Advanced Data Analysis: Fisher Discriminant Analysis

Masashi Sugiyama (Computer Science)

W8E-406, sugi@cs.titech.ac.jp

http://sugiyama-www.cs.titech.ac.jp/~sugi

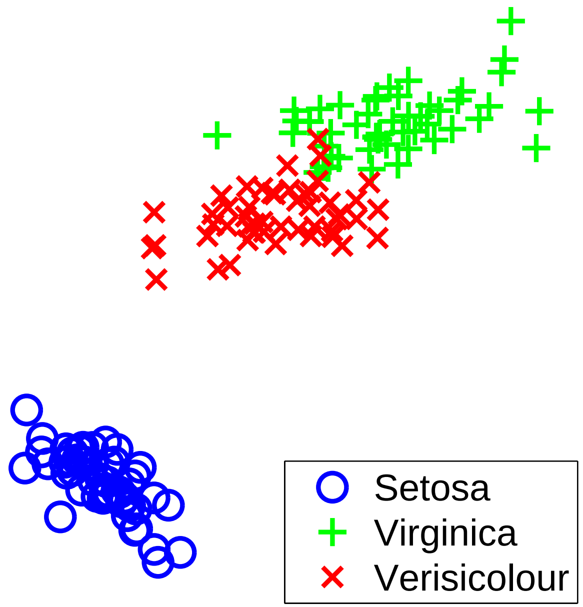
Supervised Dimensionality Reduction

- Samples $\{\boldsymbol{x}_i\}_{i=1}^n$ have class labels $\{y_i\}_{i=1}^n$:

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$$

$$\boldsymbol{x}_i \in \mathbb{R}^d$$
$$y_i \in \{1, 2, \dots, c\}$$

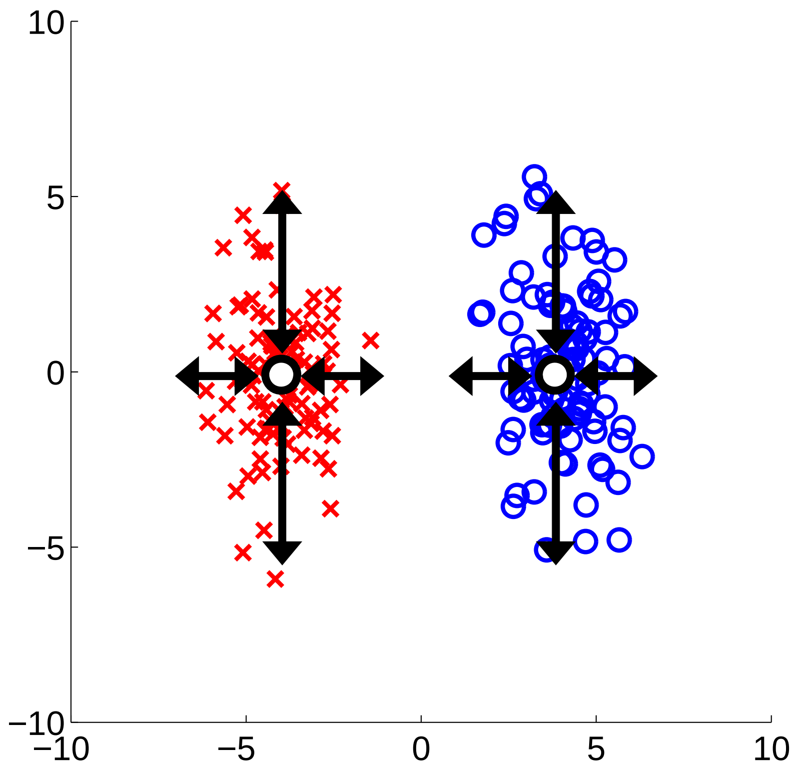
- We want to obtain an embedding such that samples in different classes are well separated from each other!



