class, this is a quadratic rule (Quadratic discriminant analysis, or QDA)

$$(\mathrm{x} - \mu_1)^T \Sigma_1^{-1} (\mathrm{x} - \mu_1) + \log|\Sigma_1| - (\mathrm{x} - \mu_2)^T \Sigma_2^{-1} (\mathrm{x} - \mu_2) + \log|\Sigma_2| < T$$

- In practice, population mean vectors $\mu_k$ and covariance matrices $\Sigma_k$ are estimated by corresponding sample quantities

# Sample mean and variance

- Class mean

$$\mu_k = \frac{1}{|C_k|} \sum_{\mathrm{x} \in C_k} \mathrm{x}$$

- Class covariance

$$\Sigma_k = \frac{1}{|C_k|} \sum_{\mathrm{x} \in C_k} (\mathrm{x} - \mu_k)(\mathrm{x} - \mu_k)^T$$

# Example

$$X_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}.$$

$$\mu = \frac{1}{3}\left(X_1 + X_2 + X_3\right) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\Sigma = \frac{1}{3}\left((X_1 - \mu)(X_1 - \mu)^T + (X_2 - \mu)(X_2 - \mu)^T + (X_3 - \mu)(X_3 - \mu)^T\right)$$

$$= \frac{1}{3}\left(\begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}\begin{pmatrix} 0 & -1 & 0 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}\begin{pmatrix} -1 & 1 & 0 \end{pmatrix}\right)$$

$$= \frac{1}{3}\left(\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

# Two-class case

- If the two classes have the same covariance matrix, $\Sigma_k = \Sigma$ , the discriminant rule is linear (Linear discriminant analysis, or LDA; FLDA for $K = 2$):
- Quadratic rule

$$(\mathrm{x}-\mu_1)^T\Sigma_1^{-1}(\mathrm{x}-\mu_1)+\log|\Sigma_1|-(\mathrm{x}-\mu_2)^T\Sigma_2^{-1}(\mathrm{x}-\mu_2)+\log|\Sigma_2|<T$$