Within-Class Scatter Matrix

- Sum of scatter within each class:

$$\boldsymbol{S}^{(w)} = \sum_{y=1}^{c} \sum_{i:y_i=y} (\boldsymbol{x}_i - \boldsymbol{\mu}_y)(\boldsymbol{x}_i - \boldsymbol{\mu}_y)^\top$$



$$\boldsymbol{\mu}_y = \frac{1}{n_y} \sum_{i:y_i=y} \boldsymbol{x}_i$$

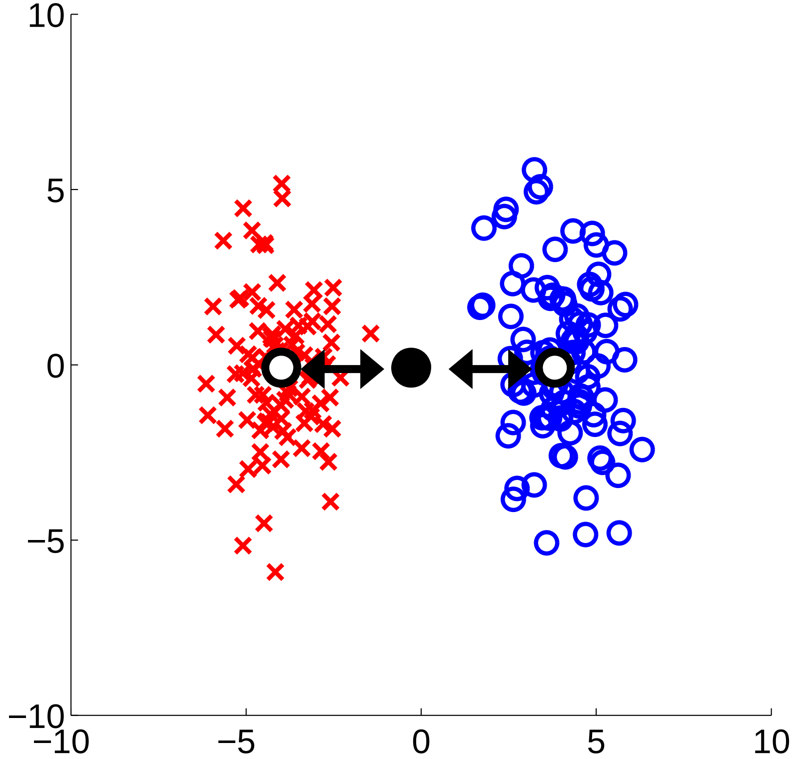
$\boldsymbol{\mu}_y$:mean of samples in class y

n_y :# of samples in class y

Between-Class Scatter Matrix

- Sum of scatter between classes:

$$\boldsymbol{S}^{(b)} = \sum_{y=1}^{c} n_y (\boldsymbol{\mu}_y - \boldsymbol{\mu})(\boldsymbol{\mu}_y - \boldsymbol{\mu})^\top$$



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

$$\boldsymbol{\mu}_y = \frac{1}{n_y}\sum_{i:y_i=y} \boldsymbol{x}_i$$

$\boldsymbol{\mu}$:mean of all samples

$\boldsymbol{\mu}_y$:mean of samples in class y

n_y :# of samples in class y

Fisher Discriminant Analysis (FDA)

Fisher (1936)

- Idea: minimize within-class scatter and maximize between-class scatter by maximizing

$$\mathrm{tr}((\boldsymbol{B}\boldsymbol{S}^{(w)}\boldsymbol{B}^\top)^{-1}\boldsymbol{B}\boldsymbol{S}^{(b)}\boldsymbol{B}^\top)$$

- To disable arbitrary scaling, we impose

$$\boldsymbol{B}\boldsymbol{S}^{(w)}\boldsymbol{B}^\top = \boldsymbol{I}_m$$

- FDA criterion:

$$\boldsymbol{B}_{FDA} = \underset{\boldsymbol{B}\in\mathbb{R}^{m\times d}}{\mathrm{argmax}}\, \mathrm{tr}(\boldsymbol{B}\boldsymbol{S}^{(b)}\boldsymbol{B}^\top)$$

$$\text{subject to } \boldsymbol{B}\boldsymbol{S}^{(w)}\boldsymbol{B}^\top = \boldsymbol{I}_m$$

FDA: Summary

- FDA criterion:

$$\boldsymbol{B}_{FDA} = \underset{\boldsymbol{B}\in\mathbb{R}^{m\times d}}{\mathrm{argmax}}\, \mathrm{tr}(\boldsymbol{B}\boldsymbol{S}^{(b)}\boldsymbol{B}^\top)$$

$$\text{subject to } \boldsymbol{B}\boldsymbol{S}^{(w)}\boldsymbol{B}^\top = \boldsymbol{I}_m$$

- FDA solution:

$$\boldsymbol{B}_{FDA} = (\boldsymbol{\psi}_1|\boldsymbol{\psi}_2|\cdots|\boldsymbol{\psi}_m)^\top$$

  - $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$:Sorted generalized eigenvalues and normalized eigenvectors of $\boldsymbol{S}^{(b)}\boldsymbol{\psi} = \lambda\boldsymbol{S}^{(w)}\boldsymbol{\psi}$

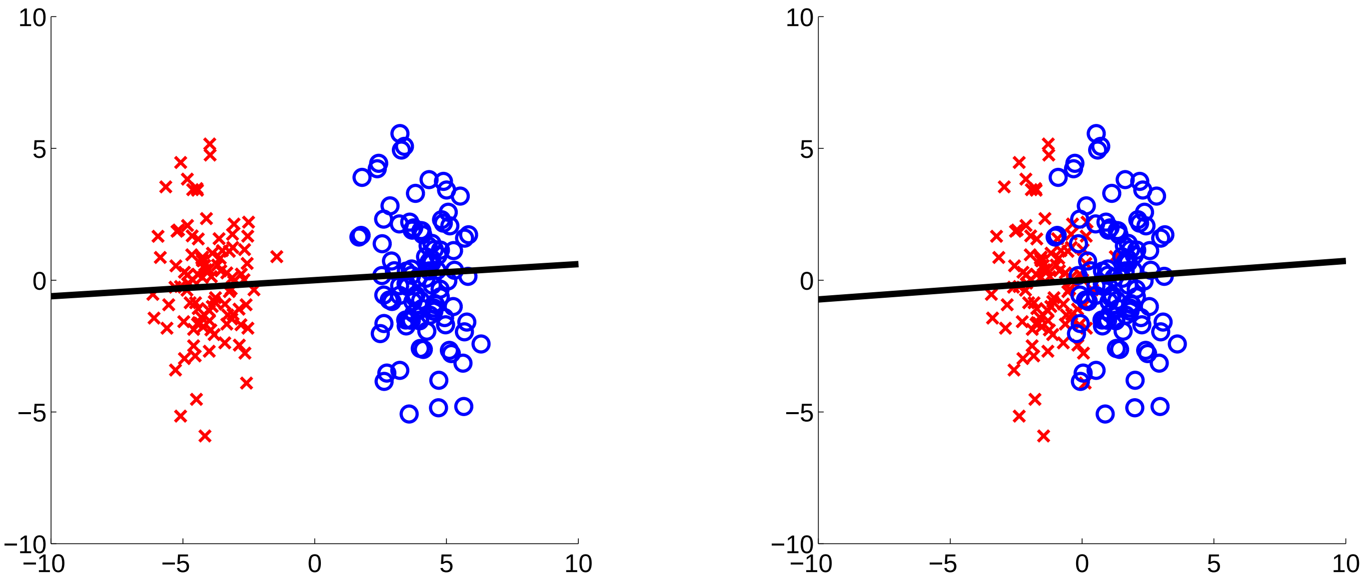
$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \qquad \langle \boldsymbol{S}^{(w)}\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle = \delta_{i,j}$$

- FDA embedding of a sample $\boldsymbol{x}$:

$$\boldsymbol{z} = \boldsymbol{B}_{FDA}\boldsymbol{x}$$

Examples of FDA

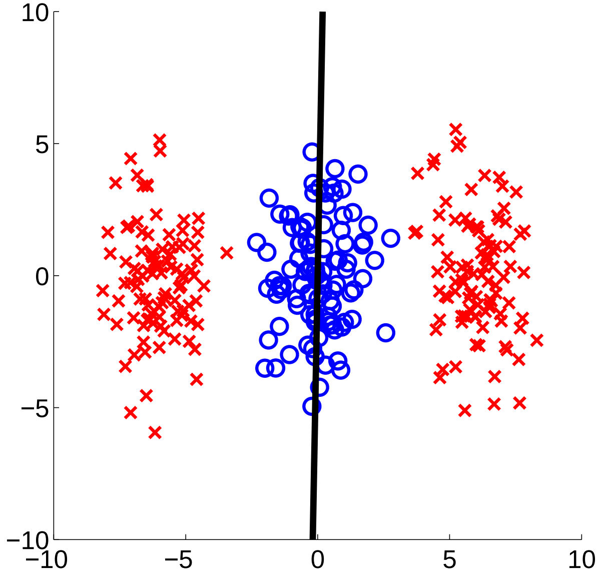
$$d = 2, m = 1 \;\; (\mathbb{R}^2 \Longrightarrow \mathbb{R}^1)$$



- FDA can find an appropriate subspace.