Become $\mathrm{x}^T\Sigma^{-1}(\mu_2-\mu_1)<c$

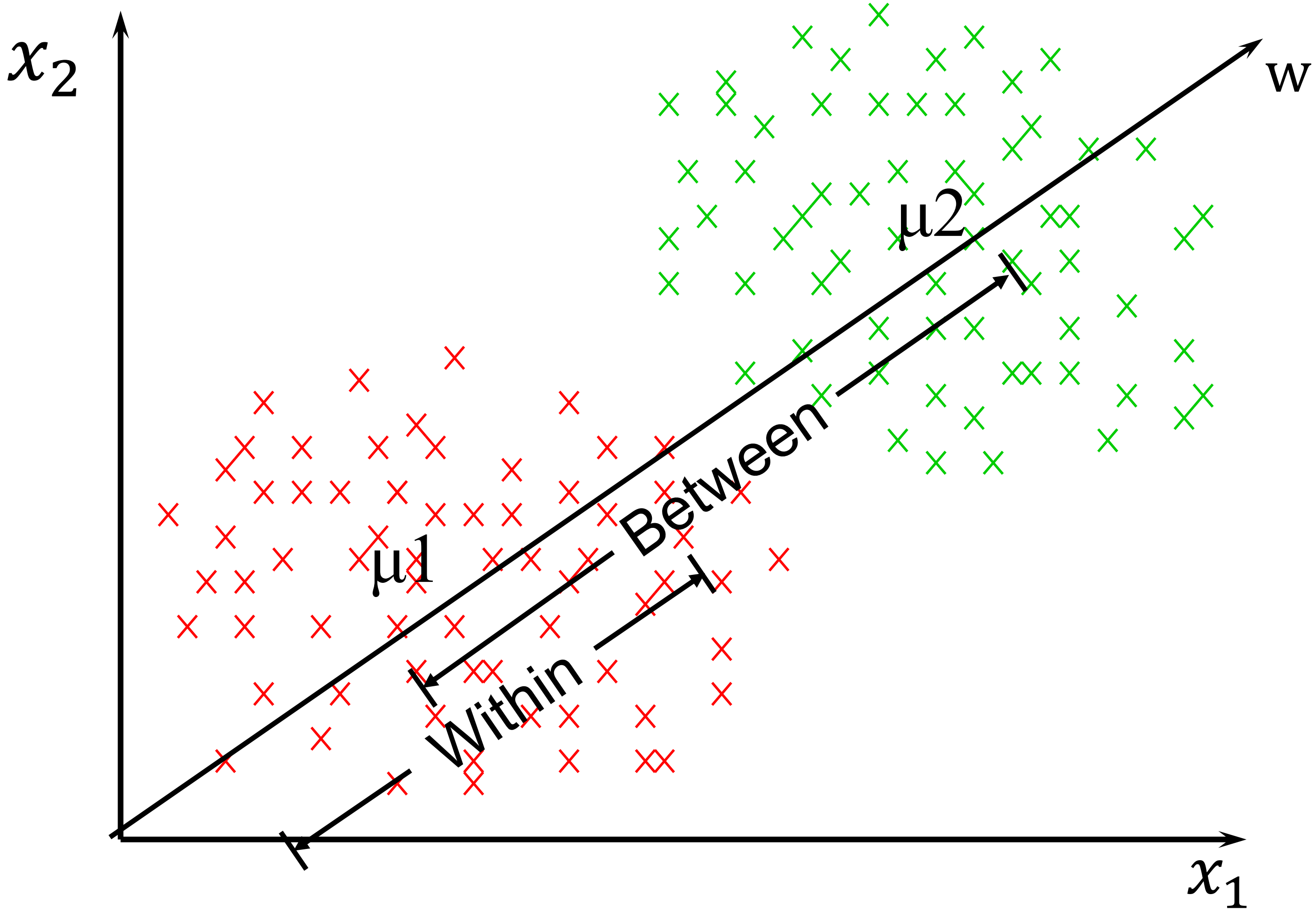
$\boxed{\mathrm{x}^T\mathrm{w}<\mathrm{c},}$ where $\mathrm{w}=\Sigma^{-1}(\mu_2-\mu_1)$

Linear rule

In practice,

$$\Sigma=\frac{1}{n}(n_1\Sigma_1+n_2\Sigma_2)$$

# Illustration




**Good separation**

Large between class distance and small within class distance

# Two-class case

- Maximize the signal-to-noise ratio

$$\max_{\mathrm{w}} \frac{\mathrm{w}^T \Sigma_{between} \mathrm{w}}{\mathrm{w}^T \Sigma_{within} \mathrm{w}}$$