Examples of FDA (cont.)

$$d = 2, m = 1 \;\; (\mathbb{R}^2 \Longrightarrow \mathbb{R}^1)$$



- However, FDA does not work well if samples in a class have multimodality.

Dimensionality of Embedding Space

- We have $\mathrm{rank}(\boldsymbol{S}^{(b)}) \leq c - 1$. (Homework)
- This means $\{\lambda_i\}_{i=c}^{d}$ are always zero.

  c :# of classes $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

- Due to the multiplicity of eigenvalues, eigenvectors $\{\boldsymbol{\psi}_i\}_{i=c}^{d}$ can be arbitrarily rotated in the null space of $\boldsymbol{S}^{(b)}$.
- Thus FDA essentially requires

$$m \leq c - 1$$

- When $c = 2$, m can not be larger than 1 !

m :dimensionality of embedding space

Pairwise Expressions of Scatter

(Homework)

- $$\boldsymbol{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{Q}_{i,j}^{(w)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\boldsymbol{Q}_{i,j}^{(w)} = \begin{cases} 1/n_y & (y_i = y_j = y) \\ 0 & (y_i \neq y_j) \end{cases}$$

- $$\boldsymbol{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{Q}_{i,j}^{(b)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\boldsymbol{Q}_{i,j}^{(b)} = \begin{cases} 1/n - 1/n_y & (y_i = y_j = y) \\ 1/n & (y_i \neq y_j) \end{cases}$$

n :# of all samples

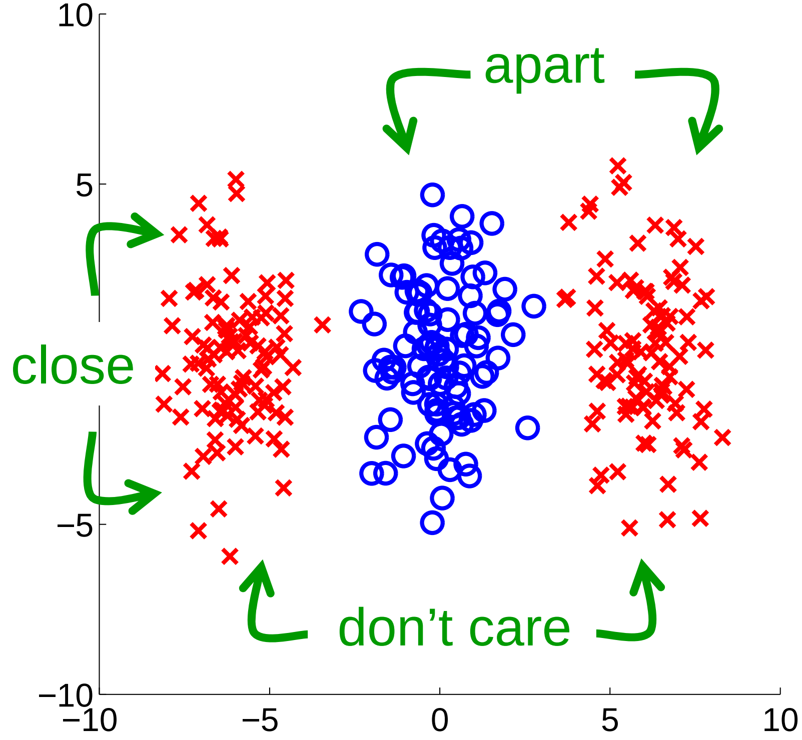
n_y :# of samples in class y

- Implication:
  - Samples in the same class are made close
  - Samples in different classes are made apart

Local Fisher Discriminant Analysis

Sugiyama (2007)

- Idea: Take the locality of data into account:
  - Nearby samples in the same class are made close
  - Far-apart samples in the same class are not made close
  - Samples in different classes are made apart