$\longrightarrow$ Between-class separation

$\longrightarrow$ Within-class separation

$$\boxed{s.t.\ \|\mathrm{w}\| = 1}$$

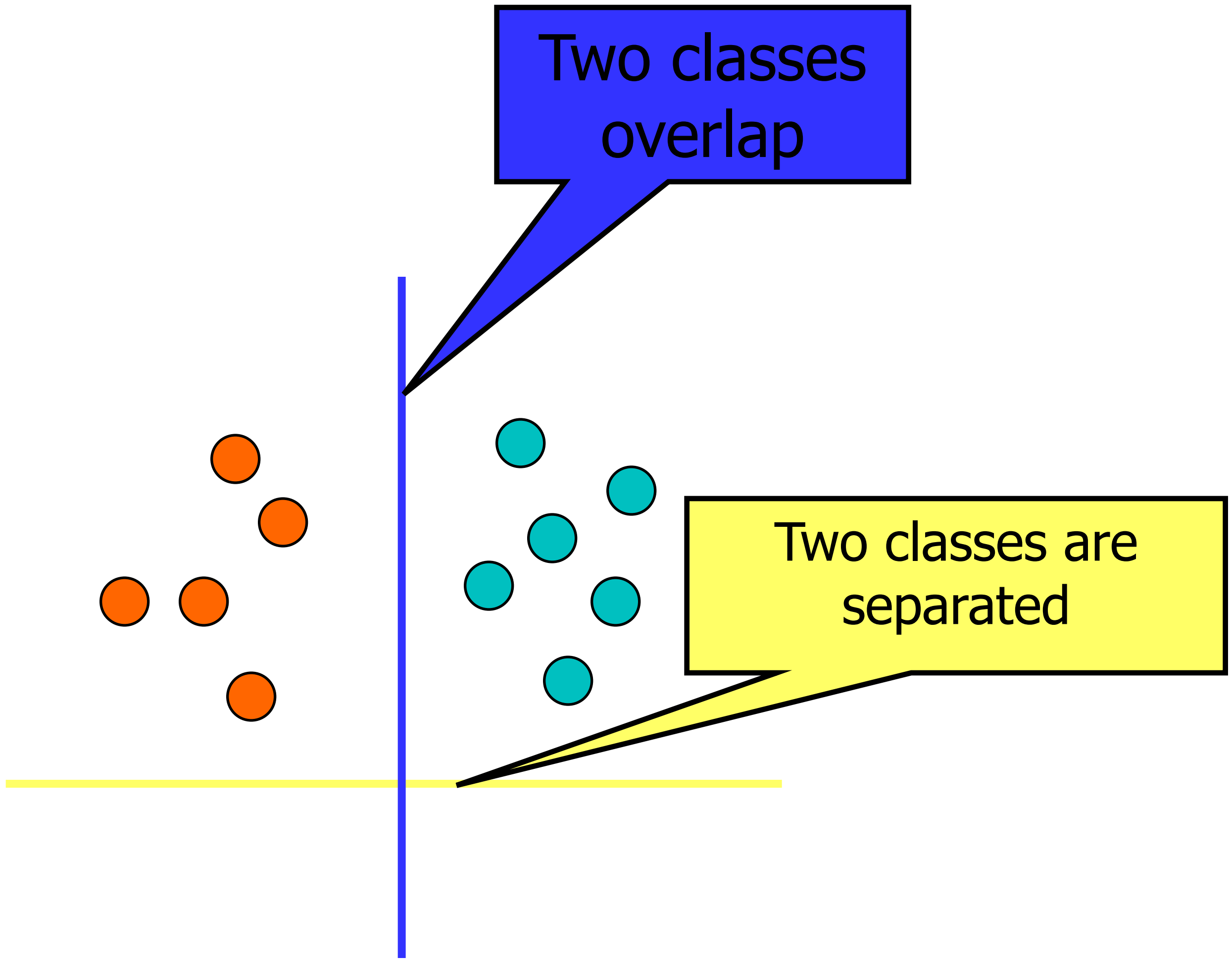
where $\Sigma_{between} = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$