LFDA Criterion

- Local within-class scatter matrix:

$$\widetilde{\boldsymbol{S}}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \widetilde{\boldsymbol{Q}}^{(w)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$\boldsymbol{W}_{i,j}$: Similarity

$$\widetilde{\boldsymbol{Q}}^{(w)}_{i,j} = \begin{cases} \boldsymbol{W}_{i,j}/n_y & (y_i = y_j = y) \\ 0 & (y_i \neq y_j) \end{cases}$$

- Local between-class scatter matrix:

$$\widetilde{\boldsymbol{S}}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \widetilde{\boldsymbol{Q}}^{(b)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\widetilde{\boldsymbol{Q}}^{(b)}_{i,j} = \begin{cases} \boldsymbol{W}_{i,j}(1/n - 1/n_y) & (y_i = y_j = y) \\ 1/n & (y_i \neq y_j) \end{cases}$$

- LFDA criterion:

$$\boldsymbol{B}_{LFDA} = \mathop{\mathrm{argmax}}_{\boldsymbol{B}\in\mathbb{R}^{m\times d}} \mathrm{tr}(\boldsymbol{B}\widetilde{\boldsymbol{S}}^{(b)}\boldsymbol{B}^\top)$$

$$\text{subject to } \boldsymbol{B}\widetilde{\boldsymbol{S}}^{(w)}\boldsymbol{B}^\top = \boldsymbol{I}_m$$

LFDA: Summary

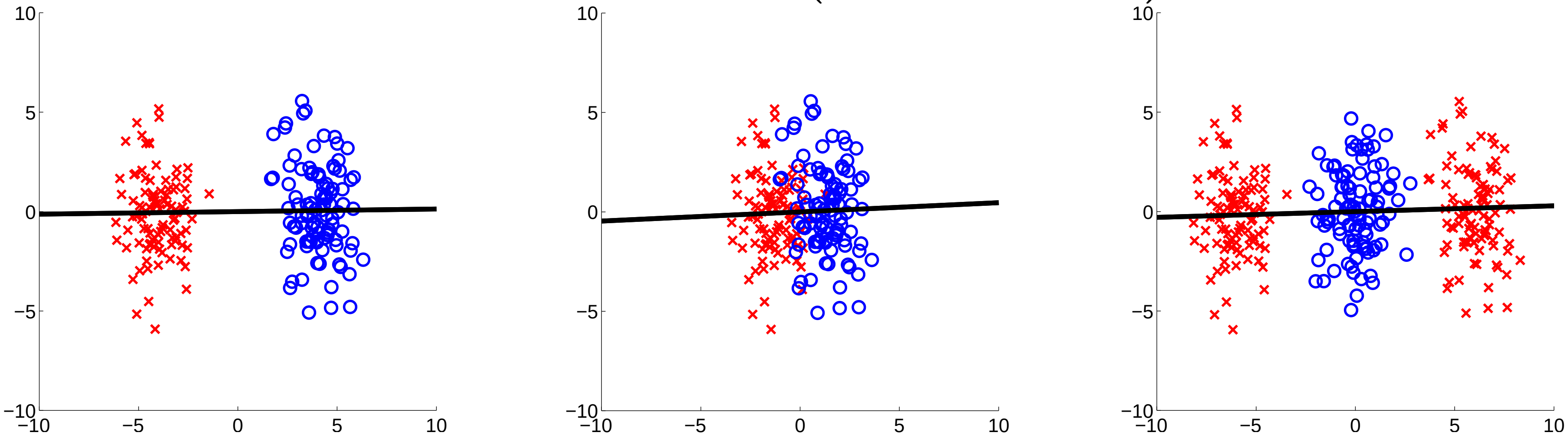
- **LFDA criterion:** $$\boldsymbol{B}_{LFDA} = \mathop{\mathrm{argmax}}_{\boldsymbol{B}\in\mathbb{R}^{m\times d}} \mathrm{tr}(\boldsymbol{B}\widetilde{\boldsymbol{S}}^{(b)}\boldsymbol{B}^\top)$$ $$\text{subject to } \boldsymbol{B}\widetilde{\boldsymbol{S}}^{(w)}\boldsymbol{B}^\top = \boldsymbol{I}_m$$
- **LFDA solution:** $$\boldsymbol{B}_{LFDA} = (\boldsymbol{\psi}_1|\boldsymbol{\psi}_2|\cdots|\boldsymbol{\psi}_m)^\top$$
  - $\{\lambda_i, \boldsymbol{\psi}_i\}_{i=1}^m$:Sorted generalized eigenvalues and normalized eigenvectors of $\widetilde{\boldsymbol{S}}^{(b)}\boldsymbol{\psi} = \lambda\widetilde{\boldsymbol{S}}^{(w)}\boldsymbol{\psi}$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \qquad \langle\widetilde{\boldsymbol{S}}^{(w)}\boldsymbol{\psi}_i, \boldsymbol{\psi}_j\rangle = \delta_{i,j}$$