$$\Sigma_{within} = \frac{1}{n}(n_1\Sigma_1 + n_2\Sigma_2)$$

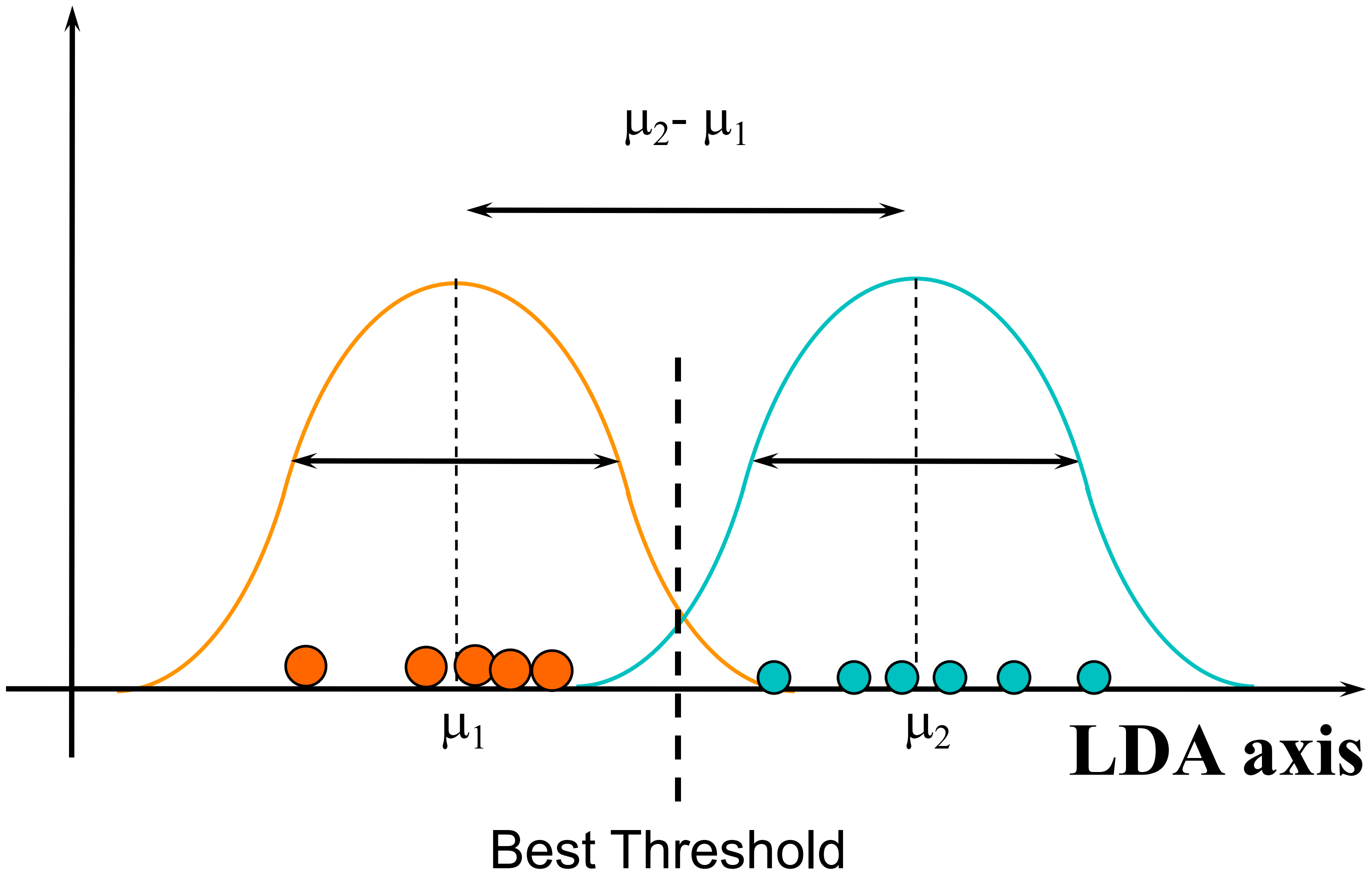
Solution is $\Sigma_{within}{}^{-1}(\mu_2 - \mu_1)$