- **LFDA embedding of a sample $\boldsymbol{x}$:** $$\boldsymbol{z} = \boldsymbol{B}_{LFDA}\boldsymbol{x}$$

Examples of LFDA

$$d = 2, m = 1 \ \ (\mathbb{R}^2 \Longrightarrow \mathbb{R}^1)$$



Note: Similarity matrix is defined by the nearest-neighbor-based method with 50 nearest neighbors.

- LFDA works well even for samples with within-class multimodality.
- Since $\mathrm{rank}(\widetilde{\boldsymbol{S}}^{(b)}) \gg c$, m can be large in LFDA.

c :# of classes

m :dimensionality of embedding space

Example of FDA/LFDA

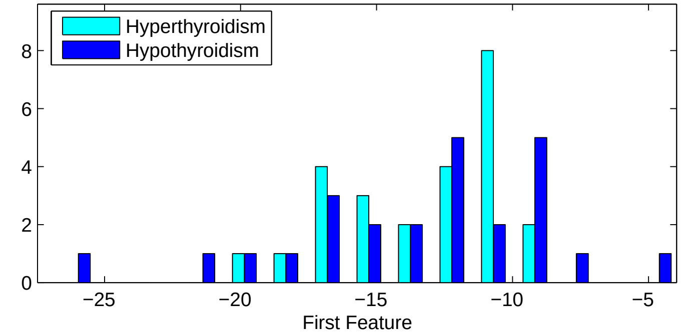
- Thyroid disease data (5-dimensional)
  - T3-resin uptake test.
  - Total Serum thyroxin as measured by the isotopic displacement method.

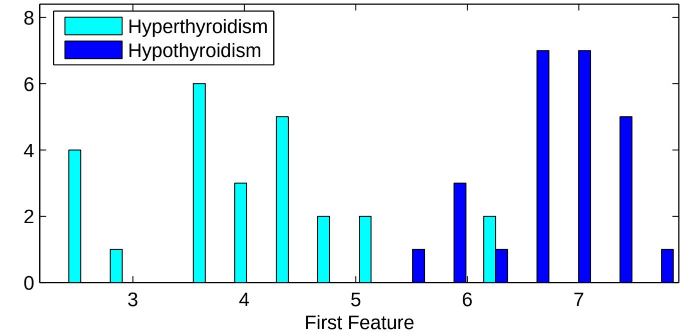
    etc
- Label: Healty or sick
- Sick can caused by
  - Hyper-functioning of thyroid
  - Hypo-functioning of thyroid

Projected Samples onto 1-D Space

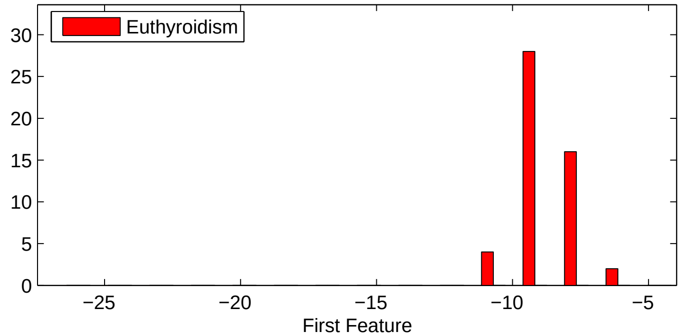
FDA

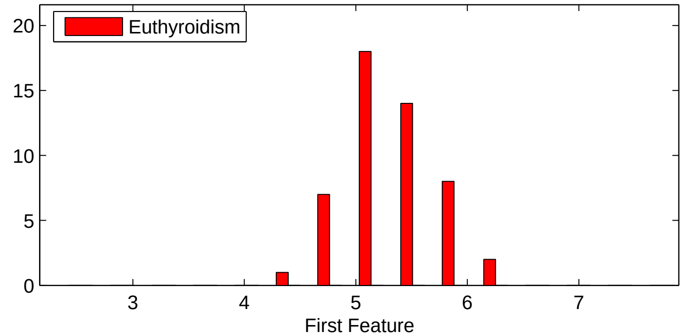
LFDA





Sick





Healthy

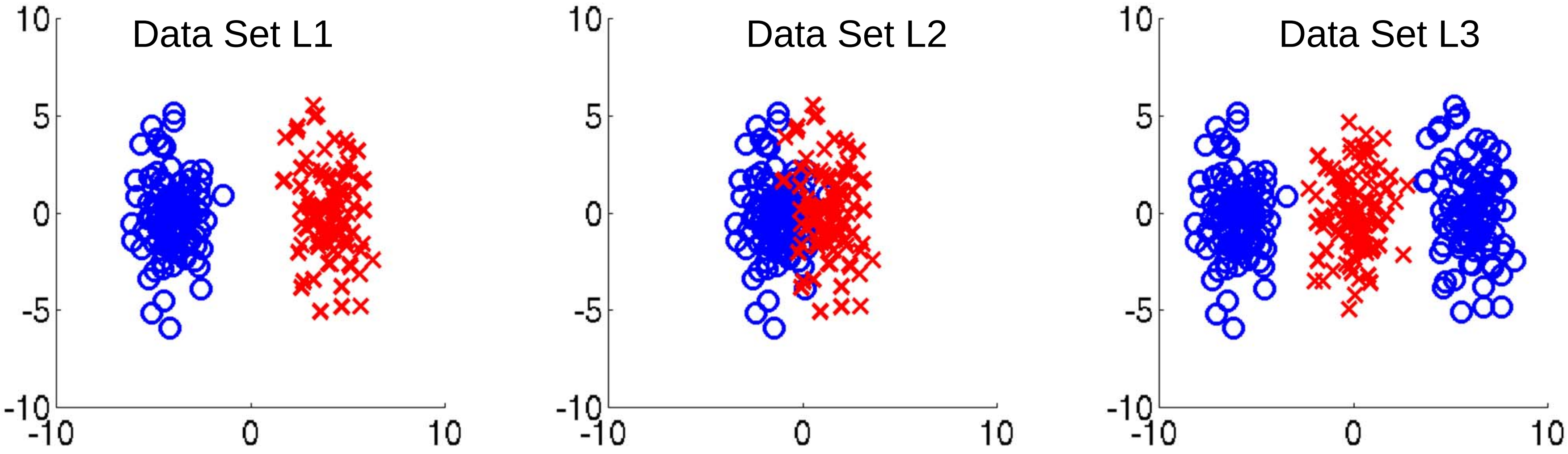
- Sick and healthy are nicely split.
- But hyper- and hypo-functioning are mixed.

- Sick and healthy are nicely split.
- Hyper- and hypo-functioning are also nicely separated.

Homework

1. Implement FDA/LFDA and reproduce the 2-dimensional examples shown in the class.

http://sugiyama-www.cs.titech.ac.jp/~sugi/data/DataAnalysis



Test FDA/LFDA with your own (artificial or real) data and analyze the characteristics of FDA/LFDA.

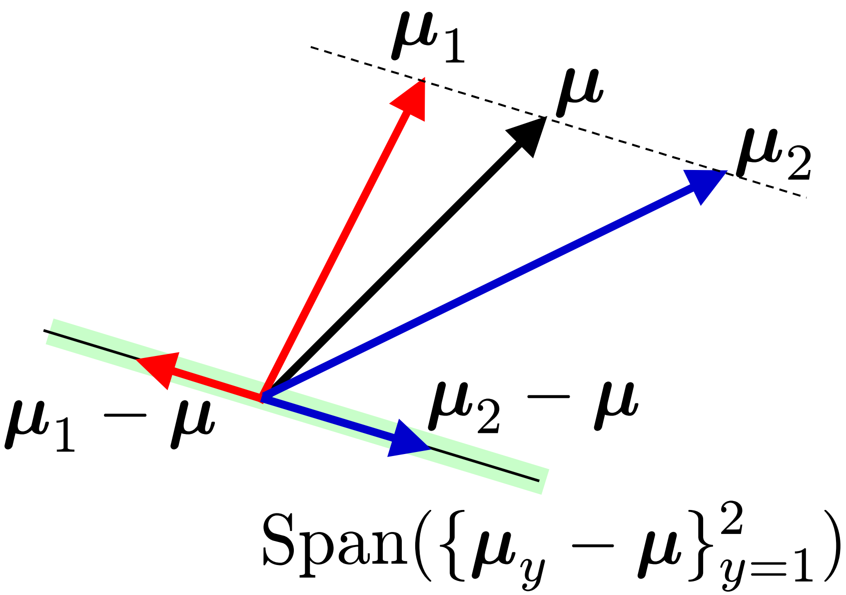
Homework (cont.)