# Two-class case (illustration)

- LDA gives the yellow direction

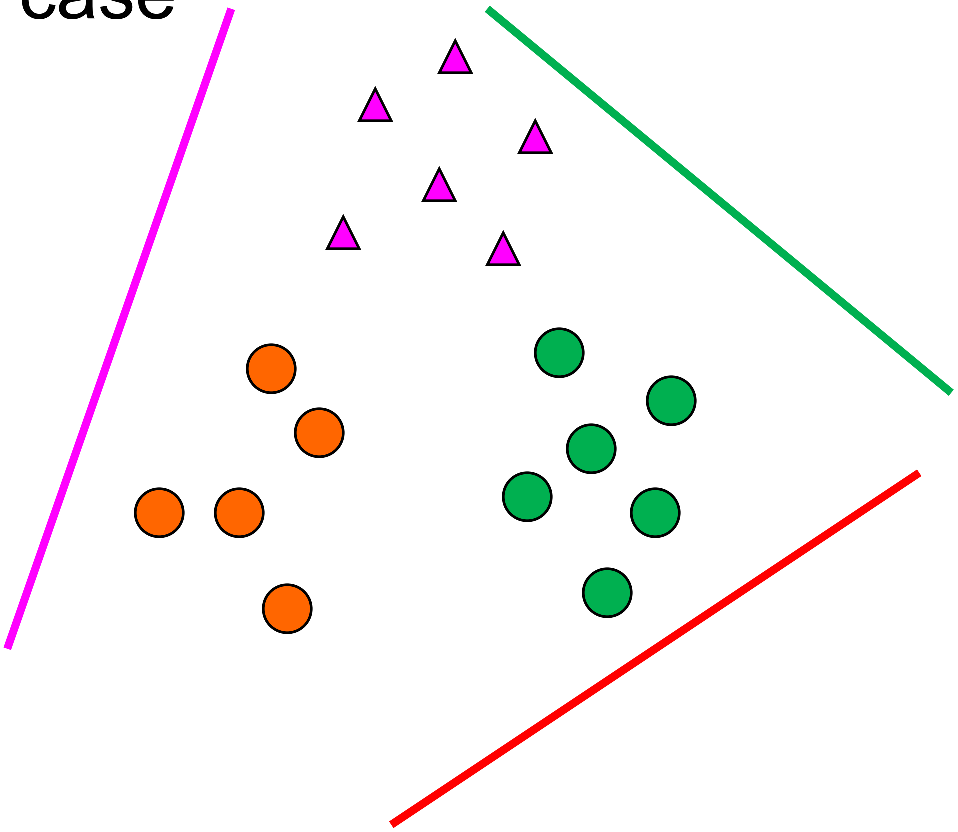


# Two-class case (illustration)



# Multi-class case

- Two approaches
  - Apply two-class LDA to each "one-versus-rest" case



# Multi-class case

**Second approach**: find multiple directions that form a low dimensional space

Transformation matrix W that projects the data to be most separable is the one that maximizes

$$\max_{\mathrm{W}} \ \mathrm{trace}\left(\frac{\mathrm{W}^T \mathrm{S}_b \mathrm{W}}{\mathrm{W}^T \mathrm{S}_w \mathrm{W}}\right)$$

$$s.t.\ \mathrm{W}^T\mathrm{W} = \mathrm{I}$$

Between-class matrix: $S_b = \frac{1}{n}\sum_{i=1}^{K} n_i(\mu_i - \mu)(\mu_i - \mu)^T$

Within-class matrix: $S_w = \frac{1}{n}\sum_{i=1}^{K} \sum_{\mathrm{x}\in C_i}(\mathrm{x} - \mu_i)(\mathrm{x} - \mu_i)^T$

# Intuition

The goal is to simultaneously maximize the between-class separation and minimize the within-class separation

The solution to:

$$\max_{\mathrm{W}} \ \mathrm{trace}\left(\frac{\mathrm{W}^T \mathrm{S}_b \mathrm{W}}{\mathrm{W}^T \mathrm{S}_w \mathrm{W}}\right)$$

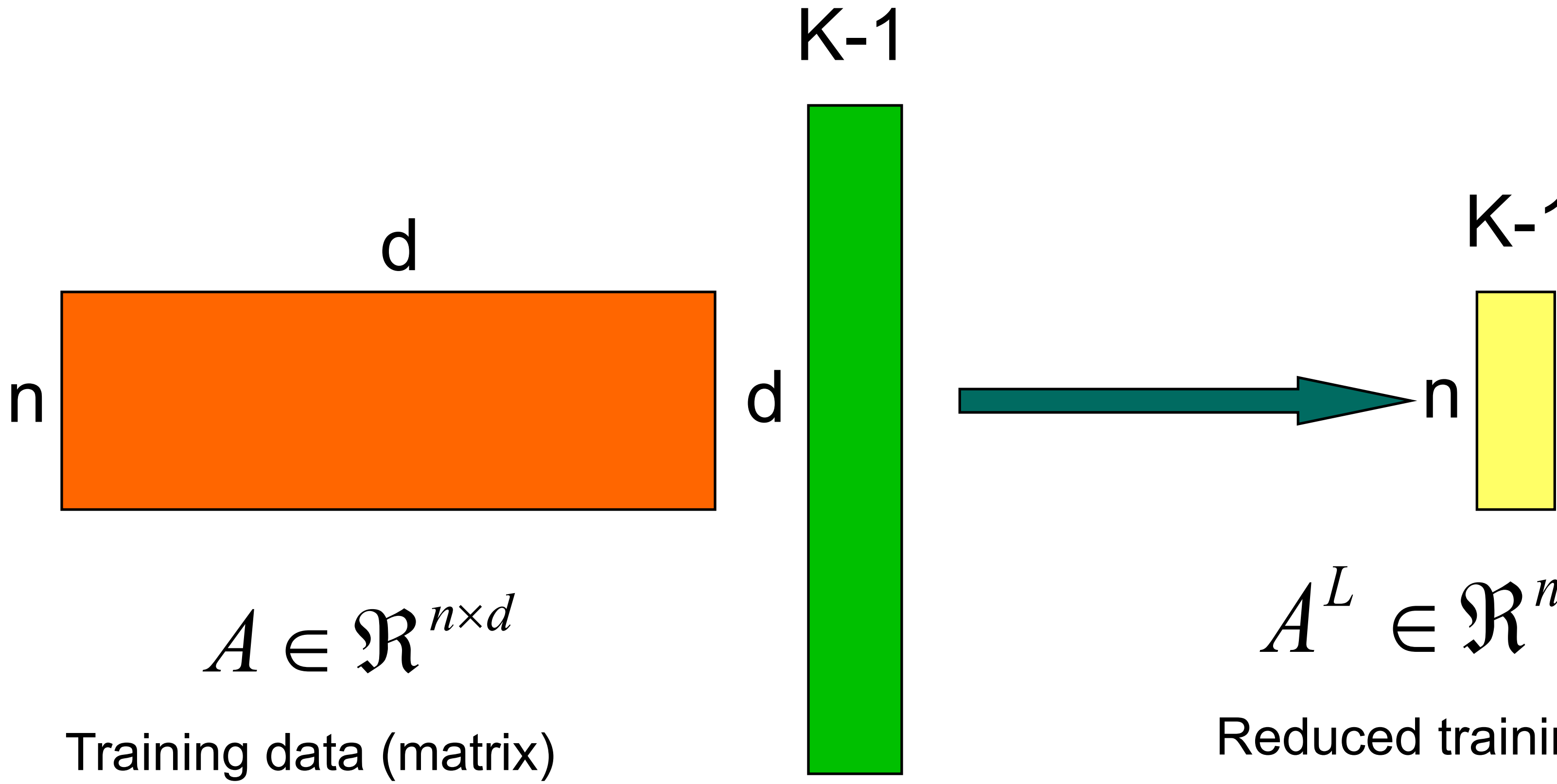
$$s.t. \ \mathrm{W}^T\mathrm{W} = \mathrm{I}$$

is the generalized eigenvalue problem:

$$\mathrm{S}_b \mathrm{w} = \lambda \mathrm{S}_w \mathrm{w}$$

where we want to find the eigenvectors associated with the first $k - 1$ largest eigenvalues.