2. Prove that $\mathrm{rank}(\boldsymbol{S}^{(b)}) \leq c-1$.

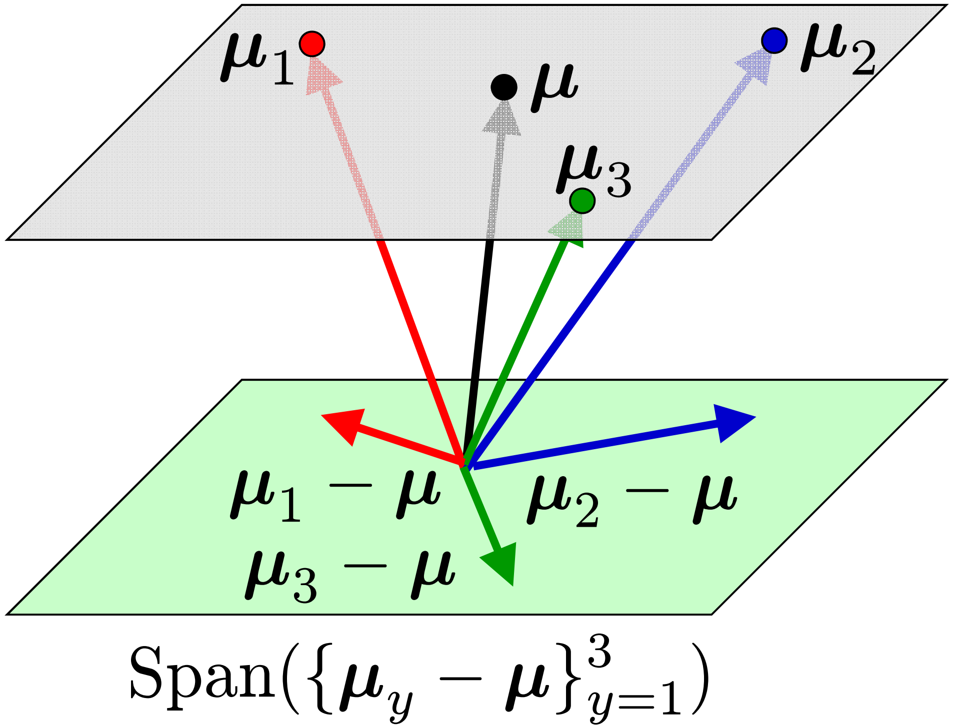
c :# of classes

Hint: Range of $\boldsymbol{S}^{(b)}$ is spanned by $\{\boldsymbol{\mu}_y - \boldsymbol{\mu}\}_{y=1}^{c}$.

- Two-class case



- Three-class case



Homework (cont.)

3. Prove that

A) $$\boldsymbol{S}^{(w)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{Q}^{(w)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

B) $$\boldsymbol{S}^{(b)} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{Q}^{(b)}_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top$$

$$\boldsymbol{Q}^{(w)}_{i,j} = \begin{cases} 1/n_y & (y_i = y_j = y) \\ 0 & (y_i \neq y_j) \end{cases} \quad \boldsymbol{Q}^{(b)}_{i,j} = \begin{cases} 1/n - 1/n_y & (y_i = y_j = y) \\ 1/n & (y_i \neq y_j) \end{cases}$$

n_y :# of samples in class y $\quad$ n :# of all samples

Hint: The use of the following mixture scatter matrix may make your life easy…

$$\boldsymbol{S}^{(m)} = \boldsymbol{S}^{(w)} + \boldsymbol{S}^{(b)} \quad \left(= \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \right)$$

Suggestion

- Read the following article for the next class:
  - B. Schölkopf, A. Smola and K.-R. Müller: Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10(5), 1299-1319, 1998.

    http://neco.mitpress.org/cgi/reprint/10/5/1299.pdf