# Graphic view of the transformation (projection)




$$A \in \Re^{n \times d}$$

Training data (matrix)

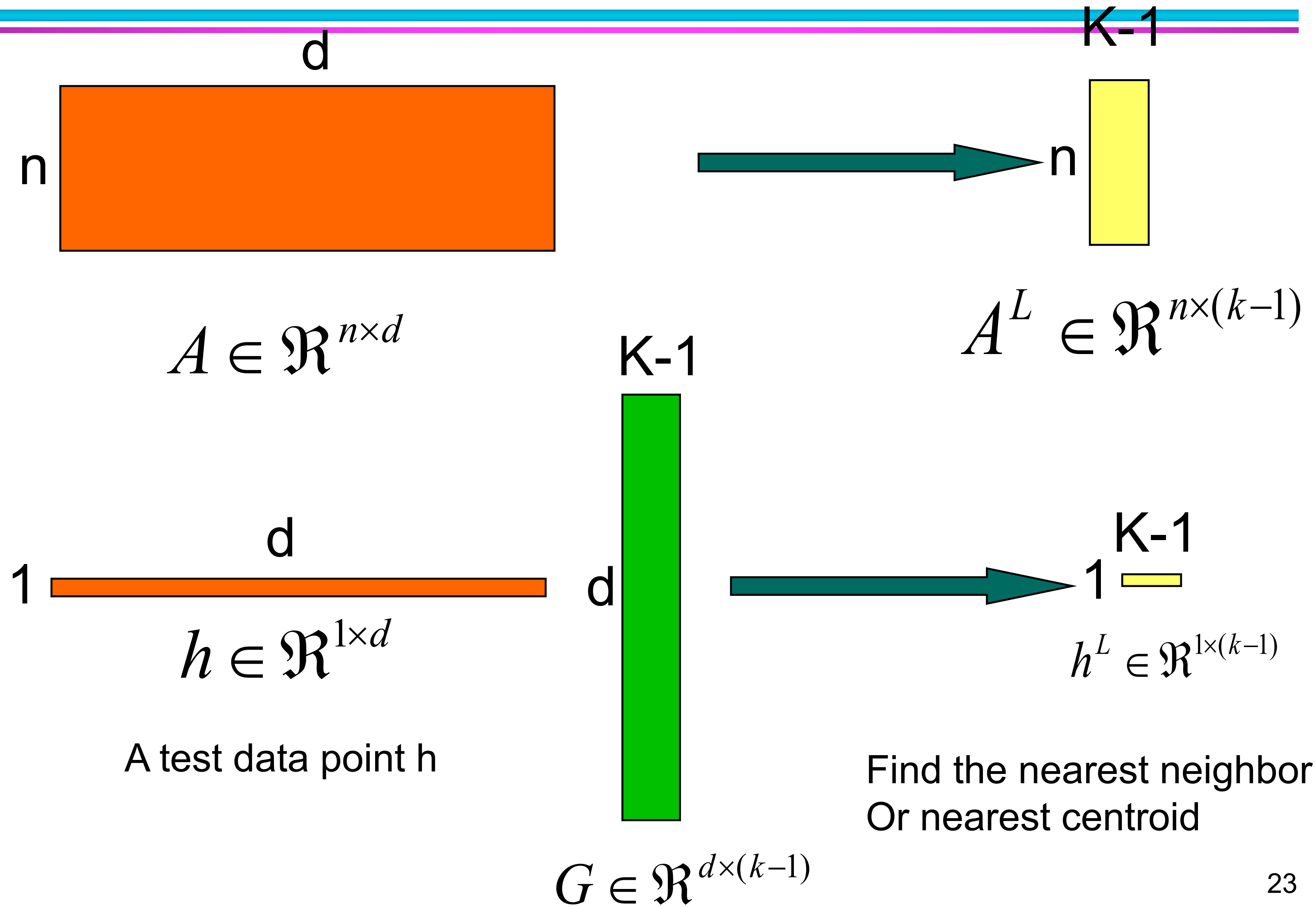
$$W \in \Re^{d \times (k-1)}$$

Transformation matrix

$$A^L \in \Re^{n \times (k-1)}$$

Reduced training data

# Graphical view of classification



# Summary

- Dimension reduction
  - Finds linear combinations of the features $X = \{x_1, \cdots, x_d\}$ with large ratios of between-groups to within-groups sums of squares - **discriminant variables**;
- Classification
  - Predicts the class of an observation $\mathrm{x}$ by the class whose mean vector is closest to $\mathrm{x}$ in terms of the discriminant variables